



STiL 2019

XII Symposium in Information and Human Language Technology
and Collocates Events

October, 15 - 18, 2019, Salvador, BA

PROCEEDINGS OF CONFERENCE

@2019 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. Re-publication of material from this volume requires permission by the copyright owners.

Editors' addresses:

Carlos Augusto Prolo

Federal University of Rio Grande do Norte - UFRN
Campus Universitário Lagoa Nova
CEP 59078-970
Caixa postal 1524
Natal/RN – Brasil
prolo@dimap.ufrn.br

Leandro Henrique Mendonça de Oliveira

Brazilian Agricultural Research Corporation - EMBRAPA
Parque Estação Biológica - PqEB s/n.
Brasília, DF - Brasil - CEP 70770-901
leandro.oliveira@embrapa.br

Acknowledgments

The Program Committee chairs acknowledge the financial support to the conference provided by the Brazilian Computer Society (SBC), the Federal University of Bahia (UFBA) and B2W Digital and promotion of CAPES e CNPq. We thank the Program Committees of the XII Brazilian Symposium in Information and Human Language Technology and Collocated Events for the reviews that they produced. Last but not least, we are grateful to the local organization led by Anne Canuto and Graçaliz Dimuro (BRACIS 2019 General Chair) and Tatiane Nogueira Rios, Ricardo Araújo Rios and Marlo Vieira dos Santos e Souza (STIL Local Chair). We also appreciate the help of Henrique Nascimento Muniz de Andrade (Getulio Vargas Foundation - Rio de Janeiro, RJ, Brazil) in updating and maintaining the STIL 2019 website.

October, 2019.

Carlos Augusto Prolo
Leandro Henrique Mendonça de Oliveira

XII Symposium in Information and Human Language Technology

This volume contains the papers presented at the XII Symposium in Information and Human Language Technology (STIL 2019) and the collocated events:

- (1) VI Workshop on Portuguese Description (JDP)
- (2) VI Student Workshop on Information and Human Language Technology (TILic)
- (3) II Evaluation of Semantic Textual Similarity and Textual Inference in Portuguese (ASSIN 2)

The event took place in Salvador at October 15 - 18, 2019, Salvador, BA

STIL is the bi-annual Language Technology event supported by the Brazilian Computer Society (SBC) and by the Brazilian Special Interest Group on Natural Language Processing (CE-PLN). The conference has a multidisciplinary nature and covers a broad spectrum of disciplines related to Human Language Technology, such as Linguistics, Computer Science, Psycholinguistics, Information Science, and others. It aims at bringing together both academic and industrial participants working on those areas.

The topics of interest centered around work in human language technology in general, such as Natural Language Resources & Tools, Corpus Linguistics, Text Classification, Sentiment Analysis and Opinion Mining, Information Extraction & Retrieval, Statistical and Machine Learning Methods, Natural language interfaces, Summarization, Terminology, Lexicology and Lexicography, to name a few.

We received 48 submissions. Each paper was reviewed by exactly three of 51 reviewers from 7 countries and 38 institutions. After a rigorous reviewing process, 16 papers were selected for oral presentation, and 10 papers were selected for poster presentation.

We thank the authors for their submissions, the program committee for their hard work, invited speakers, SBC staff and the Local and General Chairs of STIL 2019.

October 2019

STIL Chairs:

- Carlos Augusto Prolo (Federal University of Rio Grande do Norte – UFRN, Natal, RN, Brazil)
- Leandro Henrique Mendonça de Oliveira (Brazilian Agricultural Research Corporation, EMBRAPA, Brasília, DF, Brasil)

VI Journey of Description of Portuguese

The Journey of Description of Portuguese (JDP) language is a satellite event of STIL – Symposium of Information Technology and Human Language, and its sixth edition took place in Salvador/Bahia between October 15-18, 2019.

The JDP aims at gathering linguists and researchers in the field of computation, more effectively integrating these two areas, which have to work in an interdisciplinary manner in order to foster advances in automatic processing in the Portuguese language.

More specifically, this edition promoted a stronger approximation of four big areas in linguistics: corpus linguistics, applied linguistics, computational linguistics and psycholinguistics, integrating these fields as new topics to the already well-established ones in JDP, which are large areas in linguistic description: studies in phonetics and phonology, lexical studies (lexicology, lexicography and terminology), syntax studies, semantics and pragmatics studies, text and discourse studies, in diverse theoretical frameworks.

Descriptive linguistics, in special, has a huge potential to bring knowledge to automatic processing of natural languages (PNL), so as to put Portuguese in a prominent position in the global scenario, joining the group of other languages (such as English, French, Spanish) which reflected this interdisciplinarity already in the sixties.

Twenty-one studies submissions were received coming from several states in Brazil. Each article has been reviewed by three members of the Program Committee, formed by 22 members of 15 universities, one of them from Portugal and the remaining ones from Brazil. After a rigorous reviewing process, ten articles were selected for oral presentation and six to poster presentation.

We thank the authors for their submissions, to the program committee for their exemplar work, as well as to the local organization and to SBC for the support.

October 2019

JDP Chairs:

- Sandra Maria Aluísio (University of São Paulo – USP/ICMC/NILC, São Carlos, SP, Brazil)
- Computational linguistics
- Stella Tagnin (University of São Paulo - FFLCH-USP, São Paulo, SP, Brazil) - Corpus linguistics
- Lilian Cristine Hübner (Pontifical Catholic University of Rio Grande do Sul - PUC-RS, Porto Alegre, RS, Brazil) - Applied linguistics, psycholinguistics and neurolinguistics
- Gustavo Lopez Estivalet (Federal University of Paraíba - UFPB, João Pessoa, PB, Brazil) – Psycholinguistics

VI Student Workshop on Information and Human Language Technology

A picture of the future?

The VI Student Workshop on Information and Language Technology, TILic, brings together undergraduate students from the fields of Computing, Linguistics and Information Sciences, among others, engaged in research related to Natural Language Processing (NLP) and Computational Linguistics.

This is the sixth edition of the event, that attracts young researchers each year to present their research topics and share experiences and ideas with more senior researchers. Since the first edition, TILic is a satellite event of the Brazilian Symposium on Information and Human Language Technology (STIL), which in 2019 takes place in Salvador, Bahia, from October 15th to 18th.

In 2019, TILic (<https://sites.google.com/view/tilic2019>) is supported by NAACL (The North American Chapter of the Association for Computational Linguistics - <http://naacl.org>) to assist the participation of students which are authors of the papers presented at the event.

In this edition, TILic had 17 paper submissions, of which 13 are being presented as poster at the event. The papers come from 10 Brazilian institutions and 1 foreign institution, and many are the result of institutional collaborations. All papers were reviewed by 3 reviewers.

A panoramic view on the works gives an idea of what is being managed within the Brazilian research groups with regard to NLP and related areas. In general, it is difficult to distinguish recurring themes in TILic 2019 - in our portrait, diversity is the norm.

In a tentative grouping, we associate the papers of Anna Furtado & Elisa Teixeira, Wesley Silva, Márcio Dias & Nádia da Silva, and Rafael Telles et al. as those that most immediately aim to use typical NLP techniques and/or resources for social issues. Furtado & Teixeira, for example, report the creation of a multilingual terminology database from a corpus composed of booklets and forms provided to refugees or refugee applicants in Brazil. The main motivation is to facilitate the communication, especially with public agencies, and the database is part of a broader project aimed at providing linguistic and cultural accessibility. The paper of Silva et al. proposes the development of a chatbot, developed with the help of machine learning, to assist students and prospective students with the administrative and academic daily life of a university. Finally, Telles et al. use a corpus of multiple choice questions associated with high school subjects to investigate how interdisciplinarity has been affecting high school subjects in Brazil.

On the other hand, there are also themes that, if not hegemonic in the NLP community, reflect challenges reminding us that we still have - the area, and not just in Brazil - basic issues to be solved. We framed in this case the papers of Quaini et al. and de Silveira et al. . In Quaini et al., the problem addressed is the de-identification in electronic medical records, that is, the need to anonymize patient information in medical records so that researchers can use real data while maintaining patient's privacy. Silveira et al. present a work dealing with technical-scientific texts in which they report pre-processing challenges when converting from .pdf to .txt, and point out the explicit need to develop quality tools capable of providing good solutions for tables, figures, equations and footnotes, among other elements that bring unwanted material into plain text format, which hinders subsequent NLP applications.

Sentiment Analysis is represented by two papers. In the first one, Wesley Santos & Ivandré Paraboni deal with the automatic recognition of moral positions in tweets. Specifically, the authors investigate the development of models for stance recognition and polarity classification, using a

corpus of Brazilian tweets related to the subjects: 'abortion', 'marijuana', 'adulthood', 'capital punishment' and 'racial quotas'. The second paper, Luana Belisario, Luiz Gabriel Ferreira & Thiago Pardo also focus on the identification of subjectivity, and the authors reproduce the only known work in the area for Portuguese, and extend it to corpora of other domains. They report good results for machine learning and lexicon-based methods and show that these methods are significantly influenced by several factors.

Other typical NLP areas, such as Automatic Summarization and Translation, are also addressed by the papers of Henrique Fonseca, Márcio Dias & Nádia da Silva, and Karina Johansson & Helena Caseli, respectively. Fonseca et al. present a work dealing with automatic generation of multi-document summaries in which they try to identify, from heuristics, linguistic errors associated with elements of textual cohesion. In the context of translation, Johansson & Caseli's paper proposes the application of word embeddings in CAT (Computer-Assisted Translation) tools to find similarities between segments of a translation memory and those of the source sentence being translated.

Also in the area of alignment - this time between text and image - João Gabriel Barbiratto & Helena Caseli aim to improve, from a linguistic point of view, the way in which the words (or word sequences) of a text that are associated to a specific area of the accompanying image. Among the strategies implemented are the inclusion of techniques capable of dealing with multi-word units and synonyms.

Continuing the dialogue between language information and tasks, the paper of Elvis de Souza & Cláudia Freitas reports the construction of a tool (designed from a linguistic perspective) to support linguists in the context of NLP tasks related to review, edit and evaluate annotated corpus. Such an environment tries to align the work done by language specialists on the one hand and the performance of annotation systems on the other hand.

To conclude the menu of TILic 2019, the papers of Vinicius Sampaio et al. and Luísa Rocha, Cláudia Freitas & Diana Santos focus on literary texts. In Sampaio et al., the authors present a comparison between different models of named entity recognition applied to Portuguese. The work of Rocha et al. brings to TILic the field of Digital Humanities, focusing on distant reading. The paper presents the challenges in the preparation of a corpus regarding its preprocessing: the attribution of gender in proper nouns, the assignment of semantic classes, and issues related to proper nouns segmentation.

As a background, we would like to thank the careful work performed by the reviewers of TILic 2019. This is a crucial part in ensuring the quality of an event. In this edition of TILic, 31 reviewers, from linguistics and computing, with Master degree or PhD, participated in the evaluation process. Thus, TILic has a dual formative character, promoting not only the formation of research within the undergraduate level, but also the participation of young researchers as referees of scientific articles - an activity of great relevance in academia, but for which we rarely receive specific training (see, for example, Smith (1990) and Cormode (2008) on writing academic reviews in the computational context).

Many thanks, therefore, to all reviewers who have devoted their time to contribute to the quality of TILic 2019.

Congratulations to the students, as well as their advisors, for the great work!

We wish you all a good read,

REFERENCES

CORMODE, G. How NOT to Review a Paper: The Tools and Techniques of the Adversarial Reviewer. SIGMOD Record, v. 37(4), 2008.

SMITH, A. J. The task of the referee. Computer, v. 23(4), 1990.

October 2019

TILic Chairs:

- Helena Medeiros Caseli (Federal University of São Carlos – UFSCar, São Carlos, SP, Brazil)
- Cláudia Freitas (Pontifical Catholic University of Rio do Janeiro - PUC-RJ, Rio de Janeiro, RJ, Brazil)

II Evaluation of Semantic Textual Similarity and Textual Inference in Portuguese

ASSIN 2 (Avaliação de Similaridade Semântica e Inferência Textual - Evaluating Semantic Similarity and Textual Entailment) is the second edition of ASSIN, an evaluation task in the scope of the computational processing of Portuguese, targeted at Recognizing Textual Entailment (RTE), also called recently Natural Language Inference (NLI), and Semantic Textual Similarity (STS).

ASSIN 2 is an effort to offer the interested community a benchmark for computational semantic tasks in Portuguese. The previous edition of ASSIN was held in conjunction with PROPOR 2016 and had six participants, three from Brazil and three from Portugal. The present edition had nine participants, which shows that semantic processing is getting more relevant to the Portuguese Natural Language Processing community. Out of those participants, five teams are from Portugal and four from Brazil.

Differently from the previous edition, our RTE/NLI task is based on determining whether a sentence entails another or not, and thus it uses only the two labels: "entailment" and "not entailment". For the STS task, we use a scale of five points to score the "proximity of meaning" between pairs of sentences. Performance is measured with the Pearson correlation index between the gold and the submitted scores, with Mean Squared Error (MSE) as a secondary metric.

The ASSIN 2 corpus is based on SICK-BR, a translation of SICK, the dataset used in the SemEval 2014 Task 1. The data is considered to be simple: all verbs are in the present continuous, for example. Complex linguistic phenomena, such as embedded clauses, reported speech and factive verbs, were avoided. This is an attempt to offer the community a semantically annotated corpus with a different style from the one used in the previous edition of ASSIN. All the data used was manually annotated by, at least, four human annotators. For the RTE/NLI task, only the labels with a majority agreement were kept. That is to say, pairs in the dataset had at least three annotators agreeing with the golden label. For the STS task, the average of the scores given by all the annotators was measured and it is the final label of the similarity score between the sentences.

For the data annotation, we thank the Group of Computational Linguistics from São Paulo University (GLiC/USP) and the team of the Stilingue Company for their great work. We also thank several individual colleagues who decided to help, completing the annotation work. We thank all the task participants and hope that they, like us, learned much from the experience.

The best results for the RTE/NLI task were by the team Deep Learning Brazil. For the STS task, the team with the best results was IPR. Congratulations to all teams!

Looking forward to more semantic processing of Portuguese in years to come.

October 2019

ASSIN 2 Chairs:

- Livy Real (B2W Digital / GLiC-USP, São Paulo, Brazil)
- Erick Rocha Fonseca (Instituto de Telecomunicações, Lisbon, Portugal)
- Hugo Gonçalo Oliveira (Centro de Informática e Sistemas da Universidade de Coimbra, Portugal)

Local Chair

- Marlo Souza (Federal University of Bahia – UFBA, Salvador, BA, Brazil)

Program Committee STIL

- Alberto Simões (University of Minho)
- Alexandre Rademaker (IBM Research)
- Andre Adami (Universidade de Caxias do Sul)
- Arnaldo Candido Junior (UTFPR)
- Carlos Prolo (Universidade Federal do Rio Grande do Norte)
- Cassia Trojahn dos Santos (IRIT & UTM2)
- Celso Kaestner (UTFPR)
- Christopher Shulby (University of São Paulo)
- Clarissa Xavier (UFRGS)
- Cláudia Freitas (Pontifícia Universidade Católica do Rio de Janeiro)
- Diana Santos (Linguoteca/Universidade de Oslo)
- Eraldo Fernandes (Universidade Federal de Mato Grosso do Sul)
- Erick Fonseca (Instituto de Telecomunicações)
- Erick Maziero (Universidade Federal de Lavras)
- Evandro Ruiz (Universidade de São Paulo)
- Geraldo Xexéo (UFRJ)
- Gustavo Estivalet (UFPB)
- Gustavo Paetzold (University of Sheffield)
- Helena Caseli (UFSCar)
- Heliana Mello (Universidade Federal de Minas Gerais)
- Hugo Gonçalo Oliveira (Universidade de Coimbra)
- Isabel Trancoso (INESC-ID / IST)
- Ivandro Paraboni (USP Leste)
- Jorge Baptista (University Algarve)
- Leandro Henrique Mendonça de Oliveira (Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA)
- Lilian Hubner (PUCRS)
- Livy Real (B2W Digital/GLiC)
- Lucelene Lopes (Roberts Wesleyan College)
- Luciano Barbosa (Universidade Federal de Pernambuco)
- Marcelo Finger (USP/IME)
- Marcos Garcia (Universidade da Corunha)
- Maria das Graças Nunes (USP/ICMC)
- Marlo Souza (Universidade Federal da Bahia - UFBA)
- Mário Silva (INESC-ID (Instituto Superior Técnico Universidade de Lisboa)
- Nelson Neto (Federal University of Pará (UFPA))
- Norton Roman (USP/EACH)
- Osvaldo de Oliveira Jr. (University of São Paulo)
- Oto Vale (UFSCar)
- Paulo Cavalin (IBM Research Brazil)
- Renata Vieira (PUCRS)

- Roger Granada (Pontifícia Universidade Católica do Rio Grande do Sul PUCRS)
- Sandra Aluísio (USP/ICMC)
- Sergio Antonio Andrade de Freitas (Universidade de Brasília)
- Stella Tagnin (USP)
- Thiago Pardo (USP/ICMC)
- Valéria Feltrim (Universidade Estadual de Maringá)
- Valeria de Paiva (Nuance Communications USA)
- Vladia Pinheiro (Universidade de Fortaleza)

Program Commitee JPD

- Ariani di Felippo (UFSCar)
- Arnaldo Candido Junior (UTFPR, Câmpus Medianeira)
- Claudia Dias de Barros (IFSP - Câmpus Sertãozinho)
- Claudia Zaváglia (UNESP-IBILCE)
- Claudia Freitas (PUC-Rio)
- Flavia Bezerra de Menezes Hirata-Vale (UFSCar)
- Guilherme Fromm (UFU)
- José Ferrari Neto (UFPB)
- Jorge Baptista (Universidade do Algarve, Portugal)
- Juliano Desiderato Antonio (UEM)
- Leonel Figueiredo de Alencar (UFC)
- Lucilene Bender de Sousa (IFRS)
- Magali Duran (NILC)
- Mahayana Cristina Godoy (UFRN)
- Pablo Arantes (UFSCar)
- Renato Basso (UFSCar)
- Tiago Torrent (UFJF)
- Oto Araujo Vale (UFSCar)

Program Commitee TILic

- Amanda Rassi - Empresa Redação Nota 1000 (Brasil)
- Ariani Di Felippo - UFSCar (Brasil)
- Arnaldo Candido Jr. - Universidade Tecnológica Federal do Paraná, campus Medianeira (Brasil)
- Carlos Ramisch - Aix Marseille Université (França)
- Claudia Dias de Barros - IFSP/Sertãozinho (Brasil)
- Christopher Shulby - SIDI (Brasil)
- Daniel Beck - University of Melboune (Austrália)
- Débora Garcia - UFSCar (Brasil)
- Eloize Rossi Marques Seno - IFSP/São Carlos (Brasil)
- Erick Galani Maziero - UFLA (Brasil)
- Erick Rocha Fonseca - USP (Brasil)
- Evandro Fonseca - PUCRS (Brasil)
- Fernando Antônio Asevedo Nóbrega - SIDI (Brasil)
- Gabriela Wick Pedro - UFSCar (Brasil)

- Isa Mara da Rosa Alves - UNISINOS (Brasil)
- Jackson Souza - UNIFAL (Brasil)
- Larissa Freitas - UFPel (Brasil)
- Larissa Picoli - UFSCar (Brasil)
- Lucelene Lopes - PUC-RS (Brasil)
- Marcella Monteiro Lemos Couto - UFSCar (Brasil)
- Marcelo Criscuolo - IFSP/Araraquara (Brasil)
- Márcio de Souza Dias - USP (Brasil)
- Marco Antonio Sobrevilla Cabezudo - USP (Brasil)
- Marcos Treviso - USP (Brasil)
- Nathan Siegle Hartmann - USP (Brasil)
- Paula Christina Figueira Cardoso - UFLA (Brasil)
- Roana Rodrigues - UFSCar (Brasil)
- Rodrigo Souza Wilkens - University of Strasbourg (França)
- Roger Granada - PUC-RS (Brasil)
- Roney Lira de Sales Santos - USP (Brasil)
- Sandra Colovini - PUC-RS (Brasil)

Invited Speaker

A History of Research: Textual and Terminological Accessibility in Public Utility Texts in Brazil

Profa. Dra. Maria José Bocorny Finatto

Abstract: The aim of this presentation is to present an overview of the scenario of Applied Linguistics and Terminology research in Brazil. These investigations, to a certain extent, connect with the issue of accessibility of written information for the general public, especially adults with limited education. In our analysis of the national scenario, we will be looking at the history of the research group “Textual and Terminological Accessibility” (ATT). Since 2011, graduate-level researchers in our group have been working on multiple perspectives in this area, in the line of research “Lexicography, Terminology and Translation” at the Universidade Federal do Rio Grande do Sul. We have drawn insights from a number of areas, including ATT, through which we rely heavily on computer technology, as well as Corpus Linguistics and Natural Language Processing. We seek to analyze texts, discourses, terminologies, lexicon and writing conventions from different areas, under the perspective of intralingual translation. Our efforts have been carried out to support actions that may facilitate the understanding of public utility information, especially in the areas of Health care and Law. (Sponsors/Grants: CNPq / CAPES / FAPERGS / SEAD-UFRGS)

Bio: M.J.B. Finatto is a Professor of Linguistics at the Department of Linguistics and Philology (UFRGS), and a Brazilian Research Fellow with CNPq. She completed her first postdoctoral research in Natural Language Processing and Readability Assessment at the Institute for Mathematical and Computational Sciences –ICMC, of the Inter-institutional Center for Research and Development in Computational Linguistics – NILC, at the University of São Paulo, in 2011. She completed her second postdoctoral research at the University of Évora, in 2017, with the project Historical Corpora of Portuguese. Finatto has a PhD in Language Studies, Linguistics, Terminology and Terminography (2001). She holds a M.A. in Linguistics and Lexicography (1993) and a B.A. in Languages and Literature – Portuguese and German (1991). Her research interests involve Natural Language Processing, Corpus Linguistics, Descriptive Terminology, Lexicography, Lexicology, Translation, Text and Discourse Studies and Scientific Communication to the General Public.

Tutorial

Modelos de Linguagem (Word Embeddings) **Profa. Dra. Renata Vieira and MSc. Joaquim Santos**

Abstract: Recent work in the area of Natural Language Processing has been impacted by sophisticated Language Models, known as Word Embeddings. Such language models capture information from the context in which the words appear and also from the characters they are composed by. In this tutorial we will talk about the evolution of these language models, how they are generated and used. We will also talk about evaluation questions of these models and their main limitations. We will also introduce some of the current Portuguese language models that are available for free use.

Bio: Renata Vieira possui título de PhD em Informática pela University of Edinburgh (1998). É professora da PUC-RS onde atua em pesquisa e ensino na área de inteligência computacional, com ênfase em processamento de linguagem natural, representação do conhecimento, ontologias, agentes e web semântica. Renata coordena o Núcleo de Inteligência Artificial da Escola Politécnica, coordena o Laboratório de Pesquisa em Processamento de Linguagem Natural e é líder do Grupo de Pesquisa do CNPq, nessa área. Possui experiência em coordenação de projetos inter-institucionais e internacionais, e participa em diversos comitês de programa de eventos científicos nacionais e internacionais (STIL, BRACIS, PROPOR, LREC, FLAIRS, FOIS, IJCAI, ACL). Participou da criação da Comissão Especial de PLN da Sociedade Brasileira de Computação, sendo a primeira presidente dessa comissão de 2007 a 2009. Com bolsa de Pesquisador Visitante Sênior CAPES-Fulbright visitou a Universidade do Texas em Austin em 2007. Participou no comitê executivo da Association for Computational Linguistics de 2011 a 2013. Em 2017 realizou estágio pós doutoral na Universidade de Toulouse. Recebeu em duas edições consecutivas prêmio de melhor artigo no Simpósio Brasileiro de Tecnologia da Informação e Linguagem Humana. É atualmente indicada pela SBC como Conferencista Senior na área de Processamento de Linguagem Natural.

Bio: Joaquim Santos é Licenciado em Matemática pela Universidade Regional do Cariri (URCA) e atualmente é aluno de mestrado em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), sob orientação da Profa. Dr. Renata Vieira. Sua principal área de estudos é o Processamento de Linguagem Natural com ênfase no Português Brasileiro. Tem realizados pesquisas sobre Modelos de Linguagem clássicos e mais recentes (como BERT e Flair Embeddings) e os aplicado no Reconhecimento de Entidades Nomeadas para o Português, como retrata seu trabalho mais recente: “Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition”. Seus principais temas de interesse são Redes Neurais, Modelos de Linguagem, Extração de Relações e Reconhecimento de Entidades Nomeadas.

Contents

STIL Short Papers

O Uso da Linguística Computacional no processo de Atribuição de Autoria em Documentos Literários <i>Paulo Varela, Michel Albonico, Edson Justino, João Lucas Varela de Assis</i>	18
Geração automática de questões de múltipla escolha <i>Joyce Martins, Eduardo Gehrke</i>	28
Avaliação das categorias afetivas do dicionário PB-LIWC2015 <i>Flavio Carvalho, Gustavo Paiva Guedes</i>	33
“Sentimento de quê?”: uma lista de sentimentos para a Análise de Sentimentos <i>Barbara Ramos, Cláudia Freitas</i>	38
Estudando personagens na literatura lusófona <i>Diana Santos, Cláudia Freitas</i>	48
Um Conjunto de Dados Extraído do Twitter para Análise de Sentimentos na Língua Portuguesa <i>Ewerton Silva, Yuri Malheiros, Rodolffo Teles Araujo Nunes, Igor Leal Antunes, Thaís Gaudêncio do Rêgo</i>	53
Predição da Complexidade Textual de Recursos Educacionais Abertos em Português <i>Murilo Gazzola, Sidney Leal, Sandra Aluísio</i>	61
Netspeak-BR: Um léxico sobre expressões criadas na língua portuguesa brasileira para a Internet <i>Rodolpho Nascimento, Leonardo Santos, Gustavo Paiva Guedes</i>	71
Identificação de discurso ofensivo no Twitter nas eleições presidenciais de 2018 no Brasil <i>Leonardo Santos, Rodolpho Nascimento, Gustavo Paiva Guedes</i>	76

Sistemas de Recomendação e Geração de Receitas Através da Categorização Ontológica dos Ingredientes <i>Luciano Pacífico, Emilia Oliveira, Larissa Britto, Teresa Ludermir</i>	81
--	----

STIL Long Papers

Um Benchmark para Sistemas de Extração de Informação Aberta em Português <i>Florencia Malenchini, Daniel Rodrigues, Rafael Glauber, Marlo Souza, Daniela Barreiro Claro</i>	86
Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural <i>Sidney Leal, Vanessa Maia Aguiar de Magalhaes, Magali Duran, Sandra Aluísio</i>	94
Generating Sense Embeddings for Syntactic and Semantic Analogy for Portuguese <i>Jéssica Silva, Helena Caseli</i>	104
Part-of-Speech Tag Embeddings for Portuguese <i>Paulo Augusto de Lima Medeiros, Bledson Bezerra, Carlos Prolo, Antonio Thomé</i>	114

Um modelo para Sistema de Diálogo Fim-a-Fim usando Conhecimento de Senso Comum <i>Cecília Carvalho, Vladia Pinheiro, Lívio Antônio Melo Freire</i>	124
Aplicação de Reconhecimento de Entidades Nomeadas em investigação de Crimes Financeiros <i>Fabio Silva, Renata Vieira</i>	134
Detecção Automática dos Heterônimos de Fernando Pessoa por Aprendizado de Máquina <i>Hugo Abonizio, Cinthyan Renata Barbosa, Arthur Artoni</i>	144
Atribuição Autoral em Blogs e Redes Sociais <i>João Martins, José Custódio, Ivandré Paraboni</i>	154
Reavaliando o dicionário LIWC português: o caso do reconhecimento de traços de personalidade e gênero autoral <i>Ricelli Ramos, Ivandré Paraboni</i>	162
Avaliação Extrínseca de Analisadores de Dependência através da Extração de Informação Aberta <i>Maurício Wanderley, Leandro Oliveira, Daniela Barreiro Claro, Marlo Souza</i>	171
Using linguistic cues to detect fake news on the brazilian portuguese paralel corpus Fake.BR <i>Emerson Okano, Evandro Ruiz</i>	181
LexPorBr Infantil: uma base lexical tripartida e com interface Web de textos ouvidos, produzidos, e lidos por crianças <i>Gustavo Estivalet, Nathan Hartmann, Vanessa Marquiafável, Katerina Lukasova, Maria Teresa Carthery-Goulart, Sandra Aluísio</i>	190
B2W-Reviews01 - An open product reviews corpus <i>Livy Real, Marcio Oshiro, Alexandre Mafra</i>	200
A Bunch of Helpfulness and Sentiment Corpora in Brazilian Portuguese <i>Rogerio Figueiredo de Sousa, Henrico Brum, Maria das Graças Nunes</i>	209
Um Corpus de Notícias Falsas do Twitter e Verificacão Automática de Rumores em Língua Portuguesa <i>Paulo Roberto Cordeiro, Vladia Pinheiro</i>	219
New developments on processing European Portuguese verbal idioms <i>Ana Galvão, Jorge Baptista, Nuno Mamede</i>	229
JDP Papers	
Enriquecendo o corpus CM2News: Construção e Anotação de Coleções Bilíngues de Notícias <i>Yasmin Vizeu Camargo, Ariani Di Felippo</i>	239
Violações linguísticas em referências a entidades do tipo “pessoa” em extratos automáticos multidocumento <i>Luana Cristini, Ariani Di Felippo</i>	244

Anotação de unidades de informação em transcrições de fala na tarefa de reconto de narrativas em português

- Leandro Borges dos Santos, Lilian Cristiane Hübner, Anderson Dick Smidarle, Letícia Lessa Mansur, Sandra Maria Aluísio* 253

Caracterização de desvios sintáticos em um corpus de redações: o processo de anotação

- Renata Ramisch, Ariani Di Felippo* 262

Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português

- Sidney Evaldo Leal, Sandra Maria Aluísio, Erica dos Santos Rodrigues, João Marcos Munguba Vieira, Elisângela Nogueira Teixeira* 270

(Re)começando a discutir as locuções verbais

- Elvis de Souza, Cláudia Freitas* 279

Quantificando (e qualificando) o sujeito oculto em português

- Claudia Freitas, Elvis de Souza, Luisa Rocha* 288

Avaliação do uso da Diversidade Contextual e da Frequência para a Tarefa de Identificação de Palavras Complexas em Simplificação Lexical

- Nathan Hartmann, Sandra Aluísio* 294

Teste de Memória de Trabalho de Leitura: Versão Computadorizada Padronizada do Reading Span Test para o Português Brasileiro

- Laiane Vasconcelos, Priscilla Almeida, Gustavo Estivalet, José Ferrari-Neto* 303

Acesso Lexical de Formas Irregulares Flexionadas em Número em Português Brasileiro

- Jefferson Alves da Rocha, José Ferrari-Neto* 312

Conjugador verbal do português brasileiro e análise morfofonológica dos radicais

- Gustavo Estivalet* 321

Córpus 4P: um córpus anotado de opiniões em português sobre produtos eletrônicos para fins de sumarização contrastiva de opinião

- Raphael Rocha da Silva, Thiago Alexandre Salgueiro Pardo* 330

Subsídios Linguístico-Computacionais para a Revisão Gramatical Automática de Redações do Ensino Médio

- Ariani Di Felippo, Milena França, Dayse Simon, Pedro Ferreira Martins* 339

Discriminação de palavras e efeitos da variação linguística

- Raquel Meister Ko. Freitag, Victor Rene Andrade Souza* 345

TILic Papers

Análise das relações entre disciplinas do Ensino Médio do Brasil por meio de questões de vestibular com uso de técnicas de PLN

- Rafael Telles, Margarethe Steinberger-Elias, André Kazuo Takahata, Luneque Silva Junior* 354

Classificação de subjetividade para a língua portuguesa

- Luana Balador Belisário, Luiz Gabriel Ferreira, Thiago Alexandre Salgueiro Pardo* 358

Reconhecimento de posicionamentos de natureza moral em textos <i>Wesley Ramos dos Santos, Ivandré Paraboni</i>	362
Melhorias linguísticas no alinhador texto-imagem LinkPICS <i>João Gabriel Melo Barbirato, Helena de Medeiros Caseli</i>	367
Investigação do uso de word embeddings para cálculo de similaridade em memórias de tradução <i>Karina Mayumi Johansson, Helena de Medeiros Caseli</i>	372
Preparação para Leitura Distante em português: diálogos entre PLN e Humanidades Digitais <i>Luísa Rocha, Cláudia Freitas, Diana Santos</i>	377
ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente <i>Elvis de Souza, Cláudia Freitas</i>	381
Um estudo sobre desidentificação de evoluções clínicas <i>Thaila Elisa Quaini, Henrique D. P. dos Santos, Sandra C. de Abreu, Bernardo S. Consoli, Renata Vieira</i>	386
Do PDF ao TXT: Desafios na extração de informação em textos técnico-científicos <i>Aline Silveira, Elvis de Souza, Tatiana Cavalcanti, Cláudia Freitas</i>	391
Identificação Automática de Erros em Sumários Multidocumento <i>Henrique Papa A. Fonseca, Márcio de Souza Dias, Nádia Félix Felipe da Silva</i>	395
Chatbot para auxiliar os discentes nos procedimentos administrativos de uma universidade <i>Wesley Benício dos Santos Silva, Márcio de Souza Dias, Nádia F. F. da Silva</i>	400
A Brief Survey of Deep Learning based methods, against OpenNLP NameFinder for Named Entity Recognition on Portuguese Literary Texts <i>Vinicius Amaro Sampaio, Mardônio J. C. França, Paulo Bruno Lopes da Silva, Gustavo Augusto Lima de Campos, Lara Domingos Hissa</i>	404
Compilação de um Banco Multilíngue de Acolhimento a Pessoas Refugiadas <i>Anna B. D. Furtado, Elisa D. Teixeira</i>	408

O Uso da Linguística Computacional na Atribuição de Autoria em Documentos Literários

Paulo Jr. Varela¹, Edson J. R. Justino², Michel Albonico¹, João L. V. de Assis¹

¹Departamento de Informática – Universidade Tecnológica Federal do Paraná (UTFPR)
– Francisco Beltrão – PR – Brasil

² Programa de Pós-Graduação em Informática – Pontifícia Universidade Católica do Paraná (PUCPR) - Curitiba – PR - Brasil.

{paulovarela, michelalbonico}@utfpr.edu.br, justino@pucpr.br,
joao.1999@alunos.utfpr.edu.br

Abstract. This paper presents an approach based on computational linguistics for the authorship attribution in Portuguese and Spanish languages. We defined a set of three classes of linguistic attributes to delineate the stylometric profile of each author: morphological, flexor and syntactic. Then, we defined the Ibero-American databases belonging to the established authors of the literature, for training and testing. The experiments were carried out with writer-dependent approach, with the authorship verification and identification strategies. The classifier was the SVM (Support Vector Machines).

Resumo. Este artigo apresenta uma abordagem baseada na linguística computacional para a atribuição de autoria em textos de língua portuguesa e espanhola. Definimos um conjunto de três classes de atributos linguísticos para delinear o perfil estilométrico de cada autor: morfológicas, flexoras e sintáticas. Em seguida, estabelecemos as bases de dados Ibero-Americanas pertencentes à autores consagrados da literatura, para treinamento e testes. Os experimentos foram realizados com a abordagem dependente do autor, com a verificação e a atribuição de autoria como estratégias de validação. O classificador foi o SVM (Support Vector Machines).

1. Introdução

Com os meios de comunicação se tornando mais acessíveis e disponíveis, problemas relacionados à atribuição da autoria em documentos digitais se tornaram mais frequentes. Com isso, diversas áreas de aplicação da atribuição da autoria vieram à tona, tais como: a disputa da autoria de textos [Holmes 1998] [Neme *et al.*, 2015] [Varela *et al.*, 2018], detecção de plágios [Burrows *et al.*, 2014], detecção e categorização de gêneros textuais [Stamatatos *et al.*, 2001] [Argamon *et al.*, 2003], mensagens de ameaça e difamação [De Vel *et al.*, 2001] [Abbassi e Chen 2005] [Zheng *et al.*, 2006], e aplicações em casos forenses [McMenamin 2002] [Chaski 2005]. Em todos os casos, a tarefa principal é descobrir se um documento foi redigido por determinado autor ou saber o autor do documento entre diversos suspeitos. Para isso, amostras de textos de diversos autores são coletadas e armazenadas em uma base de dados, de onde são extraídas informações que fornecem um subconjunto de características de estilo.

Posteriormente, amostras desta base são confrontadas com a amostra de texto que está sendo questionada. Ao final, a ideia básica é saber se a amostra questionada e a amostra de um determinado autor foram escritas ou não pelo mesmo indivíduo.

Para solucionar este tipo de problema, onde uma análise manual não é possível, pela grande quantidade de informação a ser processada, o uso da aprendizagem de máquina é uma das mais usuais na literatura [Argamon *et al.*, 2003] [Juola 2006] [Stamatatos 2009] [Varela *et al.*, 2018]. Entretanto, além das técnicas de aprendizagem de máquina é necessário o trato das questões associadas à determinação da autoria, quer seja na língua falada ou escrita. Para isso, pode-se fazer uso da linguística computacional, que tem por função usar meios computacionais para manipulação da linguagem humana. A linguística computacional permite que a atribuição da autoria de um documento possa ser feita, pois o processo consiste em rotular/classificar cada palavra das amostras de textos de acordo com o seu nível estrutural, morfológico ou sintático, por exemplo. Com isso, é possível avaliar e identificar as características que tornam um texto, bom ou ruim para discriminar o estilo de um autor ou de um grupo de autores. Sendo assim, o estilo é considerado um elemento variável do comportamento humano que possui um conjunto de padrões gramaticais, que são conhecidos como características estilométricas [Stamatatos 2009]. Com a estilometria, a aplicação do estilo linguístico aprendido no texto é possível realizar a parametrização das características de cada autor, e assim, conseguir identificar padrões na escrita.

A verificação e a identificação da autoria podem ser executadas por intermédio da observação de atributos linguísticos, tais como os estilísticos, apresentados pelo autor do documento. Na verificação da autoria o objetivo principal é verificar se o modelo criado no treinamento é robusto o suficiente para conseguir classificar corretamente amostras de textos de um mesmo autor. A estratégia neste tipo de abordagem é um-contra-um. O autor que se deseja comparar é conhecido, realizando o processo de verificação com os modelos deste autor. O objetivo é verificar se ele acerta ou erra. No método de identificação de autoria o objetivo é confrontar a base de textos contra todos os autores, em busca de identificar quem é o autor da amostra que está sendo questionada. A estratégia nesta abordagem é um-contra-todos, ou seja, confrontar o texto questionado contra todos os modelos do treinamento, afim de tentar identificar o provável autor. Neste caso, como o confrontamento é grande e de difícil decisão por parte do classificador, é efetuada a análise dos autores melhores classificados (*Top-list*).

Este trabalho tem como base o uso de recursos estilísticos para gerar um perfil estilométrico de cada autor. Utilizamos a linguística computacional, que consiste em classificar cada palavra de uma amostra de texto, de acordo com o seu nível linguístico. A ideia principal é extrair funções sintáticas de cada palavra, necessárias para a formação de uma frase, tais como: sujeito, predicado, verbo, advérbio, objeto direto, entre outros. Para treinamento e testes foi utilizado o classificador SVM (*Support Vector Machines*). Aplicamos a abordagem em dois diferentes idiomas (português e espanhol) utilizando a base de textos literários disponibilizada em domínio público.

Neste trabalho apresentamos duas contribuições: Primeiro, apresentar uma abordagem baseada na linguística computacional que seja discriminante e aplicável em casos de atribuição de autoria. E, averiguar o comportamento do conjunto de atributos em textos literários em dois diferentes idiomas (português e espanhol). Adicionalmente,

exploramos outros questionamentos, tais como: (i) qual o comportamento da abordagem conforme a variabilidade da quantidade de informação textual de cada amostra? (ii) qual o comportamento do conjunto de atributos para os dois idiomas?

Este artigo está estruturado da seguinte maneira: A seção 1 consiste da introdução. A parte 2 descreve os materiais e métodos. A parte 3 apresenta os experimentos e a análise dos resultados. A comparação dos resultados com a literatura é apresentada na parte 4. E, na parte 5 são apresentadas as conclusões e trabalhos futuros.

2. Materiais e Métodos

Nesta seção, apresentamos as bases de dados. Posteriormente, detalhamos o conjunto de características estilométricas. E, por conseguinte, descrevemos a abordagem proposta.

2.1. Bases de Dados

Usamos duas bases de dados contendo textos de autores contemporâneos da literatura ibero-americana (línguas portuguesa e espanhola). Escolhemos tais línguas, por derivarem de uma mesma “língua-mãe”, entretanto, com diferentes evoluções em seus aspectos gramaticais e sintáticos. Além disso, as línguas são ricas linguisticamente, o que reflete diretamente no contexto dos experimentos, que é avaliar os recursos da linguística computacional como atributos discriminantes na atribuição de autoria.

Cada base de dados é composta por 100 autores distintos, com 30 amostras de textos por autor. Cada amostra possui no mínimo 1000 frases, sendo que em média cada frase é composta por 10-15 palavras. Entretanto, para avaliar melhor o comportamento da abordagem variamos a quantidade de informação, ou seja, subconjuntos aninhados por quantidade de 10, 25, 50, 75, 100, 250, 500, 750 e 1000 frases por amostra.

Para validar a abordagem, utilizamos o processo de validação cruzada. Todos os textos foram pré-processados para remover textos residuais (número de páginas, cabeçalho, rodapé, entre outros elementos que poderiam influenciar na análise do estilo de cada autor). A coleção de textos inclui contos, romances e novelas.

2.2. Conjunto de Atributos

Com o intuito de identificar os padrões de escrita de cada autor, propomos um grupo de 114 características linguísticas de estilo, divididas em 3 categorias, que representam os níveis sintáticos da frase: morfológicas (classes gramaticais), flexoras (modificações das palavras), sintáticas (regem as construções das frases) (Ver Tabela 1).

Tabela 1. Conjunto de Características Linguísticas

Grupo	Características (V_t)
Morfológicas	Substantivos, determinantes, pronomes, adjetivos, advérbios, verbos, preposições, conjunções e etc.
Flexoras	Número (singular, plural), gênero (masculino, feminino), pessoa (primeira, segunda, terceira), tempo (passado, presente, futuro), etc.
Sintáticas	Sujeito, predicado, objeto direto, objeto indireto, verbo, adjunto, complemento do objeto, etc.

Para uma análise linguística completa, uma única palavra pode ter várias funções sintáticas em uma frase, dependendo do nível de classificação. Assim, é possível que uma palavra pertença a vários grupos de características. Para otimizar o processo,

utilizamos a ferramenta desenvolvida por [Bick 2000] para realizar a rotulagem de cada palavra nos idiomas português e espanhol. Com isso, a rotulagem das características da Tabela 1 foi formada.

2.3. Abordagem

Para os experimentos propomos a abordagem dependente do autor. A abordagem dependente do autor é baseada em um modelo para cada autor, ou seja, é embasada na policotomia [Varela *et al.*, 2018] [Halvani *et al.*, 2016] [Savoy 2011]. Nesta abordagem, muitas amostras por autor são necessárias, porque o objetivo principal é enfatizar as características individuais de cada autor. Todos os autores participam das fases de treinamento e testes. Por esse motivo, o subconjunto de amostras usadas na fase de treinamento não é usado na fase de testes, mas pode ser usado como referência. Utilizamos o conceito de dissimilaridade, que está relacionada com a ideia de que cada objeto (x) é descrito por suas diferenças em relação a um conjunto de objetos (R) [Pekalska e Duin, 2002]. Neste caso, cada objeto x é representado por um vetor de dissimilaridade $D(x, R) = [d(x, p_1), d(x, p_2), \dots, d(x, p_n)]$ para os objetos $P_i \in R$.

Na Figura 1, pode ser observado uma visão geral da abordagem: (a) Todas as amostras de textos de autores conhecidos (A_c) ou autores desconhecidos (A_d), são organizadas. (b) Cada amostra de texto é segmentada em frases, e cada frase é processada para extração dos atributos. (c) É efetuado o processo de rotulagem, onde as informações sintáticas de cada autor são extraídas através do processo apresentado em [Bick 2000]. O processo de rotulagem mostra que, para cada palavra, em diferentes idiomas podem haver vários rótulos, ou seja, uma palavra pode ter várias funções sintáticas em uma única frase. (d) O conjunto de atributos é dividido em três vetores diferentes (conforme Tabela 1). Depois de rotular cada palavra, os rótulos são transformados em informações numéricas, que preencherão cada vetor. (e) Para cada texto T contendo F_k frases, os vetores V_{t_i} são criados com base na Tabela 1, onde $i \in R \wedge 1 \leq i \leq 3$ e i é o número de frases. O número de palavras N_k que compõe cada frase F_k é calculado, e o número de palavras marcadas em cada classe é computado. Os vetores de V_{t_i} , que se referem às três classes da Tabela 1, contém o número de vezes que cada atributo aparece na frase. Os vetores V_{t_i} são divididos pelo número de palavras N_k na frase, para criar os vetores F_i . Então, dado um conjunto de vetores de atributos $D_p = U^3_{i=1} F_i$, onde $p \in R \wedge 1 \leq p \leq \theta$ e θ é o número de amostras de cada autor para o procedimento de treinamento. (f) Diante disso, um conjunto de amostras genuínos é criado pela combinação de amostras de um mesmo autor $Z_{(+)}$. Um subconjunto de amostras é usado como referência e para treinamento, e outro subconjunto é usado para testes. (g) O subconjunto falso $Z_{(-)}$ é gerado pela combinação de amostras de diferentes autores. Neste caso, um vetor de atributos de um autor A é combinado com um vetor de atributos de um autor B. (h) A combinação das amostras positivas $Z_{(+)}$ e negativas $Z_{(-)}$ geram um conjunto de treinamento T_s . (i) É iniciado o processo de treinamento, onde são gerados os modelos de cada autor. (j) Então, um conjunto de vetores de testes Q_α , onde $\alpha \in R \wedge 1 \leq \alpha \leq \omega$ e ω é o número de autores, que é integrante de um subconjunto de vetores de atributos D_t , onde $t \in R \wedge 1 \leq t \leq \psi$ e ψ é o número de amostras de cada autor para os testes. O procedimento básico é calcular o vetor de dissimilaridade entre uma instância de Q_α e um subconjunto de amostras de referências R_p de um autor. (j) Um conjunto de resultados parciais Pr_{ap} é obtido, pela saída de cada um dos

classificadores (grupos da Tabela 1). (k) É tomada a decisão final pelo somatório dos resultados parciais, onde φ é o número amostras de referências utilizadas (Equação 1).

$$Fd_a = \sum_{i=1}^{\varphi} Pr_{ap} \quad (1)$$

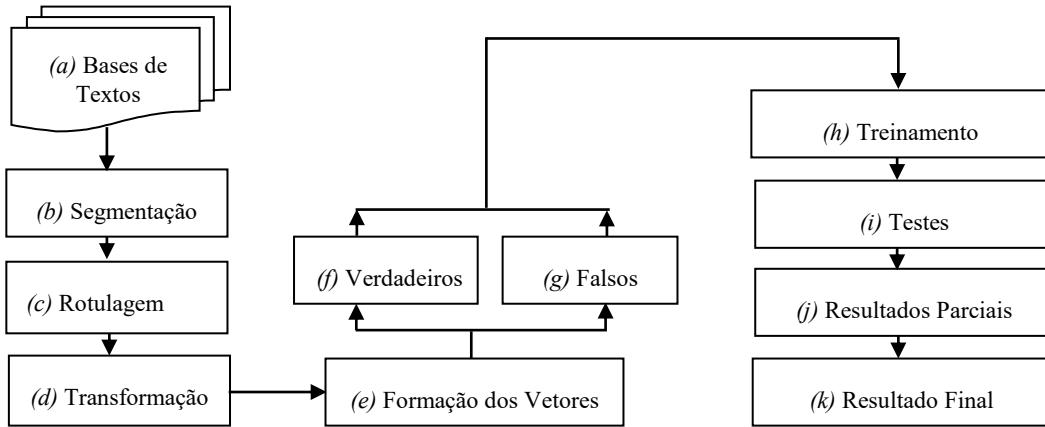


Figura 1. Visão Geral da Abordagem

3. Resultados e Discussão

Os experimentos estão divididos em duas partes: inicialmente, os resultados da verificação de autoria, e posteriormente, da identificação de autoria. Em todos os casos as taxas de acerto são representadas pela acurácia, que mostra o quanto o resultado do experimento é próximo do seu valor real, ou seja, quanto o modelo é preciso e confiável. O processo de *cross-validation* foi utilizado para dividir a base de dados. Para a aprendizagem e tomada de decisão utilizamos o SVM (*Support Vector Machines*) com *kernel* linear, que apresentou os melhores resultados em nossos experimentos e também bons resultados na literatura [Stamatatos 2009] [Varela *et. al.* 2018]. Em todos os experimentos utilizamos 9 amostras de textos como base de referência.

3.1. Verificação de Autoria

Na verificação de autoria, o objetivo principal é determinar se um texto questionado foi escrito por um determinado autor, ou seja, representa um problema de duas classes (autoria ou não-autoria). Diante disso, quando temos um vetor de características de um texto questionado Q_α , que pertence *a priori* a um autor desconhecido (A_d), o objetivo é determinar se o texto questionado pertence a um autor conhecido (A_c) ou não.

Na Tabela 2, são evidenciados os resultados da verificação de autoria para os textos em língua portuguesa e espanhola. Como é possível observar, variamos a quantidade de frases por amostra (F_k) com o intuito de observar o comportamento da abordagem com pequena ($F_k = \{10, 25, 50\}$), média ($F_k = \{75, 100, 250\}$), e grande ($F_k = \{500, 750, 1000\}$) quantidade de informação.

Em língua portuguesa observamos que as taxas de acerto para amostras com pequena quantidade de informação, variaram entre 68-73%, preconizando resultados promissores, já que dispomos de poucas informações de cada autor. Para amostras com

média quantidade de informação, a acurácia variou entre 78-92%, indicando taxas de acerto expressivas para a língua portuguesa. E, com amostras com grande quantidade de informação, as taxas de acerto ficaram entre 94-96%, indicando a robustez do conjunto de características linguísticas utilizadas neste trabalho.

Tabela 2. Resultados - Verificação de Autoria

Língua	Taxa de Acerto em % por quantidade de Frases (F_k)								
	10	25	50	75	100	250	500	750	1000
Portuguesa	68.8	70.7	73.9	78.9	87.3	92.1	94.5	96.2	96.9
Espanhola	66.5	72.1	73.4	79.4	86.5	91.7	94.2	95.9	96.4

Para a língua espanhola, os resultados foram muito semelhantes aos apresentados na língua portuguesa. Para amostras com pouca informação, as taxas de acerto ficaram entre 66-73%. Quando testadas as amostras com média quantidade de informação, a acurácia ficou entre 79-91%. E, com grande quantidade ($F_k \geq 500$) as taxas de acerto foram entre 94-96%. Esta semelhança nos resultados, podem estar diretamente ligadas às estruturas sintáticas, pois ambas as línguas possuem uma gramática e sintaxe muito próximas, indicando que existe uma constância e robustez do modelo, já que existe uma baixa variabilidade dos resultados em idiomas distintos.

Sobre os experimentos, percebe-se que conforme incrementamos o número de frases em cada amostra de texto (F_k) do autor, os resultados vão evoluindo consideravelmente. Isso significa, que quanto mais informações textuais tivermos de cada autor, melhor será a performance do modelo. Ainda que, com a limitação de frases ($F_k = 50$), que em média é um texto de 500 palavras, os resultados apresentados na verificação de autoria estão de acordo com a literatura. Observamos que as características sintáticas se sobressaem nos resultados, retornando maiores contribuições na taxa de acerto do que as características morfológicas e flexoras.

3.2. Identificação de Autoria

A identificação de autoria consiste em identificar o autor desconhecido (A_d) de um texto questionado T_d , onde $c \in R \wedge 1 \leq c \leq \xi$, onde ξ é o número de autores constantes na base de dados. Para isso, é necessária a maximização da relação $F_d = \max \{D_i(x, R_c)\}$, para obter o retorno estimado da probabilidade de acerto *a posteriori*, onde D_i representa o modelo treinado para a abordagem dependente do autor. A estimativa, indica se um texto T_d e as amostras de referência R_c pertencem ou não à um mesmo autor. Na identificação de autoria é fornecida uma lista de amostras de textos que são mais semelhantes à amostra questionada, ou seja, uma *Top-list*. Esta lista tem a função de fornecer maiores subsídios para tomada de decisão em ambientes complexos, ou seja, uma amostra será considerada correta se pelo menos uma ocorrência for listada entre as listas *Top-1*, *Top-5*, por exemplo. Apesar que o resultado almejado seja próximo de 100%, ou seja, estar no *Top-1*, muitas vezes pode se tomar decisão com base *Top-list*.

De acordo com a Tabela 3, podemos observar que em textos de língua portuguesa, a abordagem se mostrou eficiente. Para $F_k \leq 50$, as taxas de acerto variaram entre 62-64% para *Top-1*; 64-68% para *Top-3*; e, 70-72% para *Top-5*. Isso indica um resultado promissor, pois a abordagem consegue reduzir a dimensionalidade inicial de 100 autores para 5 autores com média de 70% de acerto. Nos experimentos variando F_k

$= \{75, 100, 250\}$, percebemos que para *Top-1* a acurácia foi entre 67-76%; *Top-3* entre 69-79%; e, para *Top-5* entre 74-83%. Com uma maior amplitude de informações de cada autor, ou seja, $F_k = \{500, 750, 1000\}$, os melhores resultados foram atingidos, sendo entre 81-84% para *Top-1*; 83-86% para *Top-3*; e, 86-89% para *Top-5*. Esses resultados demonstram que características morfológicas, flexoras e sintáticas são bons atributos para classificar textos questionados.

Tabela 3. Resultados - Identificação de Autoria

Língua	Top-List	Taxa de Acerto em % por quantidade de Frases (F_k)								
		10	25	50	75	100	250	500	750	1000
Portuguesa	Top-1	62.7	63.1	64.9	67.6	72.0	76.9	81.9	83.9	84.2
	Top-3	65.4	67.8	68.3	69.7	73.5	79.0	83.5	84.9	86.0
	Top-5	70.9	72.1	72.3	74.5	79.0	83.5	86.1	87.9	89.9
Espanhola	Top-1	62.1	63.7	65.0	66.9	72.4	78.3	83.2	85.4	86.7
	Top-3	66.0	66.9	67.8	69.3	74.0	80.2	85.3	86.7	88.0
	Top-5	69.7	72.0	71.3	72.9	76.8	84.5	86.9	87.6	90.1

Em língua espanhola podemos observar que os resultados para $F_k = \{10, 25, 50\}$, obtiveram taxas de acerto entre 62-65% para *Top-1*; 66-67% para *Top-3*; e, 69-71% para *Top-5*. Na avaliação com média quantidade de informação, a acurácia variou entre 66-78% para *Top-1*; 69-80% para *Top-3*; e, 72-84% para *Top-5*. Com grande quantidade de informação, a acurácia foi entre 83-86% para *Top-1*; 85-88% para *Top-3*; e, 86-90% para *Top-5*. Percebemos que as taxas de acerto em língua espanhola foram muito similares aos resultados em língua portuguesa. Novamente, isso nos mostra que a abordagem é aplicável em idiomas distintos, e também, evidencia a contribuição do conjunto de atributos para o processo de identificação de autoria.

4. Comparação dos Resultados

Na Tabela 4, comparamos a nossa abordagem com a literatura. A comparação não é exata, pois os protocolos e as bases de dados não são as mesmas. No entanto, conseguimos estimar as contribuições efetuadas pela abordagem proposta.

Primeiramente, comparamos com o trabalho desenvolvido por Pavelec *et al.*, (2008), que avaliou textos de colunas jornalísticas em língua portuguesa para a identificação de autoria. Neste caso, nossa abordagem se sobressaiu com 6% de ganho de performance. Quando comparado ao trabalho realizado por Varela *et al.*, (2011) na identificação de autoria, a abordagem proposta obteve um ganho de 13%. Na comparação com Oliveira Jr. *et al.*, (2013) que utilizou algoritmos de compressão de dados para avaliar a similaridade na verificação de autoria, nossa abordagem ficou cerca de 2% abaixo, isso pelo fato de Oliveira Jr. *et al.*, (2013) não avaliar e usar características linguísticas. Em comparação com o trabalho desenvolvido por Varela *et al.*, (2016), percebemos que os resultados são semelhantes, sendo que nossa abordagem produz taxas de acerto levemente superiores na verificação de autoria, e cerca de 4% de ganho na comparação da identificação de autoria. Já para o trabalho de Halvani *et al.*, (2016) que trabalhou com textos em língua espanhola, nossos resultados foram superiores e mais de 20%. E, por fim, comparando com Varela *et al.*, (2018) nosso trabalho mostra-se praticamente equivalente, entretanto, produzindo resultados

levemente inferiores na verificação de autoria, e com um pouco mais de 4% de perda de performance na identificação de autoria. Muito provavelmente, isso se deve por utilizarmos um escopo menor de características linguísticas.

Tabela 4. Quadro Comparativo com a Literatura

Autor	Base de Dados	Atributos	Número de Autores	Idioma	Verificação	Identificação
Pavelec <i>et al.</i> , (2008)	Jornais	Palavras-função	30	Português	-	75-83%
Varela <i>et al.</i> , (2011)	Jornais	Palavras-função	100	Português	-	76%
Oliveira Jr. et al (2013)	Jornais	Compressão	30-100	Português	77-99%	-
Varela <i>et al.</i> , (2016)	Jornais e Literários	Atributos Sintáticos	20/100	Português	96%	78-85%
Halvani <i>et al.</i> , (2016)	Variados	Variados	milhares	Espanhol	72%	-
Varela <i>et al.</i> , (2018)	Jornais e Literários	Atributos Sintáticos	20/100	Português e Espanhol	75-98%	78-93%
Abordagem Proposta	Literários	Atributos Sintáticos	100	Português Espanhol	68-97% 66-96%	62-89% 62-90%

Diante disso, podemos responder aos questionamentos propostos na introdução: *(i) qual o comportamento da abordagem conforme a variabilidade da quantidade de informação textual de cada amostra?* Percebemos que com pouca informação textual o modelo produz resultados satisfatórios, entretanto, conforme vamos aumentando a quantidade de informação sintáticas, os resultados vão ganhando performance tanto na verificação como na identificação de autoria. Isso indica, que quanto mais informações linguísticas tivermos sobre o autor, melhor será a tomada de decisão do classificador. *(ii) qual o comportamento do conjunto de atributos para os dois idiomas?* Verificamos que em ambos os idiomas (português e espanhol) o conjunto de atributos teve o mesmo comportamento quanto ao seu desempenho, mesmo quando variamos a quantidade de informação das amostras. Isso nos indica que o conjunto de características linguísticas propostas neste trabalho são robustas e eficientes para atuar em casos onde é necessário verificar ou identificar a autoria de um determinando documento.

5. Conclusões

Este trabalho teve por finalidade apresentar uma abordagem baseada na linguística computacional para atribuição de autoria em textos de língua portuguesa e espanhola. Trabalhamos com a abordagens dependente do autor, e com as estratégias de verificação e identificação de autoria. Ao todo, foram usados 114 atributos estilométricos. Tais atributos, alimentaram os vetores de dissimilaridade, que posteriormente foram utilizados para treinamento e testes através do classificador SVM. Em língua portuguesa, atingimos taxas de acerto entre 68-97% para verificação de autoria, e de 62-89% para identificação de autoria. Para língua espanhola a acurácia foi de 66-96% para a verificação de autoria e de 62-90% para a identificação de autoria.

Enfim, constatamos que o uso da linguística computacional para o reconhecimento de padrões de escrita em língua portuguesa e espanhola é viável, pois aponta resultados promissores. Percebemos que a abordagem se mostrou estável e robusta perante os experimentos. Como trabalhos futuros pretendemos avaliar a abordagem com outros tipos de textos, e, incrementar e avaliar atributos semânticos. Por conseguinte, testar a abordagem em mais línguas latinas.

Referências

- Abbasi, A. e Chen, H. (2005) “Applying authorship analysis to extremist group web forum messages”, In: IEEE Intelligent Systems, vol. 20, nº 5, p. 67-75.
- Argamon, S. Koppel, M. Fine, J. e Shimoni, A. (2003)“Gender, genre, and writing style in formal written texts”, Text, vol. 23(3), p.321–346.
- Bick, E. (2000) “The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework”, Århus.
- Burrows, S. Uitdenbogert, A. L. e Turpin, A. (2014) “Comparing techniques for authorship attribution of source code”, Software: Practice and Experience, vol. 44 (1), p.1-32.
- Chaski, C. E. (2005) “Who’s at the keyboard? - authorship attribution in digital evidence investigations”, In: International Journal of Digital Evidence, vol. 4(1), Spring.
- De Vel, O. Anderson, A. Corney, M. e Mohay, G. (2001) “Mining e-mail content for author identification forensics”, ACM SIGMOD, vol. 30, 4, p. 55–64.
- Halvani, O. Winter, C. e Pflug, A. (2016) “Authorship Verification for Different Languages, Genres and Topics”, Digit. Investig., 16(S):S33–S43.
- Holmes, D. I. (1998) “The evolution of stylometry in humanities scholarship”, In: Literary and Linguistic Computing, vol. 13, nº 3, p.111-117.
- Homem, N. Carvalho, J. P. (2011) "Authorship Identification and Author Fuzzy Fingerprints", In NAFIPS2011 - 30th Annual Conference of the North American Fuzzy Information Processing Society, El Paso, TX, USA.
- Juola, P. (2006) “Authorship attribution for electronic documents”, In: M. Olivier & S. Shenoi (Eds.), Advances in digital forensics II, p. 119–130. Boston: Springer.
- Lowe, D. e Matthews, R. (1995) “Shakespeare vs. Fletcher: a stylometric analysis by radial basis function”, Computer and the Humanities, vol. 29, p. 449-461.
- McMenamin, G. R. (2002) “Forensic Linguistics – Advances in Forensic Stylistics”, CRC Press, New York.
- Michie, D. Spiegelhalter, D. J. Taylor, C.C. (1994) “Machine Learning, Neural and Statistical Classification”, 1994.
- Neme, A. Pulido, J. R. G. Munoz, A. Hernandez, S. e Dey, T. (2015) “Stylistics analysis and authorship attribution algorithms based on self-organizing maps”, In: Neurocomputing, vol. 147, p. 147-159.
- Oliveira Jr, W. Justino, E. Oliveira, L. S. (2013) “Comparing compression models for authorship attribution” Forensic Science International, 228 (1-3), p.100-104.

- Pavelec, D. Justino, E. J. R. Batista, L. V. e Oliveira, L. E. S. (2008) “Author Identification Using Writer-dependent and Writer-independent Strategies” In: Proceedings of the 2008 ACM Symposium on Applied Computing, p. 414–418, New York, NY, USA, ACM.
- Pekalska, E. e Duin, R. P. W. (2002) “Dissimilarity Representations Allow for Building Good Classifiers”, Pattern Recognition Letters, 23(8), p.943–956.
- Savoy, J. (2011) “Who Wrote this Novel? Authorship Attribution across Three Languages – Semantic” Scholar.
- Stamatatos, E. (2009) “A survey of modern authorship attribution methods”, In: Journal of the American Society for Information Science and Technology, 60 (3), p. 538-556.
- Stamatatos, E. Kokkinakis, G. e Fakotakis, N. (2001) “Automatic Text Categorization in Terms of Genre and Author”, In: Computational Linguistics. Vol. 26 (4), p. 471-495.
- Varela, P. J. Justino, E. J. R. Bortolozzi, F. e Albonico, M. (2018) “A Computational Approach for Authorship Attribution on Multiple Languages”, In: 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro.
- Varela, P. J. Justino, E. J. R. Britto, A. e Bortolozzi, F. (2016) “A computational approach for authorship attribution of literary texts using syntactic features”, In: 2016 International Joint Conference on Neural Networks (IJCNN), p. 4835–4842.
- Varela, P. J. Justino, E. J. R. e Oliveira, L. S. (2011) “Selecting syntactic attributes for authorship attribution”, In: The 2011 International Joint Conference on Neural Networks, p. 167–172.
- Zheng, R. Li, J. Huang, Z. e Chen, H. (2006) “A framework for authorship analysis of online messages: Writing-style features and techniques”, In: Journal of the American Soc. Inf. Sci. Technol, vol. 57, 3, p. 378–393.

Geração automática de questões de múltipla escolha

Joyce Martins, Eduardo Ferreira Gehrke

Departamento de Sistemas e Computação
Universidade Regional de Blumenau (FURB) – Blumenau, SC – Brazil
joyce@furb.br, eduardofg17@gmail.com

Abstract. This article describes ChatterEDU, a chatterbot whose knowledge base is composed of multiple choice questions. The question/answer pairs are generated automatically from the morphosyntactic processing of an input text, while distractors - wrong alternatives - are generated by automated searches on virtual pages, based on the correct answer. It is possible to modify the questions before storing the information and starting the bot conversation. The application has generated suitable questions, answers and distractors for texts of any subject, as long as they consist of sentences composed by subject, verb and object. The validation was based on the evaluation of the questions generated, out of which 63% were classified as of good quality.

Resumo. Este artigo descreve o ChatterEDU, um chatterbot cuja base de conhecimento é formada por questões de múltipla escolha. Os pares perguntas/respostas são gerados automaticamente a partir do processamento morfossintático de um texto de entrada, enquanto as alternativas erradas são geradas por meio de buscas automatizadas em páginas virtuais, com base na resposta correta. Além disso, é possível efetuar alterações nas questões antes de persistir as informações e iniciar a conversa com o chatterbot. A aplicação gerou perguntas, respostas e distratores adequados para textos de qualquer tema, desde que formados por frases compostas por sujeito, verbo e objeto. A validação se deu a partir da avaliação das questões geradas, sendo 63% classificadas como de boa qualidade.

1. Introdução

O uso de perguntas/respostas no processo de ensino e aprendizagem, segundo Chi et al. [1994 apud Le; Kojiri; Pinkwart, 2011], é benéfico, uma vez que fazer perguntas estimula a autoexplicação e ajuda a identificar a falta de conhecimento sobre determinado assunto. Neste cenário, cabe destacar também a aplicação de testes com questões de múltipla escolha, que proporcionam a classificação automática de desempenho, podem abranger vários assuntos em um curto espaço de tempo e possibilitem feedbacks praticamente imediatos [Papasalouros, Kanaris e Kotis, 2008]. Entretanto, a formulação manual de questões tende a ser custosa e demandar tempo, sendo observada a importância de ferramentas que automatizem tal processo [Curto, 2010].

Le, Kojiri e Pinkwart [2011] descrevem aplicações educacionais que geram perguntas e respostas automaticamente, tendo como objetivos a aquisição de habilidades, a avaliação do conhecimento ou o diálogo com tutores. Por sua vez, Correia et al. [2010] apresentam uma técnica para gerar distratores para sistemas tutores

voltados ao ensino de um novo idioma. Entre as aplicações computacionais que simulam diálogos com tutores, pode-se citar os *chatterbots*, que são definidos como "agentes inteligentes desenvolvidos para simular uma conversa através da troca de mensagens de texto [...]" [Sganderla; Ferrari; Geyer, 2003]. Dessa forma, *chatterbots* podem ser usados no processo de ensino e aprendizagem para o esclarecimento de dúvidas ou a recapitulação de conteúdos. Assim sendo, este trabalho apresenta uma proposta para gerar automaticamente questões de múltipla escolha para a base de conhecimento de um *chatterbot* educacional.

2. ChatterEDU

O ChatterEDU é um *chatterbot* educacional web cuja funcionalidade é conversar sobre textos da educação básica, escritos em língua portuguesa. O objetivo é promover o autoestudo através da simulação de uma conversa, proporcionando para o estudante uma experiência semelhante a que teria num diálogo com um professor. Nesse contexto, inicialmente o professor deve alimentar a base de conhecimentos do ChatterEDU, informando um texto com sentenças compostas por sujeito, verbo e objeto. O ChatterEDU efetua o processamento do texto, gerando automaticamente pares perguntas/respostas e respectivos distratores. O professor tem a opção de efetuar as alterações necessárias nas perguntas/respostas e distratores gerados, adequando-os aos objetivos de aprendizagem. Feito isso, o estudante pode fazer perguntas ao robô ou responder a questionamentos, digitando a resposta diretamente ou solicitando alternativas de resposta e escolhendo uma entre as opções apresentadas. Por exemplo, para a frase *Blumenau é a terceira maior cidade de Santa Catarina.*, são gerados: (1) a pergunta *Qual a terceira maior cidade de Santa Catarina?*; (2) a resposta longa *Blumenau é a terceira maior cidade de Santa Catarina.*; (4) a resposta curta *Blumenau*; (4) os distratores *Criciúma, Florianópolis, Itajaí e Joinville*. A Figura 1 mostra um diálogo onde o estudante (identificado como VISITANTE) solicita opções de resposta ao ChatterEDU, que avalia a alternativa respondida como correta.

Figura 1. Diálogo entre ChatterEDU e estudante



O processamento do texto de entrada realizado pelo ChatterEDU possui duas etapas, também descritas por Araki et al. [2016], quais sejam: (1) geração das perguntas com as respectivas respostas corretas; (2) geração dos distratores para cada par pergunta/resposta. Al Yahia [2014] elenca três técnicas para geração de perguntas e respostas: baseada em sintaxe, baseada em semântica e baseada em *templates*. No ChatterEDU, para gerar os pares perguntas/respostas, a partir de um texto gramaticalmente correto, com sentenças compostas por sujeito, verbo e objeto, procede-se da seguinte maneira: (1) o texto de entrada é dividido em sentenças; (2) cada sentença

é analisada pelo *parser* Palavras [Bick, 2000], que determina o papel semântico¹ e a classificação morfossintática das palavras; (3) com base nos papéis semânticos, são geradas perguntas (tanto as que podem ser realizadas pelo usuário, quanto pelo robô) e respectivas respostas. O ChatterEDU processa sete papéis semânticos, sendo que para cada um é usado um pronome ou advérbio interrogativo na elaboração da pergunta. Os papéis semânticos processados são: (a) AG: agente (na pergunta usa-se *Quem* para a voz ativa, e *Por quem* para a voz passiva); (b) LOC: lugar (*Onde*); (c) LOC-TMP: localização temporal (dia, mês, ano, indicação de tempo) (*Quando*); (d) ORI-TMP: origem temporal (dia, mês, ano, indicação de tempo) (*Desde quando*); (e) EXT: extensão ou quantidade (*Quanto*); (f) EXT-TMP: período de tempo (*Quanto tempo*); (g) TH: tema (com o verbo ser, usa-se *Qual*, quando o sujeito é pessoa, usa-se *Quem*, nos demais casos, usa-se *O que*). Assim, por exemplo, para a frase *Florianópolis é a capital de Santa Catarina.*, onde *Florianópolis* tem o papel semântico TH, é gerado o seguinte par pergunta/resposta: *Qual é a capital de Santa Catarina?/Florianópolis.*

Uma vez gerados os pares perguntas/respostas, são gerados os elementos distratores, alternativas plausíveis de resposta, porém, incorretas. O método de geração de distratores também pode variar de acordo com a abordagem proposta, podendo, por exemplo, usar ontologias [Al-Yahia, 2014] ou buscá-los no próprio texto de entrada [Curto, 2010]. No ChatterEDU são realizadas buscas automatizadas por elementos distratores em páginas virtuais, utilizando a resposta correta como parâmetro da consulta. Com isso, a fonte para obtenção de distratores é o conteúdo das páginas acessadas. Definiu-se que para cada pergunta são gerados, na maioria dos casos, oito distratores. Todavia, existem situações em que este número pode ser menor por falta de termos alternativos durante as buscas online. A geração de elementos distratores pode ocorrer de duas maneiras: com ou sem numerais.

Quando um numeral é identificado na resposta, são gerados números distratores. Foram definidas regras para números no intervalo de 1000 e 2050, que têm maior probabilidade de corresponderem a anos, e para números fora desse intervalo. Assim, para a frase *O Brasil foi descoberto no ano 1500.*, com o par pergunta/resposta *Quando o Brasil foi descoberto?/No ano de 1500.*, são gerados oito distratores substituindo o termo *1500* por valores aleatórios entre 10 unidades a menos e 10 unidades a mais, ou seja, entre 1490 e 1510. Já para a sentença *O barril de petróleo vale 80 dólares.*, com o par pergunta/resposta *O barril de petróleo vale quanto?/80 dólares.*, o número *80* é substituído na resposta por um décimo (8), um quarto (20), um meio (40), três quartos (60), cinco quartos (100), dobro (160), quádruplo (320) e décuplo (800).

A segunda forma de se obter distratores ocorre quando não são identificados numerais na resposta. Nesse caso, são realizadas buscas automatizadas às páginas virtuais do Dicionário Criativo e da Wikipédia. O processamento inicia com uma conexão por meio da ferramenta JSOUP a cada um dos sites. A partir da resposta (do par pergunta/resposta) enviada como parâmetro da consulta, obtém-se páginas HTML que são processadas para extrair os elementos distratores (das *tags*). O processamento das páginas do Dicionário Criativo busca pelas *tags* `<... title=... class=...`

¹ Papéis semânticos “descrevem a relação semântica subjacente entre um verbo [...] e seus argumentos e são usados para descrever padrões léxicos e semânticos no comportamento dos verbos.” [Kipper, 2005 apud Scarton, 2013, p. xv].

"c_primary_hover...>, selecionado os termos entre title= e class como distratores. Já o processamento das páginas da Wikipédia busca pelo termo /wiki/Categoria, obtendo o link associado a essa tag. Ocorre então uma nova conexão à Wikipédia para obter a página correspondente ao link obtido. Nessa segunda página, são procuradas tags ..., sendo os termos entre e selecionados como distratores. Vale destacar que foram incluídos tratamentos especiais para situações em que a resposta contém mais de uma ocorrência de substantivo, adjetivo ou nome próprio, uma vez que para respostas compostas a geração de distratores é mais elaborada. Nesse caso, é necessário identificar um termo específico na resposta para ser substituído, mantendo o restante inalterado, como no caso da resposta *A capital de Santa Catarina.*, para a qual são buscados termos que substituam apenas *Santa Catarina*. Os distratores são então associados aos respectivos pares perguntas/respostas e a base de conhecimento é gerada. Todas as questões da base de conhecimento são formadas por: uma pergunta, uma resposta longa, uma resposta curta, um ou mais distratores.

3. Resultados preliminares

A geração de distratores foi avaliada por meio de testes efetuados com oito textos de diferentes tamanhos e assuntos. As questões compostas por pergunta, resposta e distratores foram analisadas e classificadas em uma das três categorias: (a) boa: sem necessidade de adaptações na pergunta nem na resposta, eventualmente com adaptações nos distratores, como ajustes de gênero ou número, tendo gerado ao menos um distrator adequado; (b) satisfatória: com necessidade de pequena adaptação na pergunta, incluindo substituição do pronome interrogativo por outro mais adequado, ou eventuais correções na resposta ou nos distratores, tendo gerado pelo menos um distrator adequado; (c) ruim: com necessidade de adaptação mais elaborada na pergunta ou na resposta e, consequentemente, em todos os distratores. A Tabela 1 mostra os resultados obtidos em textos sobre três assuntos: História/Geografia, Informática e Esportes. São listados também os tempos médios de processamento para a geração de perguntas, respostas e distratores.

Tabela 1. Textos analisados

assunto	tamanho do texto	questões geradas	questões boas	questões satisfatórias	questões ruins	tempo médio
T1: História/Geografia	5 frases e 41 palavras	8	6	2	0	33
T2: Informática		8	6	2	0	35
T3: Esportes		10	6	4	0	28
sub-total		26	18	8	0	32
T4: História/Geografia	9 frases e 78 palavras	14	8	6	0	78
T5: Informática		14	8	6	0	68
T6: Esportes		14	9	3	2	59
amostra 2: sub-total		42	25	15	2	68
T7: História/Geografia	16 frases e 160 palavras	24	16	5	3	107
T8: História/Geografia	24 frases e 160 palavras	33	23	8	2	166

Observa-se que: (a) o processamento do *chatterbot* é aplicável para conteúdos de História/Geografia, Informática e Esportes, visto que aproximadamente 63% das perguntas dos textos (T1 a T6) foram avaliadas como boas, 34% como satisfatórias e apenas 3% como ruins; (b) a velocidade média de processamento foi de 32 segundos

para o tratamento de um texto com 41 palavras (T1, T2, T3) e 68 segundos para um texto com 78 palavras (T4, T5, T6); (c) quanto maior o número de frases e palavras processadas, maior o tempo de processamento, visto que, para cada sentença tratada, é necessário processá-la semanticamente, o que exige conexões ao *parser* Palavras, ocorrendo também a posterior busca por distratores com conexões ao Dicionário Criativo e à Wikipédia. Essa afirmação é reforçada por meio dados do processamento de textos apenas de História/Geografia (T7, T8) com o mesmo número de palavras, mas diferente quantidade de frases. Também foi observado que frases longas tendem a prejudicar o tratamento dos papéis semânticos, gerando mais perguntas avaliadas como ruins, e respostas inconsistentes tendem a gerar distratores inconsistentes.

4. Considerações finais

A geração automática de perguntas/respostas a partir de um texto de entrada e de distratores a partir de páginas virtuais, apresenta-se como contribuições desse trabalho, em contrapartida aos *chatterbots* que são baseados em conteúdos fixos ou à geração de distratores baseados no próprio texto de entrada. Em relação à geração de distratores, notou-se que o uso da Wikipédia se apresentou mais apropriado para fornecer distratores a nomes próprios, a exemplo de pessoas e localidades, enquanto que o Dicionário Criativo foi mais assertivo para substantivos em geral. Foi possível perceber que em casos nos quais apenas parte da resposta é extraída para ser substituída, foram comuns erros de concordância na geração dos distratores. Além disso, os distratores são gerados a partir da estrutura das páginas virtuais, o que torna o funcionamento do *chatterbot* sujeito a alterações a partir do momento em que as páginas consultadas apresentarem mudanças em seus conteúdos ou na nomenclatura de *tags*.

Referências

- Al-Yahia, M. (2014), “Ontology-based multiple choice question generation”. *The Scientific World Journal*, [New York], v. 2014, p. 1-9.
- Araki, J. et al. (2016), “Generating questions and multiple-choice answers using semantic analysis of texts”. In: *Proceedings COLING 2016*. Osaka, p. 1125-1136.
- Bick, E. (2000), The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus University Press.
- Correia, R. et al. (2010), “Automatic Generation of Cloze Question Distractors”. In: *Proceedings L2WS*. Tokyo.
- Curto, S. S. L. (2010), Geração automática de testes de escolha múltipla. Mestrado. Universidade Técnica de Lisboa, Lisboa.
- Le, NT.; Kojiri, T. e Pinkwart, N. (2011), “Automatic question generation for educational applications: the state of art”. In: van Do, T.; Thi, H. e Nguyen, N. (eds) *Advanced computational methods for knowledge engineering*. Springer, Berlin, p. 325-338.
- Papasalouras, A.; Kanaris, K. e Kotis, K. (2008), “Automatic generation of multiple choice questions from domain ontologies”. In: *Proceedings IADIS*. Amsterdam, p. 427-434.
- Sganderla, R. B.; Ferrari, D. N.; Geyer, C. F. R. (2003), “BonoBOT: um chatterbot para interação com usuários em um sistema tutor inteligente”. In: *Anais XIV SBIE*. Rio de Janeiro.
- Scarton, C. (2013), VerbNet.Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. Mestrado. USP.

Avaliação das categorias afetivas do dicionário PB-LIWC2015

Flavio Carvalho¹, Gustavo Paiva Guedes¹

¹CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

flavio.carvalho@eic.cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *The Linguistic Inquiry and Word Count (LIWC) is a computer program available for natural language processing, which can use either a standard dictionary or its versions in other languages. There are versions in Portuguese based on the 2007 and 2015 dictionaries. However, there is a lack of studies evaluating the affective categories of the most recent version in Portuguese. So in this work we present an evaluation of the affective category and the emotions subcategories of the 2015 version of the dictionary. Experiments with the dictionary, using the most recent version, show better results in classifying texts and better correlation with the original.*

Resumo. *O Linguistic Inquiry and Word Count (LIWC) é uma programa de computador disponível para o processamento de linguagem natural, que pode usar tanto um dicionário padrão, quanto versões deste em outros idiomas. Existem versões em português baseadas nos dicionários de 2007 e 2015. No entanto, faltam estudos avaliando as categorias afetivas da versão mais recente em português. . Com isso, neste trabalho apresentamos uma avaliação da categoria afetiva e das subcategorias de emoções da versão 2015 do dicionário. Experimentos com o dicionário, utilizando a versão mais recente, mostram melhores resultados na classificação de textos e melhor correlação com o original.*

1. Introdução

Atualmente, na área de Análise de Sentimentos, uma das metodologias disponíveis faz uso do processamento de textos em linguagem natural com o Linguistic Inquiry and Word Count (LIWC), um programa que possui várias versões que foram melhoradas ao longo dos anos. Este programa pode utilizar tanto um arquivo de dicionário padrão, quanto versões do dicionário padrão em outros idiomas. Neste trabalho, chamamos de EN-LIWC2015 a última versão do dicionário padrão em inglês, de 2015.

A página oficial do LIWC¹ disponibiliza dicionários oficiais em diversos idiomas, dentre eles, são encontrados dois dicionários em português. Identificamos neste trabalho estes dicionários como PB-LIWC2007 e PB-LIWC2015. É possível encontrar avaliações dos dicionários em tarefas da área de Análise de Sentimentos [Balage Filho et al. 2013, Carvalho et al. 2019].

Na avaliação do PB-LIWC2007, são analisadas as categorias afetivas, incluindo as subcategorias de emoções positivas e negativas. Entretanto, não foram encontrados estudos que avaliam as categorias afetivas do PB-LIWC2015. Nesse cenário, este trabalho

¹www.liwc.net/dictionaries

tem o objetivo de preencher essa lacuna, realizando análises da categoria de afeto e das subcategorias de emoções do PB-LIWC2015.

Esse trabalho é dividido em mais 4 seções. Na Seção 2, são relacionadas informações sobre o LIWC, detalhando na Seção 3 as versões em português do dicionário. São descritos os experimentos realizados para avaliação e os resultados na Seção 4. A Seção 5 traz considerações sobre os resultados obtidos com o uso do PB-LIWC2015.

2. Linguistic Inquiry and Word Count

As palavras que utilizamos revelam informações sobre nossos estados físicos, mentais e sociais [Pennebaker et al. 2015]. Nos últimos anos, vêm sendo desenvolvidas maneiras para análises quantitativas do comportamento verbal, como a análise textual automática. Na análise quantitativa, se utiliza um programa para analisar as palavras em um arquivo de texto e fornecer resultados objetivos e replicáveis.

O Linguistic Inquiry and Word Count (LIWC) é um programa que considera que diferenças na frequência de categorias de palavras refletem em diferenças individuais do estado emocional ou de cognição. Dado um arquivo de texto, o LIWC executa a contagem de palavras, de acordo com um dicionário contendo categorias de palavras, organizadas em grupos de um domínio específico [Pennebaker et al. 2015]. Para a Língua Portuguesa, existem duas versões do dicionário do LIWC, conforme detalhamos a seguir.

3. Versões do LIWC em português

O PB-LIWC2007 é um dicionário baseado na versão em inglês de 2007 do dicionário padrão do LIWC. Vários Dicionários Bilíngues Português-Inglês foram utilizados na tradução do dicionário de 2007 do inglês para o português [Balage Filho et al. 2013]. Parte do processo de criação do dicionário não passou por revisão da tradução², o que poderia mitigar a ocorrência de alguns dos problemas encontrados no dicionário.

Dentre os problemas do PB-LIWC2007, algumas palavras não estão nas categorias apropriadas, enquanto outras estão incorretamente associadas a algumas categorias [Carvalho et al. 2018]. Outra questão relevante está relacionada à ortografia, sendo possível identificar palavras como ‘ninguén’ e ‘issos’. Também encontramos a inclusão múltipla de palavras com o mesmo radical (ou *stem*) antes de um ‘*’ (asterisco), que além de gerar inconsistências [Carvalho et al. 2018], também aumenta desnecessariamente a quantidade de palavras incluídas no dicionário e, como consequência, aumenta o tempo de análise de textos [Carvalho et al. 2019].

Uma versão mais recente do LIWC em português do Brasil é o PB-LIWC2015 [Carvalho et al. 2019], baseado no EN-LIWC2015. O PB-LIWC2015 foi desenvolvido com o uso de listas de palavras do domínio referente a cada categoria. O PB-LIWC2015 se baseou na versão em inglês, utilizando dicionários monolíngue e bilíngue para criação dos dicionários contendo as diversas categorias do LIWC.

O PB-LIWC2015 tem um total de 2.105 palavras associadas à categoria de processos afetivos. A Tabela 1 apresenta uma comparação da quantidade de palavras das

²<http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc>, conforme acessado em 1 de agosto de 2019.

principais categorias do EN-LIWC2015, PB-LIWC2007 e PB-LIWC2015, com o total de palavras na categoria de processos afetivos (*affect*) e subcategorias (*posemo*, *negemo*).

Tabela 1. Comparação da quantidade de palavras das principais categorias do EN-LIWC2015 (2015), PB-LIWC2007 (2007_pt) e PB-LIWC2015 (2015_pt)

Categoria	Abreviação	Exemplos	2015	2007_pt	2015_pt
Proc. afetivos	affect	Admirável, agonia	1.393	28.475	2.105
Emoção positiva	posemo	Bem-estar, felicidade	620	12.878	863
Emoção negativa	negemo	Incômodo, solidão	744	15.115	1.213

4. Experimentos

Numa primeira etapa da avaliação, foi efetuado um procedimento semelhante ao realizado em trabalhos que apresentaram versões do dicionário em outras línguas [Van Wissen and Boot 2017], para obtenção medidas da correlação entre as versões em português e em inglês. Utilizou-se a coleção bilíngüe Português-Inglês das edições online da revista científica REVISTA PESQUISA FAPESP³ [Aziz and Specia 2011]. Essa coleção foi denominada FAPESP-CORPUS. A análise dos textos em português foi realizada utilizando o PB-LIWC2015 e o PB-LIWC2007 como dicionários, e a análise dos textos equivalentes em inglês foi realizada utilizando o dicionário EN-LIWC2015. O resultado dessa análise são tabelas contendo, nas colunas, valores das frequências relativas de palavras nas categorias dos dicionários, para cada arquivo analisado.

Com esses valores, calculou-se a correlação entre as versões em português e em inglês, utilizando o coeficiente de τ_b de Kendall [Kendall 1938], que avalia as associações estatísticas e não depende de suposições sobre as distribuições [Noether 1981]. Os valores dos coeficientes de correlação τ_b de Kendall para ‘affect’ e suas subcategorias podem ser observados na Tabela 2. Pode-se observar que a análise dos textos em português utilizando o PB-LIWC2015 apresenta uma maior correlação com os valores da análise dos textos em inglês com o EN-LIWC2015.

Tabela 2. Resultados da análise do FAPESP-CORPUS para comparação entre EN-LIWC2015 x PB-LIWC2007 (τ_1) e EN-LIWC2015 x PB-LIWC2015 (τ_2), observando os coeficientes de correlação τ_b de Kendall nos valores das categorias de afeto

Categoria (Abreviação)	τ_1	τ_2
affect	0,46	0,59
posemo	0,34	0,56
negemo	0,55	0,67

Para a avaliação da classificação de polaridade de emoções, foram carregados no LIWC os dicionários PB-LIWC2015 e PB-LIWC2007 para analisar publicações reais de redes sociais, sendo que utilizamos para classificação somente os valores da categoria de afeto (‘affect’) e suas subcategorias. Após o processamento dos textos com cada dicionário, utilizou-se os arquivos gerados pelo LIWC para classificação com cinco algoritmos disponíveis no programa Weka [Hall et al. 2009], escolhendo o conjunto de configurações padrão.

³<http://revistapesquisa.fapesp.br/>

Os algoritmos selecionados foram o *Naive Bayes* (NB) e o *Naive Bayes Multinomial* (NBM), por serem base de referência para classificação de texto [Wang and Manning 2012], e *Random Forest* (RF) e J48, por fornecerem bons resultados na classificação de textos [Fersini et al. 2015, Gabrilovich and Markovitch 2004]. Outro algoritmo escolhido foi o *Logistic Model Tree* (LMT), por ser um dos que alcançam melhores resultados na classificação de textos utilizando recursos estilísticos do português [Aires et al. 2004]. Esse algoritmo apresenta, inclusive, bons resultados em instâncias com o LIWC como um dos recursos utilizados em algumas tarefas (e.g., detecção de sátira, detecção de sarcasmo) [Ravi and Ravi 2017]. Para a obtenção dos resultados no Weka, foi utilizada a técnica de validação cruzada de particionamentos, denominada validação *k-fold* com dez partições para obter a média da medida F, mais conhecida como *F-measure*.

O conjunto de dados utilizado para a tarefa de classificação é o PTSA-800k, que contém aproximadamente 780.000 arquivos. O conjunto está disponibilizado publicamente⁴, contendo subconjuntos de 50.000, 100.000, 200.000, 300.000, 400.000 e 500.000 dados de conteúdo textual, coletados do Twitter e rotulados como ‘positivos’ e ‘negativos’. Foi selecionado, para a tarefa de classificação, o subconjunto com 100.000 entradas, em que 50.000 são negativos e 50.000 são positivos. Os resultados são descritos na Tabela 3.

Tabela 3. Valor de *F-measure* dos algoritmos usados na classificação de polaridade de emoções para o conjunto de dados PTSA-800k, usando categorias de afeto.

	NB	NBM	J48	RF	LMT
PB-LIWC2007	0,528	0,554	0,545	0,558	0,548
PB-LIWC2015	0,716	0,835	0,921	0,933	0,940

5. Considerações

Este trabalho traz análises da categoria de afeto e as subcategorias de emoções do PB-LIWC2015 por meio de experimentos executando tarefas de classificação de textos provenientes de redes sociais. Também foi incluída uma comparação estatística de diferentes versões em português dos dicionários do LIWC. Foram utilizados conjuntos de textos que estão disponibilizados publicamente, facilitando a replicação dos experimentos e comparações com outros recursos em português, a serem realizadas no futuro.

Observa-se, com os resultados encontrados neste estudo, que um grande número de palavras em um dicionário pode não representar um impacto positivo no número de palavras a serem contadas nos textos. Os resultados deste trabalho mostram que, em Análise de Sentimentos, é possível obter melhores resultados nas tarefas de classificação com um número reduzido de palavras, desde que haja uma boa correspondência com o domínio que se deseja investigar. No caso específico, obtivemos um resultado 68,5% melhor na classificação de polaridade de emoções utilizando as categorias do dicionário PB-LIWC2015, que possui quantidade equivalente a apenas 7,4% do total de palavras nas categorias correspondentes do PB-LIWC2007.

⁴PTSA-800k pode ser obtido em <https://www.kaggle.com/augustop/portuguese-tweets-for-sentiment-analysis>

Referências

- Aires, R., Manfrin, A., Aluísio, S., and Santos, D. (2004). Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs?
- Aziz, W. and Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL)*, Cuiabá, Brasil. Sociedade Brasileira de Computação.
- Balage Filho, P. P., Pardo, T. A., and Aluísio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 215–219.
- Carvalho, F., Rodrigues, R. G., Santos, G. d., Cruz, P., Ferrari, L., and Guedes, G. P. (2019). Evaluating the 2015 Brazilian Portuguese LIWC lexicon with sentiment analysis in social networks. In *CSBC 2019 - 8º BraSNAM*, Belém, Brazil.
- Carvalho, F., Santos, G. d., and Guedes, G. P. (2018). AffectPT-br: an affective lexicon based on LIWC 2015. In *37th International Conference of the Chilean Computer Science Society (SCCC 2018)*, Santiago, Chile. IEEE.
- Fersini, E., Pozzi, F. A., and Messina, E. (2015). Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–8. IEEE.
- Gabrilovich, E. and Markovitch, S. (2004). Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In *Proceedings of the twenty-first international conference on Machine learning*, page 41. ACM.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
- Noether, G. E. (1981). Why Kendall Tau? *Teaching Statistics*, 3(2):41–43.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report, University of Texas.
- Ravi, K. and Ravi, V. (2017). A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*, 120:15–33.
- Van Wissen, L. and Boot, P. (2017). An electronic translation of the LIWC Dictionary into Dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 703–715. Lexical Computing.
- Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

“Sentimento de quê?”: uma lista de sentimentos para a Análise de Sentimentos

Barbara C. Ramos¹, Cláudia Freitas²

¹ PPGEL, PUC-Rio

Rua Marquês de São Vicente, 225 Rio de Janeiro, Brasil

² PPGEL, PUC-Rio

Rua Marquês de São Vicente, 225 Rio de Janeiro, Brasil

barbaracmramos@gmail.com, claudiafreitas@puc-rio.com

Abstract. *The present article describes the process of a manually built sentiment lexicon through large corpora. The research was conducted based on the lexical-grammatical pattern “sentimento de N” and the results were distributed within four groups. As a result, we listed over seven hundred words which are considered types of sentiment by Portuguese speakers. The aim of this lexicon is to carry out automatic detection of emotions and sentiments in large collections of electronic texts.*

Resumo. *Este artigo descreve a construção manual de um léxico de emoções elaborado a partir da exploração em grandes corpora. Realizamos buscas com o padrão léxico-gramatical “sentimento de N”, e distribuímos os resultados em quatro grupos. Como resultado final, elencamos mais de setecentas palavras que são classificadas como tipos de sentimento por falantes de português. O objetivo desse léxico é servir como subsídio para a detecção automática de emoções e sentimentos em grandes coleções de texto.*

1. Introdução

A área de Análise de Sentimento tem se mostrado de grande popularidade no ambiente de PLN. Se, originalmente, a maioria dos trabalhos se dedicava a encontrar sentimentos e opiniões associados a resenhas de produtos ou personagens políticos, em ambientes como lojas virtuais, tweets e comentários em redes sociais (veja-se, por exemplo, [Carvalho et al. 2011] e [Freitas et al. 2014] para trabalhos voltados para a língua portuguesa), mais recentemente as Humanidades Digitais também têm se interessado pelo tema, buscando detectar, em quantidade e qualidade, sentimentos e emoções em personagens e obras literárias. [Klinger et al. 2016], por exemplo, descrevem como as emoções se manifestam ao longo de duas obras de Kafka, bem como a maneira pela qual as personagens dessas obras se caracterizam em termos emocionais. [Heuser et al. 2016], por sua vez, tentam produzir uma cartografia afetiva de Londres, associando emoções e espaços geográficos em obras da literatura inglesa.

Uma abordagem frequente para resolver a tarefa é a utilização de léxicos, que, em geral, são de dois tipos: léxicos de palavras de emoções e sentimentos; e léxico de polaridades, no qual palavras e expressões, não necessariamente palavras de emoções, são classificadas de acordo com suas polaridades (positiva, negativa ou neutra).

Os léxicos de palavras de emoções e sentimentos são claramente inspirados em propostas de emoções fundamentais. Nestes casos, postula-se a existência de um certo número de emoções básicas, fundamentais ou universais, comuns a todas as culturas. Nesta abordagem encontram-se, por exemplo, as propostas de [Plutchik 1962, 2001], que elenca oito emoções; [Tomkins 1962, 1963], que elenca nove; [Ekman 1999], que elenca sete tipos básicos de emoção e, por fim, o modelo OCC [Ortony et al. 1988], que propõe 22 tipos de emoção.

No entanto, a dificuldade quanto a uma definição precisa e/ou consensual do que seja emoção e sentimento (e a divergência quanto à quantidade de emoções ilustra o ponto) vem desde antes de Cristo. Como nos lembram [Maia e Santos 2018], o que chamamos de amor era nomeado pelos gregos de diferentes maneiras: ágape, éros, ludus, pragma, philia, storge. A discussão em torno das emoções está presente, até hoje, em áreas diversas como psicologia, antropologia, filosofia e estudos da linguagem.

De uma perspectiva antropológica, é frequente uma abordagem para as emoções e sentimentos diametralmente oposta à ideia de emoções básicas: não há como se falar em universalidade, pelo contrário. [Rezende e Coelho 2010], por exemplo, argumentam que “os sentimentos são tributários das relações sociais e do contexto cultural em que emergem” (2010:11) e que na visão das ciências sociais, “as emoções, embora situadas no corpo, têm com este uma relação que é permeada sempre por significados culturalmente e historicamente construídos” (2010:33). Esta abordagem também é compartilhada pelo antropólogo, sociólogo e psicólogo David Le Breton. Em palestra sobre as emoções, em 2019, o teórico defende que as diferenças de culturas afetivas se marcam pela existência de emoções ou de sentimentos que não são confortavelmente traduzíveis em outras línguas sem possíveis erros grosseiros de interpretação. Le Breton sugere, inclusive, que seria conveniente colocar aspas em cada uso de um termo emocional para traduzir o fato de que ele não se estende de um significado próprio para outro. O perigo da tradução denota as diferenças de sentimento e de expressão de uma sociedade e de uma época. Para ele, as emoções são modos de afiliação a uma comunidade social; uma maneira de se reconhecer e de poder comunicar junto, a partir de um fundo afetivo. [Maia e Santos 2018] vão pelo mesmo caminho. Reconhecendo que a discussão dos conceitos de emoção é extensa e pouco consensual, as autoras decidiram focar seu estudo em buscar entender o motivo pelo qual essas tentativas de conceitualizar emoções podem ser tão controversas e enganosas. Para elas, também, a tradução de palavras relacionadas à emoção (*emotion*, *sensation*, *sentiment*, *feeling* e *mood*) geralmente não comporta todo o significado da palavra em sua língua fonte, pois podem ter diferentes conotações culturais nas culturas fonte e alvo. No contexto dos estudos da linguagem, [Wierzbicka 1999] critica a ontologia proposta por [Ortony et al. 1988] (e, por extensão, as demais propostas de emoções básicas) por ser etnocêntrica e focada na língua inglesa. A autora acredita, ainda, que a palavra *emoção* pode estar muito engendrada à nossa língua nativa e às línguas predominantes no cenário acadêmico. Wierzbicka opta por adotar a palavra *emotion* em seu livro, porém não como uma ferramenta neutra e livre de problemas, mas como sinônimo de “sentimentos baseados em pensamentos” (1999:12). Wierzbicka enfatiza que não é possível discutir emoção sem usar palavras, por isso é fundamental estudarmos como elas são usadas em diferentes línguas e culturas. Por fim, [Maia e Santos 2018] alegam que, mesmo com todo o trabalho, identificação e hipóteses já alcançados por psicólogos e lexicólogos, ainda não há informação suficiente para dar conta do léxico de emoções.

É com essa segunda perspectiva, que nega a universalidade das emoções, que nos alinhamos, e por isso desconfiamos de abordagens baseadas na ideia de emoções básicas, por um lado, e da tranquilidade relativa à sua tradução, por outro. Interessa-nos detectar, automaticamente, sentimentos e emoções em grandes coleções de textos, e para isso acreditamos que um léxico é um componente crucial. Como, então, prosseguir, diante de tantas incertezas? Nossa proposta para a elaboração de um léxico de sentimentos e emoções se desvia da discussão apresentada, deslocando para os falantes de uma língua a responsabilidade sobre a identificação de algo como sendo uma emoção ou sentimento. Especificamente, tiramos proveito da disponibilidade de grandes corpora anotados para perguntar, à própria língua, o que consideramos *sentimento*. Em outras palavras: fizemos uma ampla varredura nos corpora do projeto AC/DC [Santos e Bick 2000], atualmente com mais de um bilhão de palavras, utilizando o padrão léxico-gramatical “sentimento de N”. Como resultado, compilamos uma lista de 742 palavras de sentimento. Neste trabalho, detalhamos o procedimento de busca e de análise, e apresentamos dados relevantes para a Análise de Sentimento em português.

2. Análise de sentimento em português

A dificuldade de lidar com as emoções talvez explique a maior quantidade de léxico de polaridades do que de palavras de emoções ou sentimentos. No que se refere à língua portuguesa, listamos a seguir alguns dos recursos usados para a tarefa de Análise de Sentimento.

O LIWC [Tausczik e Pennebaker 2010] é um software para análise de textos, baseado em um léxico que classifica as palavras de acordo com categorias “psicologicamente significativas”. As palavras se distribuem em 4 dimensões, e uma delas, “Psychological Processes”, contém a categoria “Affective Processes”, que por sua vez compreende as subcategorias *positive emotions* (*happy, pretty, good*), *negative emotions* (*hate, worthless, enemy*), *anxiety* (*nervous, afraid, tense*), *anger* (*hate, kill, pissed*) e *sadness* (*grief, cry, sad*). A existência de uma dimensão afetiva torna o recurso interessante para a Análise de Sentimento, e nesse contexto foi feita uma avaliação do LIWC desenvolvido para a língua portuguesa [Balage et al. 2013]. Na versão em português, o *Brazilian Portuguese LIWC 2007 Dictionary* 1 foi construído por três equipes que, com o auxílio de dicionários bilíngues, inseriram de forma automática categorias preexistentes do LIWC. De acordo com as informações na página de apresentação do LIWC em português, a parcela de trabalho manual de tradução não foi revisada. Em uma comparação com outros recursos do português (OpLexicon e SentiLex), o desempenho do LIWC ficou na média dos demais recursos. O OpLexicon [Souza et al. 2012], como o nome sugere, é um léxico de opinião. Idealizado especialmente para a área de *sentiment analysis*, não contém palavras de emoção, mas termos variados associados a polaridades. O SentiLex [Carvalho e Silva 2015], idealizado para o português de Portugal, vai pelo mesmo caminho, mas tem como foco predicadores humanos, sendo composto por palavras variadas que se referem a pessoas, com a indicação da polaridade atribuída tanto à pessoa que predica quanto ao alvo da predicação. A proposta do léxico Multilingualsentiment [Chen e Skiena 2014] é agregadora: elaborar um léxico internacional de sentimentos, que atualmente comporta mais de 80 línguas. No

¹ Informações disponíveis em: <http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>.

entanto, assim como os anteriores, contém uma lista de palavras não necessariamente associadas a emoção, mas distribuídas conforme a polaridade.

O trabalho de [Santos et al. 2014] tem como alvo emoções, e não polaridades, se alinhando com nossos interesses. No entanto, para abordar a tarefa de Mineração de Emoções, as autoras tomam uma perspectiva multilíngue, se propondo a dar alguns passos iniciais na área da Mineração de Emoções Multilíngue. Para tanto, fazem uso intenso da tradução (automática): tradução do léxico de sentimentos NRC (*word-emotion association*) [Mohammad e Turney 2013] e tradução de corpus, e nisso nos distanciamos. Como conclusão, as autoras apontam o grande desafio da tarefa, evidenciada, por exemplo, na baixa concordância entre os anotadores (média de 55%) quando instados a anotar o corpus de acordo com as oito emoções do dicionário NRC, que por sua vez se inspira nas já mencionadas oito emoções de Plutchik.

A dificuldade na concordância relativa à identificação de emoções já havia sido notada em [Wiebe et al. 2005]. Nos já citados trabalhos de Análise de Sentimento na literatura, não é diferente: em [Klinger et al. 2016], considerando a atribuição de emoção a 300 palavras, em apenas 46% dos casos houve concordância, dos três anotadores. Em [Heuser et al. 2016], a divergência foi tanta que as opções de emoções para anotação precisaram ser reduzidas para os opostos “medo” e “felicidade”. Ao construirmos nosso léxico a partir do padrão “sentimento de N”, evitamos também a baixa concordância entre anotadores.

3. Metodologia

Para elaborarmos o léxico, partimos da expressão de busca “sentimento de N”² no corpus OBras [Santos et al. 2018], por meio da interface AC/DC, criada e mantida pela Linguateca³. Trata-se de uma abordagem simples, inspirada nos padrões de [Hearst 1992], mas que nos assegura que a palavra em questão está sendo entendida, pelo falante, como uma palavra de sentimento, e dessa maneira nos desviamos da polêmica sobre o que é um sentimento na língua. A escolha do OBras se deveu a dois principais motivos: trata-se de um corpus de obras literárias, e partimos do princípio de que a literatura seria um espaço propício para procurar emoções, e o tamanho (5.7 milhões palavras). No entanto, diferentemente do que esperávamos, a busca não forneceu diretamente tipos de sentimento. Após a análise de diversos casos, distribuímos os lemas encontrados em quatro grupos:

- **Grupo 1:** O “N” corresponde a sentimentos que consideramos convencionais, como *culpa*, *medo* etc.
- **Grupo 2:** O “N” corresponde a maneiras não convencionais de falar sobre sentimentos.
- **Grupo 3:** O “N” não se refere a um sentimento, mas ao possuidor do sentimento.
- **Grupo 4:** O “N” introduz um modificador do sentimento. Isto é, o “N” não nomeia um sentimento, apenas atribui a ele qualidades.

O Quadro 1 apresenta exemplos do corpus para cada um dos grupos mencionados.

² Especificamente, a expressão de busca utilizada foi [lema="sentimento"] [lema="de"] [pos!="V"]* @ [pos="N" & func="P<"] within s

³ <http://www.linguateca.pt/ACDC>

Quadro 1. Exemplos de tipos de sentimento conforme os grupos de análise

Grupo 1	<i>id="Iaiá_Garcia Prosa:romance MdA 1878 romantismo"</i> : Luís Garcia não pôde furtar-se a um sentimento de pena , ao vê-lo entrar fardado e prestes a seguir para o Sul
Grupo 2	<i>id="A_semana Prosa:crônica MdA 1892 "</i> : Atentai, mais que tudo, para esse sentimento de unidade nacional , que a política pode alterar ou afrouxar, mas que a arte afirma e confirma, sem restrição de espécie alguma, sem desacordos, sem contrastes de opinião
Grupo 3	<i>id="O_Matuto Prosa:romance FT 1878 realismo_ regionalismo_romantismo"</i> : ‘Tais eram as idéias e os sentimentos de d. Damiana
Grupo 4	<i>id="A_Alma_Encantadora_das_Ruas Prosa:crônica JdR 1807 "</i> : Esse sentimento de natureza toda íntima não vos seria revelado por mim se não julgasse, e razões não tivesse para julgar, que este amor assim absoluto e assim exagerado é partilhado por todos vós

Foram encontrados 120 lemas diferentes com essa abordagem. Para um léxico de sentimentos, nosso interesse está nos grupos 1 e 2. Ou seja, por mais que seja difícil ou discutível, em um primeiro momento, decidir se estamos diante de uma maneira convencional ou não de mencionar um sentimento, essa discussão não é, ao menos por enquanto, a discussão que nos interessa. Nosso foco está em descartar da lista aquilo que claramente não se refere a um tipo de sentimento. Assim, uma outra maneira de olhar para os grupos é vê-los simplesmente como *sentimento* (1 e 2) vs *não-sentimento* (3 e 4).

Em seguida, após termos clareza sobre a maneira de lidar com os resultados inesperados, repetimos o procedimento, dessa vez tomando como objeto todos os corpora disponíveis no projeto AC/DC – note-se que, com isso, estamos lidando com aquilo que é materializado como sentimento na língua portuguesa nas variantes brasileira, portuguesa e moçambicana, e em diferentes recortes temporais.

A busca em todos os corpora (excluindo-se o OBras) do AC/DC resultou em 2060 lema diferentes. O ponto de corte da análise dos lemas foi a frequência acima de 3, o que diminuiu o número de lemas para 853. Alguns lemas foram excluídos por serem erros de digitação ou pré-processamento (por exemplo, “germanidade” também aparecia como “germanidade6”) e 86 lemas já haviam sido analisados na etapa do corpus OBras. Além dessas exclusões, também agrupamos oito lemas por representarem apenas variações ortográficas (por exemplo “afeto” e “afecto”), com isso, ficamos com 724 lemas para a análise.

De maneira completamente independente de nossa análise, todo o material do AC/DC já havia passado por uma anotação semiautomática do campo semântico das emoções [Mota e Santos, 2015]. A anotação foi baseada em pistas lexicais obtidas automaticamente a partir de palavras disparadoras (*seed words*) retiradas de recursos lexicais variados, e então revista. Graças a essa camada anterior de anotação, podemos também comparar a eficácia de nossa abordagem (e indiretamente, podemos avaliar a anotação semiautomática já presente no AC/DC).

4. Resultados

Os resultados estão na Tabela 1. Indicamos com sema = emo todos os lemas que já estavam anotados, no AC/DC, como pertencentes ao campo semântico da emoção. Do mesmo modo, sema ≠ emo indica as palavras que nós consideramos palavras de sentimento convencionais (Grupo 1) ou pouco convencionais (Grupo 2), mas que, no AC/DC, não foram anotadas como tal.

Tabela 1: Resultados da análise levando com conta a anotação do AC/DC

		Corpus		
		OBras	Todos exceto OBras	Todos
Grupo 1	sema = emo	33 (62%)	76 (35%)	109 (40%)
	sema ≠ emo	20 (38%)	141 (65%)	161 (60%)
Grupo 2	sema = emo	0%	26 (6,1%)	26 (5,5%)
	sema ≠ emo	49 (100%)	398 (93,9%)	446 (94,5%)
Grupo 3		11	59	70
Grupo 4		07	24	31

Consideramos os resultados de nossa abordagem muito positivos: elencamos, no total, 742 palavras de sentimento, partindo de uma lista total de 843 palavras. Ou seja, o padrão “sentimento de N” é confiável, levando a 90% de palavras de sentimento. Quando compararmos com a anotação do AC/DC, vemos que, no corpus OBras, apenas 62% das palavras que consideramos palavras de sentimentos convencionais já estavam identificadas como palavras do campo semântico das emoções/sentimentos. Quando vamos para todo o material, apenas 40% de tudo o que consideramos sentimentos convencionais já continha anotação de emoção. Alguns exemplos de palavras do grupo de sentimentos pouco convencionais que encontramos e que estavam com sema = emo são *autoestima, decoro e heroísmo*.

Uma explicação para essa divergência está na própria anotação do AC/DC. Expressões do tipo “sentimento de N” contém a indicação de um sema emoção genérica (emo:gen) na própria palavra “sentimento”, e o substantivo especificador do sentimento, por outro lado, não recebe sema algum. Além disso, mas de maneira não surpreendente, no OBras, nenhuma das palavras que consideramos palavras de sentimentos pouco convencionais já estava anotada como palavra do campo semântico das emoções/sentimentos. Por outro lado, quando considerarmos o AC/DC completo, temos que 5% das palavras de sentimento que consideramos pouco convencionais já estavam anotadas como pertencente ao campo semântico das emoções, o que sugere que fomos rigorosas com a ideia de convencionalidade.

5. Discussão

Ao longo da análise dos lemas, nos chamou a atenção a distribuição das palavras de sentimento por campos lexicais pouco mencionados na literatura: (1) palavras de sentimento relacionadas à ideia de *pátria*; (2) palavras de sentimento relacionadas à *religião* e *espiritualidade*; (3) palavras do campo semântico de família e parentesco usadas para expressar sentimento. Os referidos lemas estão no Quadro 1. Destes, apenas “ufania” já estava anotado semanticamente no AC/DC, com “sema=“emo:orgulho””.

A distribuição dos lemas por campos lexicais promove reflexões interessantes. Grande parte das palavras de sentimento usadas nos corpora estão contidas outros campos lexicais que não o das emoções – nacionalismo, espiritualidade e parentesco. Além disso, lemas como “brasilidade”, “patriotismo” e “pertencimento”, que aparecem no Quadro 1, podem carregar mais de um sentimento, como por exemplo “orgulho”, “amor”, “(in)satisfação” e “(in)felicidade”.

Quadro 1: Campos lexicais identificados no léxico de emoções

Campo lexical 1: lemas referentes ao nacionalismo. Sentimento de....

americanismo – anticomunismo – antilhanía – anti-lusitanismo – bairrismo – brasilidade – cidadania – civismo – democracia – germanidade – germanismo – identidade – italianidade – mineiridade – nação – nativismo – nacionalização – nacionalidade – nacionalismo – origem – pátria – patriota – patriotismo – pertença – pertencimento – povo – raça – ufanía – ufanismo – união – xenofobia.

Exemplos:

“É que o **sentimento de brasilidade**, de patriotismo, nato do brasileiro, impera”

“Eu fui ao diante dele, afirmado que a adoção de uma nacionalidade é ato político, e muita vez pode ser dever humano, que não faz perder o **sentimento de origem** nem a memória do berço”

“Essa instituição por ele pregada e que se fez realidade entre nós precisa ser conservada sempre, não só por ser ela a defesa do território, como ainda por servir de meio eficaz e rápido para civilizar o caipira, dando-lhe o **sentimento de pátria** e ensinando-lhe amar e compreender as cores e a história da nossa Bandeira”

Campo lexical 2: lemas referentes à espiritualidade. Sentimento de...

abstenção – adoração – católico – cobiça – comunhão – devocão – doação – fanatismo – fé – idolatria – perdão – religião – religiosidade

Exemplos:

“Helena deixou-se cativar desse **sentimento de abstenção** e elevação; se alguma dor ou remorso a pungia, esqueceu-os, por um minuto ao menos, entre aquelas paredes desataviadas, diante de um padre, entre uma imagem de Jesus e as obras vivas do Criador”

“Se escrevo sobre o médium, não é por nenhum **sentimento de idolatria**, mas por reconhecimento ao trabalho de um companheiro que abdicou de si mesmo para servir a causa que abraçou”

Campo lexical 3: lemas referentes ao parentesco. Sentimento de...

avó – irmandade – irmão – mãe – maternidade – orfandade – pai – paternidade – viúvo

Exemplos:

“Ao contrário do que se pensa, o sentimento de maternidade não é natural, mas sim uma construção social e cultural, como já foi mostrado nos capítulos anteriores”

“Carlos, o marido de Edemeia é a personagem que completa junto com Carvalho, pai da protagonista, o triângulo da contenda e reproduz a figura do ser angustiado e dividido entre a manutenção das funções de vigor e poder reservadas ao sexo masculino, e o sentimento de pai e de esposo”

6. Considerações finais e desafios futuros

A literatura sobre o tema emoção e sentimento é abundante de discussões e de propostas sobre quais são os sentimentos humanos. No entanto, para realizar a análise automática, precisamos elencar de forma razoavelmente segura palavras usadas para descrever emoções e sentimentos – e, no recorte deste trabalho, na língua portuguesa. Para fugir das discussões acerca do que é um sentimento, nos apoiamos no próprio corpus, isto é, na maneira como falantes (de português) conceitualizam e verbalizam coisas como *sentimento*. Consideramos a abordagem bem-sucedida, sendo capaz de elencar um total 742 palavras de sentimento, várias delas de difícil associação imediata ao campo dos sentimentos. Por outro lado, cerca de mil lemas com três ou menos ocorrências não foram analisados, e pretendemos desenvolver estratégias para lidar com essa cauda longa.

O próximo desafio está em organizar esta lista de palavras para que, em seguida, possamos adicioná-la à anotação do AC/DC. A anotação semântica de emoção no AC/DC atualmente conta com 24 clusters. Isto é, as palavras do campo semântico das emoções e sentimentos, no AC/DC, se distribuem pelos seguintes grupos: *alívio; admirar; amor; coragem; desejo; desespero; esperança; felicidade; fúria; genérica; gratidão; humildade; infelicidade; ingratidão; insatisfação; inveja; medo; ódio; orgulho; pena; satisfação; saudade; surpresa e vergonha*.

Assim, nos interessa saber *se e como* as palavras que encontramos se encaixam nos clusters existentes e, de maneira complementar, critérios capazes de justificar a criação de novos clusters. Como resultado, teremos uma infraestrutura mais poderosa, na língua portuguesa, para investigar correlações entre sentimentos e outros campos semânticos, como os já mencionados trabalhos de [Klinger et al. 2016] e [Heuser et al. 2016].

Agradecimentos

Agradecemos à Diana Santos e aos pareceristas anônimos pelas valiosas contribuições que, certamente, deixaram o artigo mais interessante.

Referências bibliográficas

- Balage Filho, P.P., Aluísio, S.M. e Pardo, T.A.S. (2013) An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology – STIL*, pages 215–219.
- Carvalho, P., Sarmento, L., Teixeira, J. e Silva, M. J. (2011) Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. In *ACL (Short Papers)*, pages 564–568.
- Carvalho, P., Silva, M. (2015) Sentilex-pt: principais características e potencialidades. In *Linguística, Informática e Tradução: Mundos que se Cruzam*, Organizado por Simões, Barreiro, Santos, Sousa-Silva & Tagnin, *Oslo Studies in Language* 7(1):425–438.
- Chen, Y. e Skiena, S. (2014) Building Sentiment Lexicons for All Major Languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, pages 383–389.
- David Le Breton (2019) *Experiências da dor: uma antropologia*, Palestra ministrada por David le Breton, da Universidade de Estrasburgo, em 13 de março de 2019, Instituto de Estudos Avançados em Humanidades, PUC-Rio.
- Ekman, P. (1999) *Basic Emotions, Handbook of Cognition and Emotion*. Editado por Tim Dalgleish e Mick Power, John Wiley & Sons, Sussex, UK.
- Freitas, C., Motta, E., Milidiú, R. L. e César, J. (2014) Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus. In *New Language Technologies and Linguistic Research: A Two-Way Road*. Sandra Aluísio e Stella E. O. Tagnin, Cambridge Scholars Publishing, 2014, pages 128–146.
- Hearst, M. (1992) Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.
- Heuser, R.; Moretti, F.; Steiner, E. (2016) The emotions of London. In *Literary Lab Pamphlets*, <https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf>, Março.
- Klinger, R., Samat, S. S. e Reiter, N. (2016) Automatic Emotion Detection for Quantitative Literary Studies: A case study based on Franz Kafka’s “Das Schloss” and “Amerika”. In *Digital Humanities 2016: Conference Abstracts*, pages 826–828, Cracóvia, Polônia.
- Maia, B.; Santos, D. (2018) Language, emotion and the emotions: The multidisciplinary and linguistic background. In *Lang Linguist Compass*.
- Mohammad, S. M., e Turney, P. D. (2013) Crowdsourcing a word-emotion association lexicon. In *Computational Intelligence* 29(3):436–465.
- Mota, C. e Santos, D. (2015) Emotions in natural language: a broad-coverage perspective. In Linguateca, <https://www.linguateca.pt/acesso/EmotionsBC.pdf>, Maio.
- Ortony, A., Clore, G. L., e Collins, A. (1988) *The cognitive structure of emotions*. Cambridge and New York: Cambridge University Press.
- Plutchik, R. (1962) *The Emotions: Facts, Theories, and a New Model*. Random House Inc, EUA.

- Plutchik, R. (2001) “The Nature of Emotions”, *American Scientist* 89:344–350.
- Rezende, C.; Coelho, M. (2010) *Antropologia das Emoções*, Editora FGV, Rio de Janeiro.
- Santos, A. G. L., Becker, K. e Moreira, V. (2014) “Um estudo de caso de mineração de emoções em textos multilíngues”, *III Brazilian Workshop on Social Network Analysis and Mining* (BraSNAM 2014).
- Santos, D. e Bick, E. (2000) “Providing Internet access to Portuguese corpora: the AC/DC project”, *Proceedings of the Second International Conference on Language Resources and Evaluation*, Edited by Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis e Gregory Stainhauer, LREC 2000, pages 205–210.
- Santos, D., Freitas, C. e Bick, E. (2018) OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain. In *OpenCor*, Canela, RGS, Brasil.
- Souza, M.; Vieira, R.; Busetti, D.; Chishman, R. E Alves, I. M. (2012) Construction of a Portuguese Opinion Lexicon from multiple resources. In *8th Brazilian Symposium in Information and Human Language Technology*.
- Tausczik, Y., Pennebaker, J. (2010) The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. In *Journal of Language and Social Psychology*, 29(1), pages 24–54.
- Tomkins, S. S. (1962) *Affect Imagery Consciousness: Volume I, The Positive Affects*, London: Tavistock Publications.
- Tomkins, S. S. (1963) *Affect Imagery Consciousness: Volume II, The Negative Affects*, London: Tavistock Publications.
- Wiebe, J., Wilson, T., e Cardie, C. (2005) Annotating expressions of opinions and emotions in language. In *Language resources and evaluation*, 39(2-3), pages 65–210.
- Wierzbicka, A. (1999) *Emotions Across Languages and Cultures: Diversity and Universals*, pages 1–48, Inglaterra: Cambridge University Press.

Estudando personagens na literatura lusófona

Diana Santos¹, Cláudia Freitas²

¹Linguateca & ILOS, Universidade de Oslo
Pb 1003 Blindern, 0315 Oslo, Noruega

²Linguateca & PPGEL, PUC-Rio
Rua Marquês de São Vicente, 225 Rio de Janeiro, Brasil

d.s.m.santos@ilos.uio.no, claudiafreitas@puc-rio.br

Abstract. In this paper we describe some studies and tools to deal with literary characters in Portuguese. After briefly describing the framework and the tools employed, we present character networks for some novels, and some preliminary attempts to characterize them.

Resumo. Neste artigo descrevemos alguns estudos feitos sobre personagens literárias em português. Após referir a infra-estrutura usada e as ferramentas utilizadas, apresentamos redes de personagens de algumas obras, e algumas tentativas de caracterização das mesmas.

1. Introdução

Vários autores têm estudado as personagens de obras literárias no âmbito da leitura a distância. Assim, [Moretti 2011] criou redes de personagens do *Hamlet*, e [Grayson et al. 2016] fizeram o mesmo a romances chave da língua inglesa, de Charles Dickens e Jane Austen. [Klinger et al. 2016] estudaram por seu lado as personagens de duas obras de Kafka em termos de assinaturas emocionais para sete diferentes emoções, e [Bonch-Osmolovskaya and Skorinkin 2017] caracterizaram as personagens de Tolstoi em termos de papéis semânticos como agente, paciente, experienciador, possuidor, etc.

Pensamos que este é o primeiro artigo que relata a construção de redes de personagens de obras em português, embora já nos tenhamos debruçado sobre a caracterização de pessoas em textos literários, nomeadamente adjetivos associados a género em [Santos et al. 2018b].

2. O contexto

Este trabalho é feito no âmbito da Literateca [Santos 2019], um ambiente para estudar obras literárias em português usando a infraestrutura de corpos e anotação da Linguateca, mais especificamente o projeto AC/DC, que usa o PALAVRAS [Bick 2000] e outros mecanismos de anotação, e que contém, na sua versão 1.2 de 10 de julho de 2019, 730 obras completas de 167 autores diferentes.¹.

¹<https://www.linguateca.pt/acesso/corpus.php?corpus=LITERATECA>

3. Pré-processamento: identificação das personagens

Uma personagem é geralmente referida de formas muito variadas numa obra. Por exemplo, a personagem Clara n’*As Pupilas do Senhor Reitor* de Júlio Dinis é também denominada pelos seguintes nomes próprios (ou entidades mencionadas): *Clarinha* e *Clarita*. E a personagem que dá o nome ao romance epônimo de Machado de Assis, *Helena*, também é referida por *Nhanhã Helena*, *D. Helena do Vale* ou apenas *D. Helena*. Um caso ainda mais difícil de obter automaticamente é a personagem de *Dom Casmurro* que aparece sob os nomes de *João*, *Pádua*, *Joãozinho*, *Sr. Pádua* e *Tartaruga*...

Por isso, uma das primeiras tarefas num estudo sistemático de personagens em obras literárias é identificar o conjunto dos nomes próprios que correspondem a uma mesma personagem, e definir uma maneira única de referi-las. [Grayson et al. 2016], por exemplo, criaram um dicionário de personagens, com uma única entrada para cada personagem e todos os seus nomes alternativos associados. Mas, além disso, é preciso garantir que um dado nome se refere à personagem certa.

Por outro lado, nem sempre é usado o nome próprio para referir uma personagem. Para terem uma identificação mais fiável das personagens em alemão, [Krug et al. 2018] marcaram todas as referências, que podem ser um nome comum, um pronome, e, no caso do português mesmo nada, visto que a língua portuguesa é uma língua de sujeito nulo: Numa investigação sobre a omissão do sujeito em português, considerando-se todos os romances, contos e crônicas de Machado de Assis, verificamos que quase 30% das frases têm o chamado *sujeito oculto* no verbo da oração principal – os números sobem para 40% considerando verbetes biográficos de uma enciclopédia e, num corpus jornalístico, 16% das frases têm sujeito oculto [Freitas et al. 2019]. Deste modo, em português, fica claro que há muito a perder quando não entramos em conta com essa questão.

4. População de um romance

Apesar de os trabalhos sobre personagens literárias tematizarem aquelas *principais*, *secundárias* ou *tipos*, é raro o texto que não inclua outros nomes próprios, referentes a entidades ficcionais partilhadas pela cultura, entidades históricas, e entes religiosos. A perspectiva das Humanidades Digitais é especialmente relevante para iluminar estes outros tipos de personagens, normalmente diluídos em uma história, mas que ganham corpo quando considerados parte de um acervo amplo. Veja-se a tabela 1 com essas quantidades (revistas) para obras diferentes:

Obra	tamanho em palavras	históricas	literárias	religiosas
Dom Casmurro	78606	36	13	72
Helena	66241	11	4	25
As Pupilas do Senhor Reitor	114343	9	16	176
Úrsula	54292	4	3	114

Tabela 1. Número de outras personagens mencionadas em quatro romances

Deus é de longe a entidade religiosa mais mencionada nos textos em português. Na tabela 2, comparamo-lo com menções ao diabo em romances, novelas e contos.

Deus	11.175	Diabo	3.308
Jesus	1.054	Satanás	135
Cristo	677	demo	113
Nosso Senhor	290	Lúcifer	49
Santo Deus	266	Belzebu	14
Nosso Senhor Jesus Cristo	260	Satã	13
Total	13.722	Total	3.632

Tabela 2. Distribuição de termos referentes a deus e ao diabo em 359 romances, novelas e contos de língua portuguesa

Deus (13.722) é assim invocado mais de três vezes mais do que o diabo (3.632), mas a variação entre autores também é relevante: Maria Peregrina de Sousa é quem faz (relativamente) mais alusões à divindade, seguida de perto por Maria Firmina dos Reis, enquanto que Olavo Bilac é quem mais se refere ao diabo, seguido de Artur Azevedo.

5. Redes de personagens

Com base na referência por nomes próprios, já unificados, criamos um programa que calcula a coocorrência numa janela deslizante de 3000 palavras (com sobreposição de 500 palavras), e que conta as vezes que duas personagens coocorrem nessa janela. Esses valores permitem-nos desenhar uma rede não dirigida, veja-se a figura 1, referente aos romances *As Pupilas do Senhor Reitor* e *Dom Casmurro*.

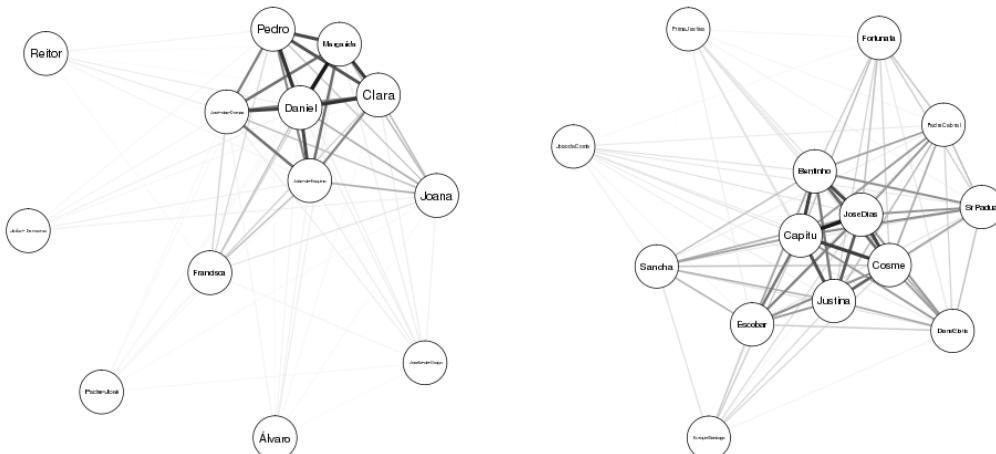


Figura 1. Redes de personagens

A simples visualização das redes relativas aos dois romances permite ver uma maior relação entre todas as personagens de *Dom Casmurro* comparada com *As Pupilas do Senhor Reitor*, que apresenta (aparentemente) cinco personagens pouco ligadas com o resto da trama.

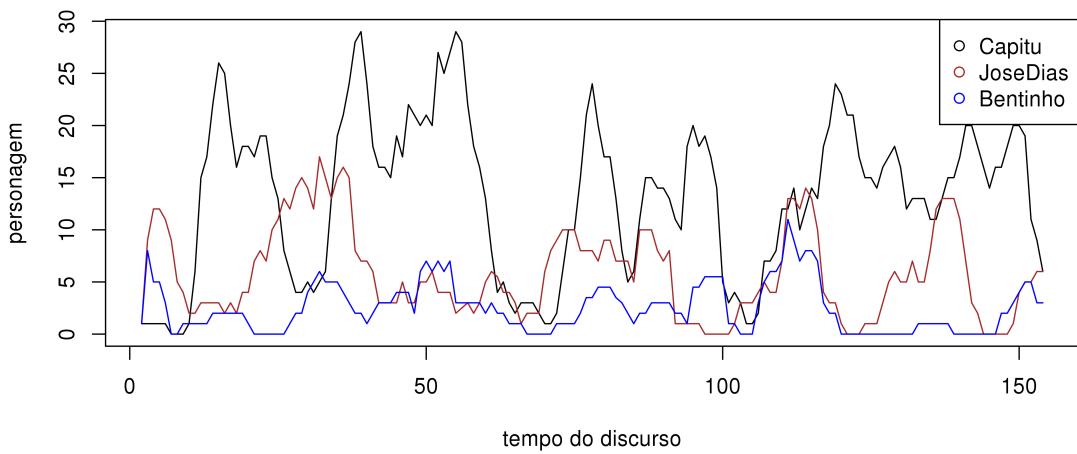


Figura 2. Algumas personagens de *Dom Casmurro* ao longo do tempo

6. Presença das personagens ao longo do enredo

Outro tipo de visualização que podemos fazer é a importância das personagens ao longo do livro, como a figura 2 mostra.

Quanto às ações desempenhadas pelas personagens, na esteira de [Bonch-Osmolovskaya and Skorinkin 2017, Archer and Jockers 2016], pesquisamos quais as ações das personagens (descritas por nomes próprios) nos quatro romances já mencionados, e pudemos constatar que a maior parte das ações eram comuns às personagens femininas e masculinas, embora as mulheres sorrissem mais e os homens respondessem mais.

7. Trabalho futuro

Pretendemos em breve associar sentimentos às personagens, como [Klinger et al. 2016], e polaridade à história, seguindo o exemplo de [Archer and Jockers 2016]. Para isso estamos a investigar as emoções em português e em texto literário, veja-se [Ramos and Freitas 2019, Santos and Simões 2019].

Ao termos revisado completamente a atribuição do género dos nomes próprios em todo o corpo OBras [Santos et al. 2018a], veja-se [Rocha et al. 2019], podemos fazer uma leitura mais distante das propriedades do género em toda a literatura a que temos acesso. Do mesmo modo, a explicitação dos sujeitos nos textos literários é um ponto que precisa ser tratado, a fim de evitar limitações das análises.

Pretendemos também obter redes sem revisão para estudar qual o mínimo de intervenção humana necessário para poder comparar centenas de obras, no âmbito da nossa filosofia de colaboração humana-máquina.

No âmbito da comparação entre várias línguas, poderemos, além do género das personagens olhar para as profissões mais mencionadas, como proposto em [Stanković et al. 2019].

Referências

- Archer, J. and Jockers, M. L. (2016). *The Bestseller Code: Anatomy of the Blockbuster Novel*. Sr. Martins's Press.
- Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, Aarhus, Denmark.
- Bonch-Osmolovskaya, A. and Skorinkin, D. (2017). Text mining War and peace: Automatic extraction of character traits from literary pieces. *Digital Scholarship in the Humanities*, 32. Supplement 1.
- Freitas, C., de Souza, E., and Rocha, L. (2019). Quantificando (e qualificando) o sujeito oculto em português. In *VI Jornada de Descrição do Português, STIL 2019*.
- Grayson, S., Wade, K., Meaney, G., Rothwell, J., Mulvany, M., and Greene, D. (2016). Discovering structure in social networks of 19th century fiction. In *Proceedings of the 8th ACM Conference on Web Science*, pages 325–326. ACM.
- Klinger, R., Suliya, S. S., and Reiter, N. (2016). Automatic emotion detection for quantitative literary studies. In *DH*.
- Krug, M., Weimer, L., Reger, I., Macharowsky, L., Feldhaus, S., Puppe, F., and Jannidis, F. (2018). Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. Technical report. DARIAH-DE Working Papers, Nr. 27.
- Moretti, F. (2011). Network theory, plot analysis. *New Left review*, 68:80–102.
- Ramos, B. and Freitas, C. (2019). "sentimento de quê?": uma lista de sentimentos para a análise de sentimentos. In *STIL 2019 - The 12th Brazilian Symposium in Information and Human Language Technology, Salvador, BA, Brazil, October, 15-18, 2019*.
- Rocha, L., Freitas, C., and Santos, D. (2019). Recursos para leitura distante em português: diálogos entre pln e humanidades digitais. In *TILic 2019*.
- Santos, D. (2019). Literature studies in Literateca: between digital humanities and corpus linguistics. In Doerr, M., Øyvind Eide, Grønvik, O., and Kjelsvik, B., editors, *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, pages 89–109. Novus forlag.
- Santos, D., Freitas, C., and Bick, E. (2018a). Obras: a fully annotated and partially human-revised corpus of brazilian literary works in the public domain. OpenCor.
- Santos, D., Freitas, C., and Lopes, J. M. (2018b). Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português. In Higuchi, S. and Ribeiro, C. J. S., editors, *I Congresso Internacional em Humanidades Digitais no Rio de Janeiro*, pages 375–383.
- Santos, D. and Simões, A. (2019). Towards a computational environment for studying literature in Portuguese. In *DH Budapest 2019, Digital Humanities Conference*.
- Stanković, R., Santos, D., Frontini, F., Erjavec, T., and Brando, C. (2019). Named entity recognition for distant reading in several european literatures. In *DH Budapest 2019, Digital Humanities Conference*.

Um Conjunto de Dados Extraído do Twitter para Análise de Sentimentos na Língua Portuguesa

Ewerton Paulo da Silva¹, Yuri Malheiros¹, Rodolffo Teles Araujo Nunes²,
Igor Leal Antunes², Thaís Gaudencio do Rêgo²

¹Departamento de Ciências Exatas – Universidade Federal do Paraíba (UFPB)
Rio Tinto - PB - Brasil

²Centro de Informática - Universidade Federal do Paraíba (UFPB)
João Pessoa - PB - Brasil

{ewerton.paulo,yuri}@dcx.ufpb.br

Abstract. *The large amount of data generated by users on social networks has attracted increasing interest in the analysis of the opinions and feelings that are being expressed. For this, one of the most widely used techniques is machine learning, which needs large datasets to work properly. However, in Portuguese, few datasets for this purpose are available, limiting the development of applications in that language. Therefore, this work aims to collect messages from Twitter and to classify their sentiments to create a dataset for sentiment analysis. Volunteers labeled 2,787 messages that are publicly available. Using the dataset, we achieved 0.4503 accuracy through machine learning, a result higher than the 0.3523 accuracy using the SenticNet lexicon.*

Resumo. *A grande quantidade de dados gerada por usuários nas redes sociais tem despertado cada vez mais o interesse na análise das opiniões e sentimentos que estão sendo expressados. Para isso, uma das técnicas mais utilizadas é a aprendizagem de máquina, que precisa de grandes conjuntos de dados para funcionar adequadamente. Entretanto, na língua portuguesa, poucos conjuntos de dados para esse fim estão disponíveis, limitando o desenvolvimento de aplicações nesse idioma. Com isso, este trabalho tem como objetivo a coleta de mensagens do Twitter e a classificação do sentimento delas para criação de um conjunto de dados para a análise de sentimentos. Voluntários rotularam 2.787 mensagens que estão disponibilizadas publicamente. Utilizando os dados colecionados, conseguiu-se 0,4503 de acurácia através de aprendizagem de máquina, resultado superior aos 0,3523 de acurácia usando o lexicon SenticNet.*

1. Introdução

Ao longo dos últimos anos, as redes sociais se tornaram uma das principais plataformas de comunicação do planeta, nas quais um número muito grande de pessoas consegue se expressar compartilhando diversos tipos de informações, sejam elas fotos, vídeos, textos, etc. A grande quantidade de dados gerada pelos usuários das redes sociais é preciosa e muitas vezes traz informações que não são percebidas com facilidade. Por exemplo, é possível extrair automaticamente de mensagens textuais sobre que assunto elas se referem, que idioma estão as mensagens e também que sentimento elas carregam: felicidade,

tristeza, excitação, raiva, etc. Este último tipo de informação é tratado especificamente pela área de análise de sentimentos [Liu and Zhang 2012] [Pang and Lee 2008].

A detecção de sentimentos por meio de computadores tem ganhado muita atenção nos últimos anos, tanto nas universidades quanto nas empresas [Liu et al. 2005] [Gamon 2004]. Um dos motivos de tal interesse é justamente o aumento da quantidade de conteúdo gerado pelas pessoas na Internet, principalmente quando elas estão expressando opinião. Entre as técnicas mais utilizadas para detecção de sentimentos está a aprendizagem de máquina supervisionada. Nela, classificadores usam dados previamente rotulados com os seus sentimentos para aprender padrões e conseguir prever novas entradas. Para treinar classificadores são necessários muitos dados, assim a disponibilidade de conjuntos de dados são essenciais para a realização de pesquisas e desenvolvimento de aplicações nessa área. Entretanto, conjuntos de dados com exemplos na língua portuguesa ainda são escassos, o que limita as aplicações voltadas para esse idioma [Moraes et al. 2015].

Tendo em vista esta necessidade, este trabalho tem como objetivo a coleta e classificação de *tweets*, que são as mensagens compartilhadas no Twitter, para criação de um conjunto de dados para análise de sentimentos na língua portuguesa. Para alcançar esse resultado foi desenvolvido um coletor de mensagens utilizando a API do Twitter. Em seguida, foi desenvolvida uma aplicação web para que voluntários pudessem classificar as mensagens coletadas em relação ao seu sentimento (positivo, negativo ou neutro). No total foram classificados 2.787 *tweets* sendo 888 positivos, 881 negativos e 1.018 neutros.

O restante do artigo está estruturado da seguinte forma. Na seção 2 são descritos os trabalhos relacionados, na seção 3 é apresentado o processo de criação do conjunto de dados, na seção 4 são apresentados resultados da avaliação do conjunto de dados e uma breve discussão e na seção 5 temos conclusão.

2. Trabalhos Relacionados

Outros trabalhos na literatura procuraram alcançar resultados similares criando conjuntos de dados em português para análise de sentimentos.

No trabalho de [Brum and Nunes 2017] um conjunto de dados de 15.000 mensagens rotuladas foi criado. Para isso, foi necessária a coleta de dados utilizando a API do Twitter focado em mensagens compartilhadas durante a exibição de programas de TV. As mensagens foram classificadas entre positivas, negativas e neutras. O processo de classificação foi realizado através de uma ferramenta web de anotação utilizada por sete participantes nativos da língua portuguesa com o auxílio de um guia da língua. O conjunto possui um total de 6.648 mensagens positivas, 3.926 neutras e 4.426 negativas. Também foram realizados experimentos com três métodos de aprendizagem de máquina. Ao final, o conjunto de dados criado foi disponibilizado por meio de um repositório público.

O PELESent [Corrêa et al. 2017] foi criado com o objetivo de ser um conjunto de dados com uma grande quantidade de *tweets*, possuindo um total de 980.067 mensagens. Pelo grande custo de realizar essa anotação por humanos, *emojis* foram utilizados para classificar as mensagens, sendo 554.623 positivas e 425.444 negativas. Para avaliação, métodos de classificação de polaridades foram treinados com o conjunto de dados e os modelos resultantes foram aplicados em cinco outros conjuntos de dados anotados manualmente.

O 7x1-PT é um conjunto de dados para análise de sentimentos com *tweets* que foram enviados ao longo da partida da Alemanha com o Brasil durante a Copa do Mundo de 2014 da FIFA [Moraes et al. 2015]. Durante o jogo entre as equipes foram buscadas mensagens que continham palavras relacionadas à Copa do Mundo, por exemplo, hexa, vencedor, etc. O conjunto de dados final foi classificado por dois anotadores humanos totalizando 2.728 *tweets*, sendo 157 positivos, 1.771 neutros e 800 negativos.

No trabalho de [de Arruda et al. 2015] foi criado um conjunto de dados de notícias extraídas de 4 grandes jornais brasileiros. Para coletar os dados, durante sete dias, às 20:00 horas, um *crawler* capturava pelo menos 20 notícias sobre política de perfis do Twitter dos meios de comunicação selecionados. Em seguida, essas notícias foram divididas em parágrafos para que quatro anotadores humanos determinassem sobre que pessoa o parágrafo se referia e o sentimento do parágrafo em relação a essa pessoa. No total foram classificados 1.447 parágrafos de 113 notícias.

Outro trabalho realizado abordando notícias brasileiras foi desenvolvido por [Dosciatti et al. 2015], nele foi construído um corpus de notícias para análise de sentimentos que poderiam ter as seguintes classificações: alegria, tristeza, raiva, surpresa, repugnância e medo. Foram coletados 2.000 textos que foram classificados por seis anotadores voluntários com experiência de no mínimo 15 anos em linguística.

Em comparação a esses trabalhos descritos, o nosso principal diferencial é que não restringimos o escopo das mensagens para um contexto específico e que coletamos mensagens por um período significativamente maior que os trabalhos citados. Além disso, na classificação, cada mensagem podia receber o julgamento de até cinco pessoas para obter uma maior consistência.

3. Criação do Conjunto de Dados

A criação do conjunto de dados foi realizada em duas etapas principais. Na primeira, foi desenvolvida uma ferramenta para coletar mensagens compartilhadas no Twitter. Na segunda, uma ferramenta web para rotular os textos foi desenvolvida e disponibilizada para que voluntários classificassem as mensagens coletadas de acordo com o seu sentimento (positivo, negativo ou neutro). A seguir mostraremos mais detalhes sobre como foram realizadas essas duas etapas.

3.1. Coleta de dados

Para realizar a coleta das mensagens foi desenvolvida uma ferramenta em Python que utiliza a API do Twitter para buscar mensagens compartilhadas na rede social de acordo com palavras-chave. Como centenas de milhões de *tweets* são enviados a cada dia [Domo 2019], temos uma variedade muito grande de textos compartilhados. Para buscar *tweets* com uma maior chance de ter algum sentimento, escolheu-se usar como palavras-chave adjetivos da língua portuguesa. Os adjetivos utilizados nas buscas foram disponibilizados pelo thesaurus TeP 2.0 [Dias-Da-Silva and de Moraes 2003].

A ferramenta de coleta foi executada em um servidor entre os dias 24/09/2018 e 06/12/2018. Durante esse período, a ferramenta selecionava aleatoriamente um adjetivo do thesaurus TeP 2.0 e efetuava uma busca no Twitter, guardando as mensagens encontradas. Após salvar os *tweets* no banco de dados, a ferramenta começava a checar se 30

minutos já se passaram, para em seguida escolher um novo adjetivo e recomeçar o processo. Os dados armazenados de cada *tweet* foram o ID do *tweet* e o seu conteúdo textual. Os passos do processo de coleta são sumarizados na Figura 1.

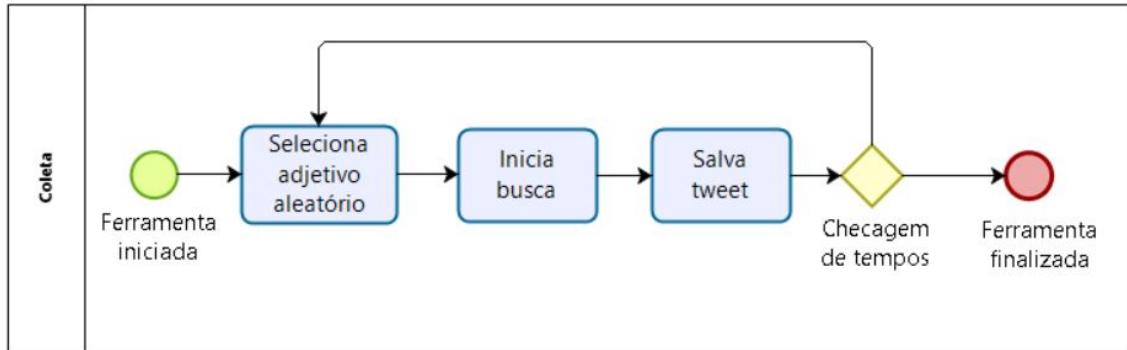


Figura 1. Passos executados para a coleta de tweets

A etapa de salvar *tweets* seguiu alguns critérios de avaliação do texto antes de armazená-los. Primeiramente, foi verificado se o texto do *tweet* já havia sido salvo no banco de dados. Quando isso acontecia, como não queríamos ter dados duplicados no conjunto de dados, o *tweet* era descartado. Em seguida, para tentar obter um maior número de mensagens com sentimentos foi calculada a polaridade da mensagem utilizando o SenticNet 5 [Cambria et al. 2018] através da biblioteca SenticNet API¹.

A polaridade é um número que varia no intervalo de -1 a 1. Mensagens com polaridades próximas de 0 são consideradas neutras, sem sentimento. Quanto mais próxima de -1 a polaridade de uma mensagem for, mais ela é negativa e, no caso contrário, quanto mais próxima de 1 a polaridade, mais positiva é a mensagem. Para calcular a polaridade de uma mensagem foi realizada uma média das polaridades das palavras da mensagem presentes no SenticNet. Com isso, polaridades entre -0,003 e 0,003 foram consideradas neutras no processo de coleta e, por isso, foram descartadas.

Após todo o processo de coleta, foram armazenados no banco de dados 641.471 *tweets* para serem classificados posteriormente em relação aos seus sentimentos. A Tabela 1 mostra três exemplos de *tweets* coletados.

Tabela 1. Exemplos de tweets coletados

Tweet
esse gel p dor é milagroso
um vampiro aventureiro e um vampiro mimado https://t.co/rVdBZaFlZv
Dividido entre a tristeza, o ódio e o veneno que existem dentro de mim

3.2. Classificação

Com os *tweets* coletados e salvos em um banco de dados, a etapa subsequente foi a classificação das mensagens por voluntários. Para isso, foi desenvolvida uma ferramenta web para que os voluntários julgassem se mensagens tinham sentimentos positivos, negativos ou neutros. Ao iniciar a ferramenta é apresentado para o usuário um *tweet* escolhido

¹<https://github.com/yurimalheiros/senticnetapi>

aleatoriamente do banco de dados. Abaixo do *tweet* são disponibilizados três botões para o usuário interagir, cada um relacionado a uma das possíveis classificações. Quando o usuário seleciona um sentimento, a ferramenta grava a classificação e exibe uma nova mensagem para avaliação. A tela exibida para o usuário pode ser visualizada na Figura 2.

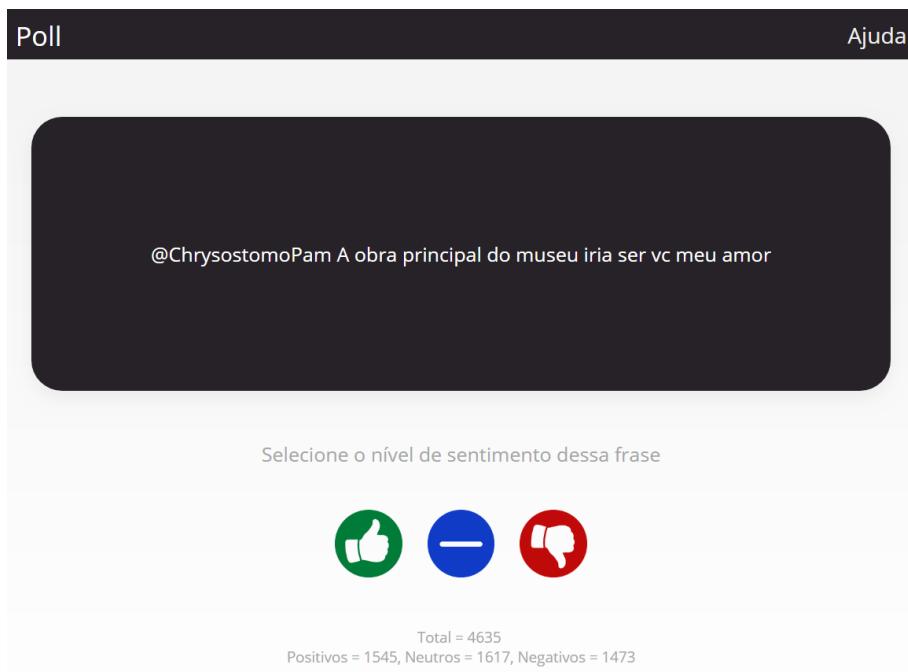


Figura 2. Tela da ferramenta para classificação de tweets

A ferramenta foi disponibilizada para os alunos dos cursos de Bacharelado em Sistemas de Informação e Licenciatura em Ciência da Computação da Universidade Federal da Paraíba classificarem as mensagens do conjunto de dados. Para tornar o processo mais confiável, um *tweet* poderia ser classificado até cinco vezes, sendo sua classificação final o sentimento mais escolhido. Assim, tentou-se dar maior consistência às classificações atribuídas, pois um julgamento destoante de um voluntário poderia ser corrigido pelos julgamentos restantes. Outro ponto importante é que foram coletados 641.471 *tweets*, um número muito grande para ser classificado pelos voluntários, portanto na ferramenta foi usado um subconjunto de 10.000 *tweets* escolhidos aleatoriamente a partir dos *tweets* coletados. Por fim, verificações manuais periódicas foram realizadas nas classificações para detecção de comportamentos indesejáveis, por exemplo, muitas atribuições de um único sentimento para muitas mensagens em pouco tempo.

Ao final da etapa de classificação foram realizadas 1.545 avaliações positivas, 1.473 negativas e 1.617 neutras. Com isso, no total foram classificados 2.787 *tweets* sendo 888 positivos, 881 negativos e 1.018 neutros. O conjunto de dados está disponível publicamente através do *GitHub* em um arquivo *CSV* que contém o ID de cada *tweet* e o seu sentimento correspondente². A Tabela 2 traz uma amostra de três *tweets* que estão no conjunto de dados finais e os seus sentimentos.

²<https://github.com/arialab/tash-pt>

Tabela 2. Exemplos retirados do conjunto de dados

Texto	Sentimento
queria deixar registrado o quanto eu fiquei feliz por ter recebido um áudio da sophi falando que tinha feito um desenho pra mim	positivo
É feliz quem sonha. Más só tem sucesso quem se dispõe a pagar o preço para transformar sonho em realidade...	positivo
Anotado, vou querer ver a tatuagem depois e o cabelo tbm	neutro
Eu cometí o terrível erro de beber uma caneca de café agr de tarde	negativo
Sem querer eu descubro as coisas, esse doido mente muitoooooooooooo, não sei como eu fiquei com esse inútil	negativo

4. Avaliação do conjunto de dados

Antes de qualquer treinamento utilizando o conjunto de dados, realizou-se o pré-processamento do texto coletado com auxílio da biblioteca NLTK [Loper and Bird 2002]. Ele consistiu em quatro etapas: remoção de *links* por meio de expressão regular, remoção de *stopwords* em português, aplicação de *stemming* e remoção de acentuação. Após realizadas essas etapas, o texto foi vetorizado utilizando o *TFIDF Vectorizer* da biblioteca *Scikit-learn* [Pedregosa et al. 2011], formato adequado aos classificadores.

Com o texto pré-processado, foi utilizada validação cruzada com *5-folds* para avaliar o resultado dos classificadores *LogisticRegression*, *MultinomialNB*, *SGDClassifier*, *LinearSVC* e o *NuSVC*, implementados pela biblioteca *Scikit-learn*. Nela, 80% do conjunto de dados (2.229 exemplos) foi utilizado para treinamento do classificador, e 20% para teste (558 exemplos). Além disso, para fins de comparação, os exemplos do conjunto de testes também foram classificados usando o SenticNet. Nesse caso, para obter o sentimento de um *tweet*, foi calculada a média das polaridades das palavras do *tweet* existentes no SenticNet. Com o resultado em mãos, a classificação foi atribuída seguindo as regras:

- Se a média das polaridades for maior que 0,1, o *tweet* tem sentimento positivo;
- Se a média das polaridades for menor que -0,1, o *tweet* tem sentimento negativo;
- Se a média das polaridades estiver entre -0,1 e 0,1, o *tweet* tem sentimento neutro.

A Tabela 3 traz os resultados da validação cruzada e da classificação utilizando o SenticNet. Na primeira coluna tem-se a técnica de classificação e na segunda a média da acurácia dos *5 folds*.

Tabela 3. Resultados das classificações

Classificador	Acurácia
LogisticRegression	0,4503
MultinomialNB	0,4415
SGDClassifier	0,4250
LinearSVC	0,4365
NuSVC	0,4216
SenticNet	0,3523

Analizando os resultados, percebe-se que a maior acurácia foi do classificador *LogisticRegression*, alcançando o valor 0,4503. Entretanto, a maioria dos classificadores obteve resultados semelhantes entre 0,42 e 0,45. Assim, considerando que uma classificação

aleatória de 3 classes (positivo, negativo e neutro) tende a 0,33 de acurácia, então todos os classificadores ficaram acima dessa taxa base. Em relação ao SenticNet, a acurácia foi de 0,3523, só um pouco acima da taxa base de 0,33, mostrando um desempenho ruim dessa abordagem para classificação.

A presença de linguagem informal nos *tweets*, com gírias e erros de ortografia prejudica o desempenho da classificação usando lexicons como o SenticNet, que analisa conceitos preestabelecidos. Dessa forma, no contexto de redes sociais, classificadores de aprendizagem de máquina supervisionada treinados com os dados coletados das próprias redes sociais tem o potencial de conseguir um desempenho melhor que lexicons, sendo mais adequados para problemas nessa área.

5. Conclusões

Neste trabalho foi desenvolvido um conjunto de dados para análise de sentimento na língua portuguesa utilizando mensagens do Twitter. Esse conjunto de dados tem potencial para servir de base para novas pesquisas e aplicações em análise de sentimentos no nosso idioma. Para isso, coletaram-se mensagens compartilhadas no Twitter que foram classificadas manualmente por voluntários. O conjunto de dados final possui 2.787 mensagens, sendo 888 positivas, 881 negativas e 1.018 neutras, e está disponível publicamente.

Foram realizados testes com classificadores de aprendizagem de máquina treinados com o conjunto de dados desenvolvido. A maioria dos classificadores testados obtiveram resultados semelhantes, sendo 0,4503 o máximo de acurácia conseguida pelo *LogisticRegression*. Ao comparar com o SenticNet, todas as abordagens de aprendizagem de máquina obtiveram acuráncias superiores.

Como foram coletadas mais mensagens do que as classificadas pelos voluntários, como trabalhos futuros, pretende-se classificar mais *tweets*, para aumentar o tamanho do conjunto de dados. Além disso, testes mais rigorosos com classificadores de aprendizagem precisam ser realizados, incluindo técnicas como redes neurais.

6. Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Universidade Federal da Paraíba pelo auxílio financeiro através do Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Referências

- Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.
- Cambria, E., Poria, S., Hazarika, D., and Kwok, K. (2018). Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Corrêa, E. A., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., and Brum, H. B. (2017). Pelesent: Cross-domain polarity classification using distant supervision. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 49–54. IEEE.

- de Arruda, G. D., Roman, N. T., and Monteiro, A. M. (2015). An annotated corpus for sentiment analysis in political news. In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 101–110.
- Dias-Da-Silva, B. C. and de Moraes, H. R. (2003). A construção de um thesaurus eletrônico para o português do brasil. *ALFA: Revista de Linguística*, 47(2).
- Domo (2019). Data never sleeps 6. <https://www.domo.com/learn/data-never-sleeps-6>. Acessado em: 18-05-2019.
- Dosciatti, M. M., Ferreira, L. P. C., and Paraiso, E. C. (2015). Anotando um corpus de notícias para a análise de sentimentos: um relato de experiência (annotating a corpus of news for sentiment analysis: An experience report). In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 121–130.
- Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 611–617. Association for Computational Linguistics.
- Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.
- Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.
- Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.
- Moraes, S. M., Manssour, I. H., and Silveira, M. S. (2015). 7x1pt: um corpus extraído do twitter para análise de sentimentos em língua portuguesa. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 21–25. SBC.
- Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Predição da Complexidade Textual de Recursos Educacionais Abertos em Português

Murilo Gazzola¹, Sidney Evaldo Leal¹, Sandra Maria Aluisio¹

¹Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP

mgazzola@icmc.usp.br, sidleal@gmail.com, sandra@icmc.usp.br

Abstract. In 2016, UNESCO stated the priorities for the use of Open Educational Resources (OER), highlighting the main research challenges. The lack of quality of OER is a challenge to be overcome. In a data analysis of the Integrated Platform of the Ministry of Education (MEC-RED) in May 2018, 41% of the resources did not have the teaching stage metadata, making it difficult to search OER, use and edit them. The Textual Complexity task can help identify texts that have linguistic complexity appropriate to specific series, allowing to complete the teaching stage in MEC-RED. In this article, we evaluate the impact of the textual genre in the evaluation of textual complexity, using a model trained in a large corpus of didactic texts and tested in 2 OER datasets of MEC-RED. The best trained model (F-measure 0.804) had an F-measure of 0.518 in a set of OER of the same genre and 0.389 of F-measure for the animation/simulation and practical experiment, two genres of interest in this research.

Resumo. Em 2016, a UNESCO escreveu em seu relatório as prioridades para o uso de Recursos Educacionais Abertos (REA), destacando os principais desafios de pesquisas. A falta de qualidade dos REA é um desafio a ser superado. Em uma recolha na Plataforma Integrada do Ministério da Educação (MEC-RED) de maio de 2018, 41% dos recursos não possuíam classificação da etapa de ensino, dificultando sua busca, uso e edição. A tarefa Complexidade Textual pode ajudar a identificar textos que tem complexidade linguística adequada a séries específicas, permitindo completar a etapa de ensino. Neste artigo, avaliamos o impacto do gênero textual na avaliação da complexidade textual, com um modelo treinado em um grande córpus de textos didáticos e testado em 2 conjuntos de REA da MEC-RED. O melhor modelo treinado (0.804 de F-measure) teve uma F-measure de 0.518 em um conjunto de REA de mesmo gênero e 0.389 de F-measure para os recursos do tipo animação/simulação e experimento prático, dois gêneros de interesse nesta pesquisa.

1. Introdução

O termo Recursos Educacionais Abertos (REA) foi cunhado em 2002 pela Unesco [UNESCO, 2002] no fórum sobre o impacto de cursos de ensino superior aberto em países em desenvolvimento. Os REA podem ser caracterizados como materiais de ensino, aprendizagem e pesquisa em qualquer meio de armazenamento, que estão disponíveis através da licença que permite quatro liberdades mínimas conhecidas como 4R: Revisar, Reusar, Recombinar e Redistribuir [Wiley et al. 2014]. Portanto, devido à importância do tema

[Miao et al. 2016] [Wiley et al. 2014], a Organização das Nações Unidas (ONU) definiu os principais problemas relacionados ao desenvolvimento e uso de REA: i) *o problema de qualidade dos REA*; ii) *o problema da descoberta, ou seja, como encontrar REA*; iii) *o problema da sustentabilidade*, isto é, como financiá-los; iv) *o problema de localização e recontextualização* dos REA; e v) *o problema do remix*, isto é, a dificuldade de identificar a granularidade da mudança de conteúdo por outras pessoas e o nível de mudança. Consideramos que a qualidade é um fator importante para garantir uma educação e ensino de qualidade, utilizando esses materiais. Assim, buscamos critérios e métodos da área de Processamento de Línguas Naturais (PLN) para avaliar a qualidade dos REA. Os principais trabalhos que avaliaram a qualidade de REA ou materiais semelhantes aos REA não trazem uma análise linguística, nem avaliam o conteúdo de REA, apenas tratam os seus metadados para julgar a qualidade ou usam indicadores subjetivos [Bethard et al. 2009], [Dalip et al. 2011], [Leary et al. 2011], [Cechinel et al. 2011], [Ahmed and Fuge 2017].

A plataforma de Recursos Educacionais do Ministério da Educação (MEC), conhecida como Plataforma Integrada MEC¹ (MEC-RED), faz parte de um dos compromissos do Brasil na *Open Government Partnership* (OGP) para fortalecer práticas que envolvem a transparência dos atos governamentais e promovem a participação social e o acesso à informação pública [MEC 2019b]. Além disso, a MEC-RED centralizou todos os materiais do Portal do Professor², Banco Internacional de Objetos Educacionais³, Domínio Público⁴ e TV Escola⁵ [MEC 2019a]. Na MEC-RED, há quatro filtros de busca por recursos: componentes curriculares, tipos de recurso, etapas de ensino e palavra-chave, embora a plataforma tenha outros metadados como título, pessoas que favoritam o recurso, escala de estrelas, URL para download associado ao material, descrição do recurso, autor do envio, autor do material, *tags* associadas ao material e tipo de recurso. No início de maio de 2019, a MEC-RED contava com 31.488 recursos, mas as avaliações feitas neste trabalho foram realizadas com informações de uma recolha realizada em maio de 2018, quando a plataforma possuía 28.026 recursos. Destes últimos, 6.966 recursos (41%) não possuíam a informação sobre etapas de ensino preenchido. Particularmente o metadado etapas de ensino permite a busca por material relacionado com a sua complexidade textual e conceitual, para a recuperação de material adequado a uma das etapas do Sistema Educacional Brasileiro. Assim, espera-se o seu preenchimento correto, sendo um item importante para se avaliar a qualidade de um recurso.

Sabemos que certos conteúdos são estudados em séries específicas do Ensino Fundamental I e II, Ensino Médio e Ensino Superior e que a cada nova etapa novos gêneros textuais (por exemplo, romances, crônicas, fábulas, ensaios, anúncios, editoriais e reportagens de jornal, cartas, relatórios, anedotas, dentre outros) são trabalhados e que também os próprios componentes curriculares (ou disciplinas) trazem características de complexidade textual variadas, como, por exemplo, os textos de ciência trazem uma terminologia técnica, textos de história trazem ideologias e interpretação de eventos com grande número de personagens e locais geográficos, os de matemática símbolos e organização textual novos e conceitos mais abstratos [Fang 2016]. Neste trabalho, propomos iden-

¹<https://plataformaintegrada.mec.gov.br/>

²<http://portaldoprofessor.mec.gov.br>

³<http://objetoseducacionais2.mec.gov.br>

⁴<http://www.dominiopublico.gov.br/>

⁵<http://tvescola.org.br/>

tificar automaticamente a etapa de ensino, via uma tarefa do PLN chamada predição automática da complexidade textual (*text readability*, em inglês) que estuda as características lexicais, sintáticas, semânticas e discursivas [Fang 2016] que podem impactar na dificuldade/facilidade de um texto ser lido por um aluno, que tem um conjunto de conhecimentos prévios, no contexto de uma dada atividade escolar. Como a MEC-REC possui 14 tipos de recursos, excluímos a análise dos áudios, imagens, infográfico, mapas e vídeos, por não se apresentarem no formato textual. De especial interesse neste trabalho são os tipos de recursos animação/simulação, aplicativo móvel, jogos, experimento prático e software educacional. Entretanto, 100% dos aplicativos móveis e 99,47% dos jogos não apresentam informação sobre a etapa de ensino, inviabilizando a compilação de um grande córpus de REA para o treinamento para um preditor de complexidade textual.

Vários trabalhos da literatura de predição de complexidade textual utilizam materiais de séries escolares ([Vajjala and Meurers 2014], ([Hartmann et al. 2016, Wagner Filho et al. 2016b]), considerando a série na qual o texto é usado como substituto (*proxy*) para a sua complexidade linguística; neste artigo também adotamos essa abordagem. Neste trabalho, primeiro a investigar a predição automática da complexidade de textos para REA, buscamos avaliar as *features* importantes para a tarefa, usando um arcabouço de análise multinível, assim como faz o ambiente Coh-Metrix [Graesser and McNamara 2011], envolvendo métricas que tratam de: (i) palavras, (ii) sintaxe, (iii) conexão entre sentenças no discurso. Essas *features* são discutidas na Seção 3.2. Como não há nenhum grande córpus disponível publicamente para predizer a complexidade textual para as etapas do ensino no Brasil, compilamos um córpus anotado com as quatro etapas (Seção 3.1).

As seções a seguir são organizadas da seguinte maneira: os trabalhos relacionados são apresentados na Seção 2; na Seção 3 são apresentados os detalhes do córpus, as *features* e os métodos de aprendizado de máquina e seleção de *features* utilizados no trabalho; e na Seção 4 os resultados da avaliação intrínseca e extrínseca, usando o melhor modelo que foi avaliado no grande córpus em dois conjuntos de REA de gêneros diferentes.

2. Trabalhos Relacionados

[Graesser and McNamara 2011, Graesser et al. 2011] desenvolveram a ferramenta Coh-Metrix⁶ para língua inglesa que analisa textos usando métricas dos vários níveis da língua e que estão alinhadas com um arcabouço teórico de compreensão discursiva multinível. Em seu trabalho, utilizaram extratos de textos com média de 288,6 palavras para extraírem suas métricas, porém não informam a quantia mínima de palavras em cada extrato, enquanto nossa proposta traz uma quantia mínima de 300 palavras e uma média de 448 palavras, para viabilizar o cálculo de métricas como o Índice Flesch, por exemplo. Utilizam 53 métricas textuais, enquanto nossa proposta extrai 79 métricas dos extratos de textos. Usaram um grande córpus da língua inglesa com 37.520 extratos fornecido pelo *Touchstone Applied Science Associates* (TASA), já neste artigo tratamos a língua portuguesa para avaliação intrínseca de um preditor de complexidade textual, além da avaliação extrínseca na MEC-RED, considerando gêneros textuais semelhantes e diferentes do preditor criado.

[Scarton and Aluísio 2010] classificaram de forma binária os textos (simples versus complexos), usando o Coh-Metrix-Port com 40 métricas textuais. Usaram 4 córpuses

⁶<http://tea.cohmetrix.com/>

para treinamento/teste: textos jornalísticos do jornal Zero Hora (ZH) dos anos de 2006 e 2007, textos reescritos para crianças da seção “Para o seu filho ler” (PSFL) do ZH e textos do gênero científico do Ciência Hoje (CH) e Ciência Hoje das Crianças (CHC). O trabalho apresenta resultados da classificação binária usando o SVM do Weka com precisão do melhor classificador treinado de 97%, porém sabemos que classificadores binários que medem uma grande distância de idade (crianças versus adultos) são mais simples do que classificadores com mais classes. Neste trabalho, usamos 4 classes e avaliamos o preditor de forma intrínseca também extrínseca, usando córpus de gêneros de texto distintos e semelhantes ao preditor.

[Hartmann et al. 2016] reportam a classificação da complexidade de textos do gênero didático em português para cinco anos do Ensino Fundamental (3º, 4º, 5º, 6º e 7º anos). O córpus compilado é formado por textos de diversas fontes e possui 7.645 textos compilados de Livros Didáticos, NILC Corpus, Testes do SARESP, CHC, FSP, PSFL e Mundo Estranho; sem indicação da quantidade mínima de palavras que esses textos possuem. Apresentam uma classificação mais fina do que a tratada neste artigo, e utilizam 108 métricas para a criação do modelo preditivo. Já a nossa proposta utiliza 79 métricas e trata a divisão por etapas escolares. Utilizam apenas o classificador SVM implementado no libsvm que obteve 56% de acurácia.

[Wagner Filho et al. 2016b] reportam a previsão automática do nível escolar da Wikilivros, considerando 3 níveis escolares (nível 1, 2 e 3). O córpus possui 77 textos e usaram 7 *features* para avaliação da inteligibilidade, com Regressão Logística⁷. Nossa proposta se diferencia, trazendo uma avaliação com *Logistic Regression*, *SVM*, *Random-Forest* e *Multilayer Perception*, além de usar 79 métricas textuais. Também realizamos uma avaliação extrínseca nos dados da plataforma MEC-RED, dada a motivação inicial do trabalho de avaliar a qualidade de REA no Brasil. Também utilizamos para o treinamento do preditor a Wikilivros (Wikibook em português) e realizamos uma avaliação intrínseca usando *cross-validation*. [Wagner Filho et al. 2016a] trazem uma continuação do trabalho de [Wagner Filho et al. 2016b], considerando a língua portuguesa do Brasil e a língua inglesa para a criação do córpus de trabalho. Usaram 9.829 textos da língua portuguesa com níveis de inteligibilidade mistas de 2 e 3 níveis compostos por Wikilivros, É Só o Começo⁸ (ESOC), PSFL, ZH, BrEscola⁹. Para a língua inglesa, usaram Wikibooks, Simple Wikipedia (SW) [Coster and Kauchak 2011] e Biografias Britânicas (BB); os níveis de inteligibilidade também são mistos, variando de 2,3 e 4 níveis. Usaram o Weka para geração dos modelos de classificação, usando os métodos SVM, Regressão Logística, DecisionStump, RandomForest, com *cross-validation*. A quantidade de *features* totaliza 134 para o idioma português e 89 para o inglês. Os melhores resultados foram com o SVM e a Regressão Logística. O trabalho [Wagner Filho et al. 2016a] avalia a generalização do modelo de classificação para inteligibilidade em diferentes níveis e em dois idiomas. Porém, os resultados são ruins quando utilizam 3 classes, sendo melhores para classes binárias e do mesmo nível de complexidade, por exemplo, textos apenas para crianças. Este trabalho diferencia do nosso, pois usamos 4 classes de complexidade textual e avaliamos gêneros textuais diferentes.

⁷Modelo SimpleLogistic da ferramenta Weka.

⁸Contrasta obras da literatura clássica brasileira com versões adaptadas.

⁹Córpus de materiais educativos para crianças e adolescentes.

Tabela 1. Córpus de livros-textos da Língua Portuguesa compilado: Marcha criança, Tudo É Linguagem, Projeto Porta Aberta, Projeto Ápis, Português, Buriti, Porta Aberta, Mundo Amigo, Nos Dias de Hoje, Projeto Teláris, CNEC Educação.

	Ensino Fund I	Ensino Fund II	Ensino Médio	Ensino Superior	Total	Dataset 1 (D1)	Dataset 2 (D2)
Fontes de textos	Livros -texto + PSFL	Livros-texto, SAEB, E-Book CNEC Educação	Wikilivros, ENEM 2015, 2016 e 2017	Wikilivros			
Docs	296	325	627	819	2.067	60	40
Sents	5.258	5.598	9.316	10.416	30.588	720	540
MTSP	20.58	24.31	29.81	39.15	31.35	31.17	46.27
Type	63.081	75.698	134.788	177.054	450.621	10870	9281
Token	101.911	127.705	241.267	342.534	813.417	20040	16216
TTR	0.618	0.592	0.558	0.516	0.553	0.542	0.572
D1	10	10	10	10	40		
D2	10	10	20	20	60		

3. Materiais e Métodos

3.1. Córpus dos Quatro Estágios Escolares do Sistema Educacional Brasileiro

Compilamos um grande córpus que abrange textos utilizados em diferentes etapas de ensino (ou níveis escolares) do Sistema Educacional Brasileiro, organizado nas seguintes etapas: Ensino Fundamental I (1º ao 5º ano), Ensino Fundamental II (6º ao 9º ano), Ensino Médio e Ensino superior. Essas quatro etapas de ensino são as mesmas utilizadas na MEC-RED para classificar os REA nos Estágios Escolares.

O córpus¹⁰ inclui: livros-texto, notícias da Seção *Para Seu Filho Ler* (PSFL) do jornal Zero Hora que apresenta algumas notícias sobre os mesmos tópicos do Zero Hora, mas escritas para crianças de 8 a 11 anos de idade , Exames do SAEB , Livros Digitais do Wikilivros em Português , Exames do Enem dos anos 2015, 2016 e 2017. Nossa córpus de trabalho compreende 2.067 extratos (min = 300 palavras, max = 596 palavras, média = 448) dos textos do córpus compilado. Como pode ser visto na Tabela 1, nosso córpus não é balanceado, pois o número de textos do Ensino Médio possui aproximadamente o dobro da quantidade do Ensino Fundamental I e do Ensino Fundamental II, por exemplo. Para resolver esse problema, foi utilizado o método ClassBalancer do Weka¹¹ antes da execução dos métodos de aprendizado de máquina (cf. mais detalhes na Seção 4).

3.2. Métricas de Complexidade Textual

A seleção inicial das métricas para a avaliação da complexidade textual baseou-se no estudo de [Graesser and McNamara 2011] que utilizou 53 métricas do Coh-Metrix agrupadas nas relacionadas às palavras, sentenças e conexões entre sentenças. Para selecionar métricas similares para o português, escolhemos duas ferramentas disponíveis publicamente: Coh-Metrix-Port [Scarton et al. 2010], Coh-Metrix-Dementia [da Cunha 2015] e o trabalho [dos Santos et al. 2017]. A Figura 1 mostra um recorte¹² das 79 métricas

¹⁰Disponível em <http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources>

¹¹<https://www.cs.waikato.ac.nz/ml/weka/>

¹²A lista completa está disponível em https://github.com/gazzola/corpus_readability_nlp_portuguese

Palavras	Sentenças	Conexões entre Sentenças
<p>1.adjective_ratio: proporção de Adjetivos em relação à quantidade de palavras.</p> <p>2.adverbs: proporção de Advérbios em relação à quantidade de palavras.</p> <p>3.pronoun_ratio: proporção de pronomes em relação à quantidade de palavras.</p> <p>4.first_person_pronouns: proporção de pronomes pessoais nas primeiras pessoas em relação à quantidade de pronomes pessoais.</p> <p>5.third_person_pronouns: proporção de pronomes pessoais nas terceiras pessoas em relação à quantidade de pronomes pessoais do texto.</p> <p>6.content_density: proporção de palavras de conteúdo em relação à quantidade de palavras funcionais do texto.</p> <p>7.conn_ratio: proporção de conectivos em relação à quantidade de palavras do texto.</p> <p>8.add_neg_conn_ratio: proporção de conectivos aditivos negativos em relação à quantidade de palavras.</p> <p>9.add_pos_conn_ratio: proporção de conectivos aditivos positivos em relação à quantidade de palavras.</p> <p>10.cau_pos_conn_ratio: proporção de conectivos causais positivos em relação à quantidade de palavras.</p> <p>11.concretude_mean: média dos valores de concretude das palavras de conteúdo.</p> <p>12.familiaridade_mean: média dos valores de familiaridade das palavras de conteúdo.</p> <p>13.imageabilidade_mean: média dos valores de imageabilidade das palavras de conteúdo.</p>	<p>1.words_per_sentence: média de palavras por sentença.</p> <p>2.sentence_length_min: quantidade Mínima de palavras por sentença.</p> <p>3.sentence_length_max: quantidade Máxima de palavras por sentença.</p> <p>4.sentence_length_standard_dev: desvio Padrão da quantidade de palavras por sentença.</p> <p>5.mean_noun_phrase: média dos tamanhos médios dos sintagmas nominais nas sentenças.</p> <p>6.min_noun_phrase: mínimo entre os tamanhos de sintagmas nominais do texto.</p> <p>7.max_noun_phrase: máximo entre os tamanhos de sintagmas nominais do texto.</p> <p>8.std_noun_phrase: desvio-padrão do tamanho dos sintagmas nominais do texto.</p> <p>9.words_before_main_verb: quantidade Média de palavras antes dos verbos principais das orações principais das sentenças.</p> <p>10.passive_ratio: proporção de orações na voz passiva analítica em relação à quantidade de orações do texto.</p> <p>11.yngve: Complexidade Sintática de Yngve.</p> <p>12.frazier: Complexidade Sintática de Frazier.</p> <p>13.dep_distance: distância na árvore de dependências.</p>	<p>1.adj_cw_ovl: Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes.</p> <p>2.adj_arg_ovl: Quantidade média de referentes que se repetem nos pares de sentenças.</p> <p>3.arg_ovl: Quantidade média de referentes que se repetem nos pares de sentenças adjacentes.</p> <p>4.adj_stem_ovl: Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças.</p> <p>5.stem_ovl: Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes.</p> <p>6.ttr: Proporção de types (despreza repetições de palavras) em relação à quantidade de tokens (computa repetições de palavras).</p> <p>7.content_word_diversity: Proporção de types de palavras de conteúdo em relação à quantidade de tokens de palavras de conteúdo.</p> <p>8.verb_diversity: proporção de types de verbos em relação à quantidade de tokens de verbos.</p> <p>9.lsa_adj_mean: similaridade média entre pares de sentenças adjacentes.</p> <p>10.lsa_adj_std: desvio padrão de similaridade entre pares de sentenças adjacentes.</p> <p>11.lsa_all_mean: média de similaridade entre todos os pares de frases.</p> <p>12.lsa_all_std: Desvio padrão de similaridade entre palavras</p> <p>13.lsa_givennes_mean: média de givennes de cada sentença do texto, a partir da segunda sentença</p>

Figura 1. Recorte de 39 das 79 métricas usadas nesta pesquisa

incluídas; agrupadas naquelas relacionadas às palavras, sentenças e conexões entre sentenças. Entretanto, 17 métricas do estudo de [Graesser and McNamara 2011] não foram adaptadas para o português, seja por falta de recursos linguísticos ou ferramentas precisas de PLN. Elas são listadas aqui para futuras pesquisas: conectivos adversativos, *meaningfulness*, verbos causais, ações intencionais, eventos e partículas, similaridade sintática (sentenças no parágrafo), sobreposição de palavras de conteúdo em todas as sentenças, dissimilaridade de PoS entre sentenças e dissimilaridade de palavras entre sentenças, coesão causal, temporal e intencional, repetição de tempo e de aspecto verbal, log da frequência de palavras, sobreposição de verbo adjacente e sobreposição de verbo no modelo LSA em sentenças adjacentes. Para suprir essa falta, novas foram anexadas, como, por exemplo, Complexidade de Yngve e de Frasier, Distância de Dependência, dentre outras.

3.3. Métodos de Seleção de Features Avaliados

Foram avaliados o *Correlation-based Feature Selection* (CFS) e o *Least Absolute Shrinkage and Selection Operator* (Lasso); o CFS resultou em 34 features (Tabela 2). Foram realizados experimentos de predição usando as 34 features selecionadas pelo CFS com os classificadores Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM) e Multilayer Perceptron (MLP) (Tabela 3).

Tabela 2. Features selecionadas pelo método CFS

1	noun_ratio	12	verbs_ambiguity	23	idade_aquisicao_1_25_ratio
2	pronoun_ratio	13	yngve	24	idade_aquisicao_55_7_ratio
3	verbs	14	std_noun_phrase	25	idade_aquisicao_25_4_ratio
4	negation_ratio	15	passive_ratio	26	imageabilidade_std
5	min_cw_freq	16	concretude_25_4_ratio	29	imageabilidade_55_7_ratio
6	first_person_pronouns	17	concretude_4_55_ratio	30	sentence_length_std_deviation
7	conn_ratio	18	familiaridade_std	31	verb_diversity
8	tmp_neg_conn_ratio	19	familiaridade_4_55_ratio	32	adj_mean
9	tmp_pos_conn_ratio	20	familiaridade_55_7_ratio	33	span_mean
10	adjectives_ambiguity	21	idade_aquisicao_mean	34	content_density
11	adverbs_ambiguity	22	idade_aquisicao_std		

Tabela 3. Resultados da classificação com as features selecionadas pelo CFS

Classificador	Ensino Fundamental I	Ensino Fundamental II	Ensino Médio	Ensino Superior	F-Measure (Weighted Avg.)
SVM	85,30%	60,90%	74,80%	83,90%	0.777
MLP	80,30%	63,10%	73,30%	83,30%	0.767
Logistic Regression	86,00%	63,10%	75,10%	84,10%	0.783
RandomForest	87,10%	67,60%	76,00%	85,00%	0.798

O experimento com o Lasso teve como entrada os dados normalizados e o parâmetro alpha ajustado com 0.2. O método selecionou 8 *features* consideradas mais representativas para o conjunto de dados: proporção de pronomes, pronomes de primeira pessoa, proporção de palavras de conteúdo do texto com familiaridade entre 4 e 5.5, desvio padrão da imageabilidade, proporção de palavras de conteúdo do texto com imageabilidade entre 4 e 5.5, proporção de palavras de conteúdo do texto com imageabilidade entre 5.5 e 7, desvio padrão do comprimento da sentença e densidade de conteúdo. Nos experimentos, os melhores resultados foram com o classificador RandomForest, considerando a seleção das *features* do Lasso, com média ponderada de *F-Measure* de 69.6.

4. Avaliação Intrínseca e Extrínseca

4.1. Avaliação da Complexidade Textual no Córpus de Textos Didáticos

Para avaliar a tarefa de classificação da complexidade textual em nosso córpus, anotado com quatro etapas de ensino e com 79 métricas, quatro classificadores foram escolhidos, com base nos trabalhos relacionados, que foram revisados na Seção 2. Os algoritmos selecionados do Weka foram: SVM, MLP, LR e RF. Para a avaliação do melhor modelo, a validação cruzada foi usada com valor *10-folds*. Os melhores resultados foram do SVM, que alcançou uma média ponderada de *F-Measure* de **0.804**; o resultado do RF foi 0.794, da MLP foi de 0.698 e da LR foi de 0.802, caracterizando um empate técnico com o SVM. Comparando as previsões dos modelos com seleção de via Lasso e CFS e treinado com as 79 *features*, foi possível observar que o desempenho dos modelos com seleção de features é inferior ao do modelo com todas as features. Os resultados por nível escolar do classificador SVM com todas as *features* pode ser visto na Tabela 4. Como havia desbalanceamento da classe nível escolar, usamos o ClassBalancer [Jain et al. 2018]. Esse método reutiliza instâncias para que a soma total de pesos em todas as instâncias seja equilibrada. Desta forma, ficamos com 516,8 instâncias em cada classe.

Tabela 4. Resultados da Classificação do SVM com todas as features

	Precisão	Precisão c/ Balanceamento	Recall	Recall c/ Balanceamento	F-Measure	F-Measure c/ Balanceamento
Ensino Fund I	81.60%	85.0%	91.2%	92.6%	0.861	0.886
Ensino Fund II	69.8%	75.0%	65.50%	75.7%	0.676	0.754
Ensino Médio	80.10%	80.5%	71.80%	69.1%	0.757	0.743
Ensino Superior	83.40%	81.6%	88.50%	85.2%	0.859	0.834

4.2. Avaliação da Complexidade Textual com REA da MEC-RED

Para avaliar a utilidade e robustez do melhor modelo treinado no grande córpus descrito na Seção 3.1 em predizer a complexidade textual de REA, realizamos uma avaliação extrínseca em dois conjuntos de REA: com gêneros diferentes do modelo treinado e com mesmo gênero textual (cf. Seção 3). Utilizamos o melhor modelo de classificação que foi o SVM com 79 *features*. O *dataset 1* é composto por 60 REA de experimentos práticos e animações/simulações. Na avaliação de robustez, a média ponderada da *F-Measure* foi de 0.389. O *dataset 2* é composto por 40 textos do gênero textual didático; na avaliação de robustez, a média ponderada da *F-Measure* foi de 0.518. Em uma análise detalhada dos textos disponíveis, verificamos que eles possuem muitos erros ortográficos e a anotação da Etapa Escolar estipulada pelos autores dos materiais que disponibilizaram no MEC-RED parecia equivocada para alguns REA. Sendo este o primeiro trabalho a avaliar a Etapa de Ensino de REA, antevemos novas pesquisas para validar a predição automática. Por exemplo, fazer uma correção gramatical nos textos e utilizar REA para os quais os metadados *pessoas que favoritam o recurso*, e *escala de estrelas sejam usados* nos indique que os recursos são usados nas escolas. Há também algumas melhorias para a tarefa como inclusão de novas métricas de complexidade, como as 17 métricas do trabalho de [Graesser and McNamara 2011], citadas na Seção 3.2, que não foram incluídas neste estudo atual.

5. Conclusões e Trabalhos Futuros

Em resumo, este artigo explorou métodos automáticos para predizer o metadado etapa de ensino da plataforma MEC-RED, embora este trabalho possa ser utilizado para outras plataformas. Foi criado um grande córpus para modelar a tarefa de complexidade textual e assim avaliar o modelo com textos de gêneros textuais didáticos e outros como animação/simulação e experimento prático. A avaliação intrínseca mostrou um ótimo desempenho para o modelo treinado (*F-measure* de 0.804). Foi feita uma seleção de *features* usando 2 métodos de redução de dimensionalidade que foram comparados com o modelo treinado com todas as 79 *features*, além de usarmos um método de balanceamento de classes. Por fim, foi possível observar o impacto dos gêneros textuais na complexidade textual para predizer a Etapa Escolar. A avaliação extrínseca usando recursos da MEC-RED mostrou que a tarefa é difícil e merece ser melhor explorada. Os trabalhos futuros consistem em (i) estudar métricas linguísticas que distinguem os tipos de recursos animação/simulação, aplicativo móvel, jogos, experimento prático e software educacional, que são de especial interesse para essa pesquisa, para explorar novas *features* para os modelos e (ii) explorar novos métodos para a tarefa de complexidade textual como as arquiteturas neurais avaliadas em [Nadeem and Ostendorf 2018] para tentar mitigar os problemas de desempenho dos modelos treinados com engenharia de *features*.

Referências

- [Ahmed and Fuge 2017] Ahmed, F. and Fuge, M. (2017). Capturing winning ideas in online design communities. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1675–1687, New York, NY, USA. ACM.
- [Bethard et al. 2009] Bethard, S., Wetzer, P., Butcher, K., Martin, J. H., and Sumner, T. (2009). Automatically characterizing resource quality for educational digital libraries. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 221–230. ACM.
- [Cechinel et al. 2011] Cechinel, C., Sanchez-Alonso, S., and Garcia-Barriocanal, E. (2011). Statistical profiles of highly-rated learning objects. *Comput. Educ.*, 57(1):1255–1269.
- [Coster and Kauchak 2011] Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.
- [da Cunha 2015] da Cunha, A. L. V. (2015). Coh-metrix-dementia: análise automática de distúrbios de linguagem nas demências utilizando processamento de línguas naturais. Master's thesis, Universidade de São Paulo, ICMC - USP São Carlos.
- [Dalip et al. 2011] Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2011). Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality (JDIQ)*, 2(3):14.
- [dos Santos et al. 2017] dos Santos, L. B., Duran, M. S., Hartmann, N. S., Jr., A. C., Paetzold, G. H., and Aluísio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. *CoRR*, abs/1705.07008.
- [Fang 2016] Fang, Z. (2016). Text complexity in the us common core state standards: A linguistic critique. *Australian Journal of Language and Literacy*, 39(3):195–206.
- [Graesser and McNamara 2011] Graesser, A. C. and McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, 3(2):371–398.
- [Graesser et al. 2011] Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.
- [Hartmann et al. 2016] Hartmann, N., Cucatto, L., Brants, D., and Aluísio, S. (2016). Automatic classification of the complexity of nonfiction texts in portuguese for early school years. In *International Conference on Computational Processing of the Portuguese Language*, pages 12–24. Springer.
- [Jain et al. 2018] Jain, S., Kotsampasakou, E., and Ecker, G. F. (2018). Comparing the performance of meta-classifiers—a case study on selected imbalanced data sets relevant for prediction of liver toxicity. *Journal of computer-aided molecular design*, pages 1–8.
- [Leary et al. 2011] Leary, H., Recker, M., Walker, A., Wetzler, P., Sumner, T., and Martin, J. (2011). Automating open educational resources assessments: a machine learning

generalization study. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 283–286. ACM.

[MEC 2019a] MEC (2019a). Sobre a plataforma MEC-RED. <https://plataformaintegrada.mec.gov.br/sobre>. Acessado: 2019-01-24.

[MEC 2019b] MEC (2019b). Termos de serviços - plataforma mec-red. <https://plataformaintegrada.mec.gov.br/termos-de-uso>. Acessado em: 2019-01-24.

[Miao et al. 2016] Miao, F., Mishra, S., and McGreal, R. (2016). *Open educational resources: policy, costs, transformation*. UNESCO Publishing.

[Nadeem and Ostendorf 2018] Nadeem, F. and Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

[Scarton et al. 2010] Scarton, C., Gasperin, C., and Aluísio, S. (2010). Revisiting the readability assessment of texts in portuguese. *Advances in Artificial Intelligence – IBERAMIA - Volume 6433 of Lecture Notes in Computer Science*, pages 306–315.

[Scarton and Aluísio 2010] Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

[UNESCO. 2002] UNESCO. (2002). Forum on the impact of open courseware for higher education in developing countries:: final report.

[Vajjala and Meurers 2014] Vajjala, S. and Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297.

[Wagner Filho et al. 2016a] Wagner Filho, J. A., Wilkens, R., and Villavicencio, A. (2016a). Automatic construction of large readability corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 164–173.

[Wagner Filho et al. 2016b] Wagner Filho, J. A., Wilkens, R., Zilio, L., Idiart, M., and Villavicencio, A. (2016b). Crawling by readability level. In *International Conference on Computational Processing of the Portuguese Language*, pages 306–318. Springer.

[Wiley et al. 2014] Wiley, D., Bliss, T., and McEwen, M. (2014). Open educational resources: A review of the literature. pages 781–789.

Netspeak-BR: Um léxico sobre expressões criadas na língua portuguesa brasileira para a Internet

Rodolpho da Silva Nascimento¹,
Leonardo Ferreira dos Santos¹, Gustavo Paiva Guedes¹

¹CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracana, 229 - Rio de Janeiro - RJ - Brasil.

rodolpho.nascimento@eic.cefet-rj.br, leonardo.santos@eic.cefet-rj.br
gustavo.guedes@cefet-rj.br

Abstract. *Recognition of expressions in social networks that do not belong to a formal vocabulary helps to perform tasks of polarity identification and sentiment analysis in corpus. This paper aims to present the creation of the “Netspeak-BR” lexicon to assist in recognition of these expressions. For the creation, comments from the social networks Tumblr, Twitter, and Youtube were considered, totaling approximately 526,612 documents. After preprocessing and filtering, over 400 expressions were labeled, translating an informal term to the formal term.*

Resumo. *O reconhecimento de expressões em redes sociais que não pertencem a um vocabulário formal auxilia a execução de tarefas de identificação de polaridade e análise de sentimentos em corpus. O presente trabalho tem por objetivo apresentar a criação do léxico “Netspeak-BR” para auxiliar no reconhecimento dessas expressões. Para a criação, foram considerados comentários extraídos das redes sociais Tumblr, Twitter e Youtube, totalizando aproximadamente 526.612 documentos. Após pré-processamento e filtragem, foram rotuladas mais de 400 expressões, traduzindo um termo informal para o termo formal.*

1. Introdução

As redes sociais disponíveis na Internet simbolizam uma extensão virtual da vida de seus usuários [Sajadi et al. 2018]. Nesse cenário, o conceito de comunidade representa um espaço em que indivíduos com interesses similares se reúnem para discussões e pesquisas sobre temas variados. A comunicação apresentada nas redes sociais apresenta uma diversidade de transformações, desde a estruturação de sentenças até termos e expressões. Uma melhor compreensão do vocabulário empregado em redes sociais, se apresenta como essencial para diversas tarefas processamento de linguagem natural (PLN), dentre elas, a análise de sentimentos (AS).

Na Internet, os internautas desfrutam de um alto grau de liberdade, visitando tantos sites quanto possível, baixando músicas, expressando opiniões sobre assuntos de diversos interesses, dentre outros. Durante este curso de comunicação, as pessoas se sentem mais confortáveis e livres para se expressarem de uma maneira mais flexível e irrestrita com um único propósito, que é o da comunicação em si [Leng 2012].

A utilização de palavras informais, abreviações ou simbologia demonstra ser uma forma de interação social entre os usuários, possibilitando maior transparência na

comunicação [Komesu and Tenani 2009]. Alguns estudos buscam atenuar estes ruídos no processo de normalização dos dados [Duran et al. 2014], traduzindo termos não identificados em termos conhecidos [Schlippe et al. 2010] [Bard et al. 2017]. Tal atividade torna-se importante em tarefas de pré-processamento, principalmente em textos provenientes de redes sociais e fóruns de Internet. Dessa maneira, é importante buscar o aprimoramento e a efetividade em tarefas de pré-processamento de textos na web dada a quantidade de documentos em formato digital [Guimarães et al. 2015].

Na área da Computação Afetiva (CA), os termos inseridos em um texto são de grande importância para reconhecimento de emoções [Picard 2000]. Por exemplo, pode-se destacar o crescente número de léxicos afetivos criados para a língua portuguesa [Cruz et al. 2017]. No entanto, um termo inserido fora das normas ortográficas, porém amplamente utilizado por usuários de Internet, pode gerar ruídos em tarefas que objetivam inferir emoções (e.g., *scrr = socorro; pfv = por favor*). Com isso, é de grande importância a obtenção de artifícios que ajudem a minimizar tais ruídos.

A utilização de um léxico contendo um mapeamento de termos exclusivos de Internet seria de grande contribuição para a CA, traduzindo os termos para a forma reconhecida pelas normas ortográficas. Entretanto, para o melhor do nosso conhecimento em tarefas de pré-processamento, o uso de léxicos em português-BR contendo termos informais exclusivamente utilizados por usuários de redes sociais e fóruns de Internet, contendo o mapeamento para termos reconhecidos pelas normas ortográficas, não foi encontrado.

O presente trabalho tem por objetivo apresentar a criação do léxico “Netspeak-BR”, que consiste na análise de 526.612 comentários na língua portuguesa coletadas nas redes sociais Tumblr, Twitter e Youtube. O Netspeak-BR se propõe em ser um léxico para tradução de termos informais para termos formais, se apresentando como uma alternativa em tarefas de pré-processamento de textos de Internet.

O restante desse trabalho é organizado em mais 4 seções: na seção 2 é descrita as etapas de pré-processamento realizado nos comentários; na seção 3 é apresentada a metodologia para identificação das expressões e montagem do conjunto de dados; a seção 4 descreve as características preliminares do léxico e, por fim, na seção 5, são apresentadas as considerações a respeito do presente trabalho e propostas para trabalhos futuros.

2. Pré-processamento

Para este estudo, foram coletados dados publicados em três redes sociais: Tumblr, Twitter e Youtube. Um *script* desenvolvido em Python foi utilizado com a responsabilidade de consumir as APIs¹ disponibilizadas pelas redes sociais para extração de conteúdo em português-BR. Com o objetivo de atingir um público-alvo amplo, foi utilizado como *string* de busca o termo “futebol”, para compor um *corpus* com diversos estilos linguísticos. Como resultado, foram gerados 2.328 registros do Tumblr, 21.659 registros do Twitter e 519.519 do Youtube, descartando textos repetidos e unificando-os em um único *dataset* contendo 526.612 registros, formando o conjunto de dados inicial para o estudo.

Para tratamento dos dados, em cada comentário foi efetuado o *case folding* (conversão para letras minúsculas), bem como a remoção de pontuações e links por meio de expressão regular. Devido a abrangência da *string* de busca, alguns comentários

¹Application Programming Interface

apresentaram-se em outras línguas (e.g., espanhol) e, portanto, foram descartados para o experimento.

Um dicionário na língua portuguesa do Brasil foi utilizado como apoio neste trabalho. Contendo 320.139 termos, este dicionário foi extraído do corretor ortográfico LibreOffice² e foi indexado de duas formas: a primeira em tabela de dispersão (*hashtable*) e a segunda em 23 blocos, referente às letras do alfabeto. Os blocos foram agrupados por palavras contendo a mesma letra inicial, representando o agrupamento das palavras por cada letra (e.g., A: abacate, abacaxi, abafrão; B: babá, bacharel, bactério).

3. Montagem do Netspeak-BR

Para este estudo, foi utilizado um computador com processador Intel Core i3-3230 3.3GHz, 4 cores, 16Gb RAM, e sistema operacional Windows 10 64 bits. O conjunto de dados foi dividido em 4 partes iguais, referente ao número de *cores* e o processamento foi distribuído em *threads*, sendo executado de forma paralela. No final, os dados foram sincronizados gerando como resultado, um *dataset* de 3.308 registros, em um processo que durou 23 horas para finalizar.

No processamento dos dados, os textos foram normalizados e *tokenizados*, sendo representados no modelo espaço vetorial (*Bag Of Words*). Para cada termo, foi verificada a sua existência no dicionário. Caso o termo não fosse reconhecido neste dicionário, o processo de lematização era acionado e, em seguida, o termo era novamente checado no dicionário. O objetivo era descartar conjugações verbais, plurais e outros termos não incorporados no dicionário indexado. Como ferramenta de apoio para a lematização, foi utilizada a biblioteca Spacy³.

Caso o termo ainda não fosse identificado pelo dicionário, o processo de correção ortográfica era ativado, em que este considerava como “correta” a primeira letra do termo desconhecido. Em seguida, iniciava-se a varredura de todas as palavras do bloco no dicionário contendo a respectiva letra inicial, computando a similaridade entre as palavras por meio do método Jaro-Winkler [Winkler 1999]. As palavras pré-selecionadas foram aquelas que pontuaram no mínimo 80%, e para classificação final, ao término da varredura do bloco, a palavra com maior pontuação era adotada como “palavra corrigida”.

Se após todas essas etapas o termo ainda não fosse reconhecido no dicionário, o mesmo era anotado e sua frequência era computada, na medida em que fosse apresentando-se no *dataset*. Neste ponto, a *thread* sincronizava o resultado com o processo principal, acessando uma região crítica controlada por um semáforo⁴. Ao término das execuções de todas as *threads*, um arquivo CSV contendo os termos desconhecidos e suas frequências foi gerado, totalizando 3.308 termos.

Em seguida, ocorreu o processo de análise humana. Termos com menos de 10 ocorrências eram desconsiderados, pois o objetivo não era extrair termos informais mencionados esporadicamente, mas aqueles com maiores frequências, indicando ampla utilização pelos usuários. Entidades nomeadas foram identificadas manualmente e removidas em seguida. Erros ortográficos não corrigidos pelo processo, por não atender ao

²https://cgit.freedesktop.org/libreoffice/dictionaries/plain/pt_BR/pt_BR.dic

³<https://spacy.io/>

⁴Recurso da computação paralela que mantém a integridade de escrita em variáveis compartilhadas acessadas por sub-processos.

Tabela 1. Dez termos com o maior número de ocorrências

Termo	Tradução	Ocorrências
vc	você	12.367
ta	está	7.530
vcs	vocês	7.442
pq	por que	6.298
to	estou	3.659
deos	Deus	3.497
fake	mentira	3.428
mt	muito	2.620
tmb	também	2.381
mlk	moleque	2.374

critério mínimo de viabilidade (e.g., termos com a letra inicial errada *siumes* = *ciúmes*), também foram identificados pela análise humana e removidos.

Houve casos em que alguns termos apresentaram um significado duvidoso, e que fez-se necessário a realização de uma pesquisa mais criteriosa no *dataset*, buscando avaliar o contexto no qual foi inserido, chegando a um entendimento de seu significado. No processo final, foi apresentado um conjunto de dados⁵ contendo 429 termos e seus significados. A Tabela 1 ilustra 10 termos com o maior número de ocorrências.

4. Características do Netspeak-BR

O Netspeak-BR revelou um padrão de comunicação desenvolvido de forma implícita, bastante utilizado por usuários de Internet. Termos com um significado da língua portuguesa são utilizados pelos internautas objetivando significados bem diferentes (e.g., bafo=hálito; *bapho*=fofoca). Há termos criados pelo “neologismo popular” em que “*adorei*” é substituído por “*dorei*”. Houve também abreviações de termos importantes, que quando tratados, podem levar a uma melhor acurácia nos resultados de tarefas de análise de sentimentos (e.g., *plmds*=pelo amor de Deus; *pls*=por favor; *scrr**=socorro; *agt*=agitador). Os *emoticons* também foram reconhecidos e também podem revelar importantes características, em especial, símbolos utilizados para expressar emoções (e.g., :) =feliz; :(=triste).

5. Considerações finais e trabalhos futuros

O Netspeak-BR, é um léxico de apoio em tarefas de pré-processamento para análise de textos provenientes de redes sociais e fóruns de Internet em que a linguagem informal é amplamente utilizada. Seu processo de elaboração busca identificar novos termos informais utilizados na Internet, baseado em um vocabulário desenvolvido pelos próprios usuários. Para trabalhos futuros, pretendemos desenvolver um método formal para construção de léxicos de Netspeak. Objetiva-se também aprimorar o processo de identificação do termo utilizando outras *strings* de busca, com o objetivo de identificar novos termos, ampliando assim o léxico proposto.

⁵<https://github.com/LaCAfe/NetSpeak/blob/master/NetSpeak.csv>

Referências

- Bard, P. T., Luis, R. L., and Moraes, S. M. W. (2017). Normalizador de texto para língua portuguesa baseado em modelo de linguagem. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 142–150, Porto Alegre, RS, Brasil. SBC.
- Cruz, P. P., Rodrigues, R., Belloze, K., and Guedes, G. P. (2017). Uma Revisão Sistemática sobre Léxicos Afetivos para o Português do Brasil. In *XXIII Conferência Internacional sobre Informática na Educação (TISE2017)*, Fortaleza, Brazil.
- Duran, M., Avanço, L., Aluisio, S., Pardo, T., and Nunes, M. (2014). Some issues on the normalization of a corpus of products reviews in portuguese. pages 22–28.
- Guimarães, G. T., Meirose, M. V., and Moraes, S. M. W. (2015). ngramas de caractere como técnica de normalização morfológica para língua portuguesa :um estudo em categorização de textos. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 211–220, Porto Alegre, RS, Brasil. SBC.
- Komesu, F. and Tenani, L. (2009). Considerações sobre o conceito de "internetês" nos estudos da linguagem. *Linguagem em (Dis)curso*, 9:621 – 643.
- Leng, Z. (2012). A study of the features of internet language. In *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, pages 97–100.
- Picard, R. W. (2000). *Affective computing*. MIT Press.
- Sajadi, S. H., Fazli, M., and Habibi, J. (2018). The affective evolution of social norms in social networks. *IEEE Transactions on Computational Social Systems*, 5(3):727–735.
- Schlippe, T., Zhu, C., Gebhardt, J., and Schultz, T. (2010). Text normalization based on statistical machine translation and internet user support. pages 1816–1819.
- Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census.

Identificação de discurso ofensivo no Twitter nas eleições presidenciais de 2018 no Brasil

**Leonardo Ferreira dos Santos¹, Maria Clara Lippi¹,
Rodolpho da Silva Nascimento¹, Gustavo Paiva Guedes¹**

¹CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracana, 229 - Rio de Janeiro - RJ - Brasil.

leonardo.santos@eic.cefet-rj.br, mariaclara.lippi@gmail.com

rodolpho.nascimento@eic.cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. Social networks allow its users a range of benefits besides encouraging the freedom of speech but also exposes them to situations of high stress, conflicts, and arguments, facilitating the spread of offensive discourse. The present work seeks to analyze a dataset containing 1.000 tweets published during the electoral period of 2018 about presidential candidates and propose a model for identifying offensive discourse in that context. Preliminary results are promising, consistent with the present literature, raising hypotheses for future work.

Resumo. Redes sociais possibilitam a seus usuários uma série de benefícios além de incentivar a liberdade de discurso, porém também os expõem a situações de elevado estresse, conflitos e discussões divergentes, fomentando a disseminação de discurso ofensivo. O presente trabalho busca criar um conjunto de dados com 1.000 tweets publicados durante o período eleitoral de 2018 e propor um modelo para identificação de discurso ofensivo nesse contexto. Resultados preliminares se apresentam promissores, coerentes com a literatura presente, levantando hipóteses para trabalhos futuros.

1. Introdução

Nos dias de hoje, as redes sociais se tornaram uma extensão virtual da vida cotidiana de diversas pessoas em todo mundo. Incentivam a liberdade de discurso [Tsesis 2017], o que permite uma série de benefícios para os usuários, como o compartilhamento de informações, troca de opiniões e debates sobre interesses em comum. No entanto, também dificulta a existência de um consentimento sobre assuntos acerca dos membros que as compõem, possibilitando a criação de grupos com opiniões divergentes [Parsegov et al. 2017].

A comunicação em redes sociais segue o formato já difundido em aplicativos de mensagens instantâneas [Kizza and Yang 2014]. Uma das formas possíveis é a troca de mensagens de texto. Esta forma de comunicação apresenta um conjunto de características frequentes: não é estruturada [Chen et al. 2012], uso excessivo de abreviações, gírias e erros gramaticais [de Pelle and Moreira 2017] e alto grau de não-formalismo [Murakami et al. 2009]. Essas características diminuem a qualidade geral da mensagem enviada, criando um cenário favorável a falhas na comunicação, abrindo espaço

para mal-entendidos, comunicação estressante e prática do comportamento antissocial [Haythornthwaite 2005]. Todas essas características fomentam a disseminação de discursos ofensivos [Pereira 2018].

A presença de discurso ofensivo em redes sociais impacta negativamente a experiência dos usuários assim como fomenta eventos e crimes fora da Internet [Burnap and Williams 2014]. Nesse contexto, o presente trabalho apresenta duas principais contribuições: (*i*) construção de conjunto de dados anotado em Português do Brasil, composto por 1.000 *tweets* coletados durante o período da eleição presidencial de 2018 e anotados perante julgamento de 3 juízes; (*ii*) proposta de modelo de classificação de discurso ofensivo considerando textos na língua portuguesa.

As demais seções estão dispostas da seguinte maneira: na Seção 2 é apresentada a metodologia para análise do conjunto de dados; na Seção 3 são discutidos os resultados encontrados; por último, a Seção 4 apresenta as conclusões, limitações e discussão sobre trabalhos futuros.

2. Conjunto de dados

Para o presente trabalho foram extraídos 1.000 *tweets* da rede social *Twitter* durante o período eleitoral das eleições presidenciais de 2018. O processo de análise dos dados e classificação se baseou na metodologia proposta por trabalho realizado em [de Pelle and Moreira 2017]. Como critério para busca, foram filtrados *tweets* na língua portuguesa considerando a ocorrência de hashtags amplamente usadas no período eleitoral, por exemplo: “#Bolsonaro2018”, “#BoulosPresidente”, “#SomosTodosPDT”, “#OBrasilFelizDenovo”, “#Alckmin”, etc. Após a extração, os *tweets* foram analisados por 3 alunos de pós-graduação que desempenharam o papel de juízes, sem considerar um conceito único para a definição de discurso ofensivo. Para cada *tweet*, a rotulagem foi realizada com base na resposta da seguinte pergunta: “O *tweet* em questão apresenta características de discurso ofensivo?”. O resultado final pode ser observado na Tabela 1.

Juiz	Possui	Não possui
	discurso ofensivo	discurso ofensivo
1	58	942
2	311	689
3	191	809

Tabela 1. Resultado final da avaliação de três juízes com relação aos tweets extraídos da rede social Twitter.

Ao término da rotulação dos *tweets*, foi calculada a medida de Fleiss Kappa [Fleiss 1971] com o propósito de identificar o grau de concordância entre os distintos julgamentos. Para interpretação do resultado foi considerada uma escala aceita na literatura [Landis and Koch 1977] e detalhada na Tabela 2. O resultado encontrado (0,546) permite interpretar que ocorreu uma concordância moderada entre os juízes perante o teor dos *tweets*.

Para efeitos de classificação, um *tweet* foi rotulado com a presença de discurso ofensivo se pelo menos 2 dos 3 juízes assim o consideraram. Apesar da revisão das rotulações, foram considerados 165 *tweets* como pertencentes à classe “Positiva”, isto é, com a

Grau de concordância	Interpretação
Menor que 0,0	Nenhuma concordância
0,01 - 0,20	Leve concordância
0,21 - 0,40	Concordância equilibrada
0,41 - 0,60	Concordância moderada
0,61 - 0,80	Concordância substancial
0,81 - 1,00	Concordância quase perfeita

Tabela 2. Possíveis interpretações para o resultado da aplicação da métrica Fleiss Kappa.

presença de discurso ofensivo e 835 *tweets* como pertencentes à classe “Negativa”. Cada *URL* em um *tweet* faz referência a um complemento visual ou externo. Como o principal objetivo é analisar o discurso ofensivo escrito pelos usuários, essa característica foi desconsiderada. Sendo assim, todas as *URL*’s presentes no conjunto de dados foram removidas e então foi aplicada a *tokenização* do conteúdo resultante. Posteriormente, a abordagem TF-IDF foi empregada para designação dos termos mais relevantes. Nesse contexto, as *hashtags* foram mantidas, pois observou-se que, em muitas entradas, o conteúdo do *tweet* era composto apenas destas.

3. Resultados e discussão

Para a classificação do conjunto de dados gerado foi considerado o uso da ferramenta Weka¹ e os seguintes classificadores: Naïve Bayes (NB), Naïve Bayes Multinomial (NBM), Logistic Model Tree (LMT), J48 e Support Vector Machine (SVM). Todos os classificadores foram treinados considerando os parâmetros de configuração *default* e a técnica de validação cruzada com 10 partições. Nesse cenário, após o conjunto de dados ser dividido em 10 partições de igual tamanho, uma partição é eleita como o conjunto de testes enquanto as demais irão compor o conjunto de treinamento. Esse processo é repetido 10 vezes, de forma que toda partição seja considerada como um conjunto de testes do modelo. Os resultados podem ser observados na tabela 3.

Classificador	Precisão	Abrangência	F_1
NB	0,762	0,773	0,767
NBM	0,891	0,892	0,892
SVM	0,785	0,832	0,788
LMT	0,802	0,839	0,799
J48	0,731	0,829	0,761

Tabela 3. Resultado obtidos após aplicação dos algoritmos de classificação.

Ao analisar as métricas e as matrizes de confusão para cada um dos modelos de classificação, é possível observar que o desbalanceamento presente no conjunto de dados influenciou para a obtenção de resultados promissores. As tabelas 4 à 8 evidenciam o comportamento. Os resultados mais significativos foram obtidos com o algoritmo NBM. Embora também tenha sido impactado pelo desbalanceamento do conjunto de dados, o classificador foi capaz de identificar uma maior quantidade de *tweets* pertencentes à classe

¹<https://www.cs.waikato.ac.nz/ml/weka/>

Tabela 4. Matriz de confusão referente ao classificador NB.

		Predito	
		Positiva	Negativa
Esperado	Positiva	42	123
	Negativa	104	731

Tabela 5. Matriz de confusão referente ao classificador NBM.

		Predito	
		Positiva	Negativa
Esperado	Positiva	110	55
	Negativa	53	782

Tabela 6. Matriz de confusão referente ao classificador SVM.

		Predito	
		Positiva	Negativa
Esperado	Positiva	35	130
	Negativa	29	806

Tabela 7. Matriz de confusão referente ao classificador LMT.

		Predito	
		Positiva	Negativa
Esperado	Positiva	25	140
	Negativa	21	814

Tabela 8. Matriz de confusão referente ao classificador J48.

		Predito	
		Positiva	Negativa
Esperado	Positiva	4	161
	Negativa	13	822

“Positiva”, quando comparado aos demais algoritmos. No total, foram classificados corretamente 66,7% dos *tweets* ofensivos.

Os demais algoritmos não apresentaram resultados satisfatórios quando comparados ao NBM. O algoritmo Naïve Bayes, embora tenha apresentado resultados mais discretos por conta da expressiva ocorrência de falso-positivos e falso-negativos, conseguiu identificar apenas 42 *tweets* do total de 165 com a presença de discurso ofensivo. Os algoritmos baseados em árvores de decisão (LMT e J48) apresentaram alta capacidade de identificar *tweets* pertencentes à classe “Negativa” somente. O algoritmo J48 somente conseguiu classificar corretamente 4 *tweets* ofensivos. Por fim, o algoritmo SVM apresentou resultados similares aos algoritmos baseados em árvores de decisão, porém com desempenho superior discreto na identificação de *tweets* com discurso ofensivo, identificando apenas 35 corretamente do total.

4. Conclusão

O presente trabalho teve como objetivo a identificação de discurso ofensivo durante o período eleitoral de 2018 considerando a rede social *Twitter*. Para atingir tal objetivo, foram coletados 1.000 *tweets* publicados e posteriormente anotados por 3 juízes. Os resultados mais significativos foram obtidos com o algoritmo NBM e o qualificam como uma alternativa promissora para identificação de discurso ofensivo na língua portuguesa.

Como trabalhos futuros, pretende-se expandir o conjunto de dados e a participação de mais juízes. Um ponto a ser considerado é a definição de um conceito comum de discurso ofensivo como ponto de partida. Embora considerado satisfatório, o nível de concordância encontra-se abaixo de trabalhos similares na língua portuguesa. Ao considerar o vocabulário presente nos *tweets*, observa-se como oportunidade a exploração de características psicolinguísticas por meio da ferramenta LIWC [Pennebaker et al. 2007].

Referências

- [Burnap and Williams 2014] Burnap, P. and Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.
- [Chen et al. 2012] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE.
- [de Pelle and Moreira 2017] de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM) in conjunction with Congresso da Sociedade Brasileira de Computação-CSBC. Sociedade Brasileira de Computação, São Paulo, SP, Brazil*, pages 510–519.
- [Fleiss 1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- [Haythornthwaite 2005] Haythornthwaite, C. (2005). Social networks and internet connectivity effects. *Information, Community & Society*, 8(2):125–147.
- [Kizza and Yang 2014] Kizza, J. M. and Yang, L. (2014). Social history of computing and online social communities. In *Encyclopedia of Social Network Analysis and Mining*, pages 1790–1800. Springer.
- [Landis and Koch 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- [Murakami et al. 2009] Murakami, A., Matsuzawa, H., and Nasukawa, T. (2009). Candidate synonym support device for generating candidate synonyms that can handle abbreviations, misspellings, and the like. US Patent 7,483,829.
- [Parsegov et al. 2017] Parsegov, S. E., Proskurnikov, A. V., Tempo, R., and Friedkin, N. E. (2017). Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285.
- [Pennebaker et al. 2007] Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. Austin, TX: liwc. net, 135.
- [Pereira 2018] Pereira, V. G. (2018). *Using supervised machine learning and sentiment analysis techniques to predict homophobia in portuguese tweets*. PhD thesis.
- [Tsesis 2017] Tsesis, A. (2017). Terrorist incitement on the internet. *Fordham Law Review*, 86(2):367.

Sistemas de Recomendação e Geração de Receitas Através da Categorização Ontológica dos Ingredientes

Luciano S. D. Pacífico¹, Emília G. Oliveira¹, Larissa F. S. Britto¹, Teresa B. Ludermir²

¹Departamento de Computação (DC)
Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

²Centro de Informática – CIn
Universidade Federal de Pernambuco(UFPE) – Recife, PE – Brasil

{emilia.galdino,larissa.feliciana,luciano.pacifico}@ufrpe.br, tbl@cin.ufpe.br

Abstract. *Automatic recipe recommendation systems are important tools, both in terms of providing personalized recipes for users who have food restrictions, such as allergic and intolerant people, as well as in helping reducing food waste, once such systems would assist their users towards the right use leftover food and ingredients. This work proposes new approaches to the development of automatic recipe recommendation and generation systems based on the categorization of ingredients into ontologies. The systems to be developed will use nutritional information regarding the ingredients, as well as information related to the flavors resulting from their combination, as part of the decision process to perform better recommendations.*

Resumo. *Sistemas de recomendação automática de receitas são ferramentas importantes, tanto em termos de fornecerem receitas personalizadas para usuários que possuem algum tipo de restrição alimentar, como pessoas com alergias e intolerâncias, quanto no auxílio à redução do desperdício de alimentos, dado que tais sistemas poderiam auxiliar seus usuários no uso correto de sobras de alimentos e ingredientes. Este trabalho propõe novas abordagens para o desenvolvimento de sistemas automáticos de recomendação e geração de receitas baseados na categorização ontológica dos ingredientes. Os sistemas a serem desenvolvidos farão uso de informações nutricionais relacionadas aos ingredientes, da mesma forma que farão uso de informações relacionadas aos sabores resultantes da combinação desses ingredientes, como parte de seus processos decisórios para a elaboração de melhores recomendações.*

1. Introdução

O compartilhamento massivo de receitas, feito através de meios de comunicação como páginas web e comunidades online, tem facilitado a busca por refeições das mais diversas categorias culinárias do mundo, auxiliando seus usuários no processo decisório sobre a escolha do alimento desejado em uma determinada ocasião. Porém, esta grande disponibilidade de informação faz com que a busca por uma receita específica, dentre tantas opções, seja uma tarefa difícil de ser realizada. Além disso, os respositórios existentes possuem poucas receitas que atendam grupos que possuam alguma restrição alimentar (como alérgicos, intolerantes, vegetarianos, veganos, etc.).

Visando a resolução desses problemas, sistemas automáticos de recomendação de receitas têm se mostrado de grande importância por possibilitarem uma busca mais personalizada aos seus usuários, considerando aspectos como tipo de refeição (café da manhã, almoço ou jantar), ingredientes disponíveis e valor nutricional dos alimentos, tendo tais aspectos influência direta no tipo de retorno oferecido por tais sistemas. Além da praticidade, esses sistemas também funcionam como ferramentas úteis na redução do problema do desperdício de alimentos, uma vez que receitas podem ser recomendadas a partir do fornecimento de uma lista dos ingredientes previamente disponíveis em um dado momento. No intuito da minimização do problema da escassez de receitas disponíveis para públicos que apresentam algum tipo de dieta restritiva, a geração automática de receitas torna-se uma opção viável, tendo sido adotada em vários trabalhos da literatura de recomendação de receitas [Ooi et al. 2015, Lo et al. 2015, Nirmal et al. 2018]. Para uma geração precisa de novas receitas, faz-se necessário o uso de informações detalhadas relacionadas aos ingredientes que a compõem, de modo que receitas saborosas e que apresentem um valor nutricional balanceado sejam elaboradas pelos sistemas automáticos. Uma forma de agregar tais conhecimentos e relações entre os ingredientes aos sistemas automáticos de recomendação e geração de receitas é pela elaboração de ontologias desses ingredientes. Uma ontologia é uma forma de definição de informação baseada no mapeamento hierárquico entre elementos contidos em um mesmo domínio, assim como no mapeamento das relacionamentos existentes entre estes elementos. Para a elaboração de uma ontologia de ingredientes, faz-se necessário o cruzamento de vários tipos de informações relacionadas aos mesmos, tais como informações sobre seus valores nutricionais, composições químicas, categoria alimentar, dentre outras.

Neste trabalho será apresentada uma proposta de sistema de recomendação e geração de receitas, baseado na análise da categorização ontológica de seus ingredientes. Nesse contexto, ontologias relacionadas aos ingredientes serão desenvolvidas através do cruzamento de informações contidas em diferentes bases de dados, tendo em vista que não existe um repositório único que contenha todas as informações úteis sobre os ingredientes que possam vir a ser utilizadas em sistemas automáticos. O trabalho está organizado da seguinte forma: trabalhos relacionados serão apresentados brevemente na Seção 2; em seguida, os esforços a serem realizados no desenvolvimento dos sistemas de geração automática de receitas serão descritos na Seção 3; as conclusões serão apresentadas na Seção 4.

2. Uma Breve Revisão da Literatura

Sistemas de recomendação são ferramentas computacionais que visam o direcionamento resultados de buscas, de acordo com as preferências de um determinado público. Um sistema de recomendação se baseia na filtragem dos dados presentes em um determinado domínio, tendo como objetivo o retorno apenas de informações que sejam consideradas úteis em um determinado contexto. Tais sistemas têm sido aplicados largamente no contexto de recomendação de receitas [Ooi et al. 2015, Ge et al. 2015, Harvey and Elsweiler 2015, Elsweiler et al. 2017, Nirmal et al. 2018].

Na literatura relacionada à recomendação de receitas é possível encontrar exemplos de ontologias sendo utilizadas de formas distintas, seja como uma forma de remover ambiguidades presentes nos ingredientes [Shino et al. 2016], como também para realizar o mapeamento de relações entre receitas e informações nutricionais [Ting et al. 2014].

O cruzamento de informações sobre os ingredientes também é realizado no trabalho de Nirmal et al. [Nirmal et al. 2018] (embora não haja uso direto de ontologias). Nesse trabalho, os autores propõem um sistema de filtragem para a geração de receitas pela análise das relações entre seus ingredientes, obtido através do cruzamento de bases de dados de informações nutricionais e composição química dos sabores desses ingredientes. Outras fontes da literatura fizeram uso de aprendizagem de máquina na tarefa de recomendação de receitas, tanto com o objetivo de identificar a relação entre ingredientes e estilos culinários [Jayaraman et al. 2017], quanto para encontrar ingredientes que possam representar substituições mais adequadas em níveis de valores nutricionais para uma determinada receita [Gorbonos et al. 2018].

3. Metodologia

Nesta seção serão descritos os processos necessários para a categorização dos ingredientes e criação das ontologias, assim como também as etapas de desenvolvimento do sistema de recomendação e geração de receitas a ser elaborado. O processo de recomendação e geração de receitas se divide em várias etapas, como é mostrado na Figura 1. As principais etapas do desenvolvimento do sistema serão brevemente descritas nas próximas seções (Seção 3.1, Seção 3.2 e Seção 3.3).

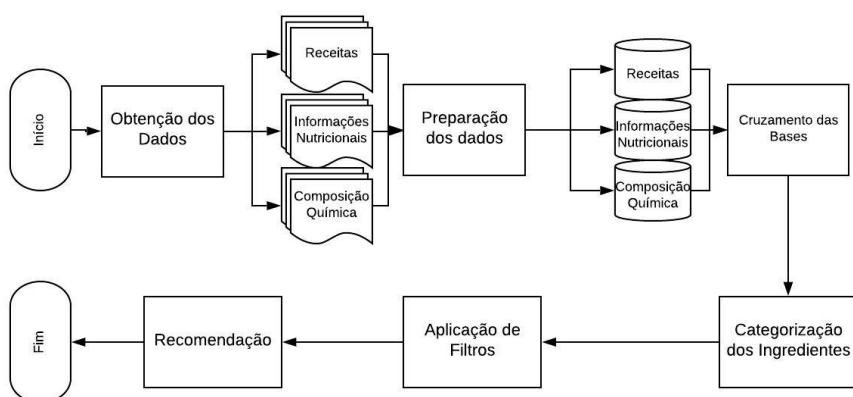


Figura 1. Etapas para a recomendação de receitas

3.1. Categorização de Ingredientes

A categorização dos ingredientes é uma etapa importante do sistema, pois é aplicada diretamente no processo de substituição de ingredientes para a geração de receitas, assim como na busca por correlações entre diferentes ingredientes. O tratamento já realizado para a categorização dos ingredientes extraídos das receitas que constituem a base atual do projeto se baseia no seguinte critério: as categorias de ingredientes foram divididas de forma a separar os grupos de ingredientes alérgenos em tipos diferentes, de modo que os grupos mais comuns de alérgenos, como leite, amendoim, crustáceos, trigo e peixes, são separados de outras categorias. Esta forma de elaboração de categorias possibilita o encontro com mais facilidade das receitas que contém um ou mais alérgenos, facilitando tanto o processo de recomendação de receitas quanto de substituição desses ingredientes. Além do critério descrito anteriormente, também foram considerados aspectos socioculturais, como a existência de dietas restritivas (como vegetarianos e veagnos), e também

religiões que não consomem bebidas alcoólicas. Visando a geração mais precisa de novas receitas pela substituição de ingredientes, pretende-se realizar a substituição do sistema de categorização atual pela elaboração de ontologias relacionadas aos ingredientes (vide Seção 3.2).

3.2. Ontologias

No contexto deste trabalho, as ontologias serão usadas como uma ferramenta para traçar relações entre os ingredientes de modo que seja possível fazer substituições em receitas sem que características importantes dessas receitas (como sabor e valor nutricional) sejam perdidos. A importância das ontologias está no fato de permitirem uma análise mais aprofundada das semelhanças e diferenças entre os ingredientes, fazendo uso de informações como a hierarquia estabelecida entre os mesmos, de forma que melhores soluções para problemas de combinação de ingredientes em receitas sejam elaboradas, algo que não poderia ser obtido apenas com a categorização simples dos ingredientes. Para que os ingredientes possam ser categorizados de forma ontológica, informações como composição química e as valores nutricionais serão necessárias, e para que estas informações possam ser usadas, será preciso fazer o cruzamento de diferentes bases de dados. Este processo irá permitir que sejam geradas receitas que tenham uma combinação de ingredientes com um sabor agradável e não tóxica, por exemplo [Nirmal et al. 2018]. Tanto a etapa de cruzamento de bases de dados, quanto a etapa de categorização dos ingredientes serão usados para a construção das ontologias.

3.3. Recomendação e Geração de Receitas

Os sistemas automáticos de recomendação e geração de receitas a serem desenvolvidos serão elaborados através da compreensão das propriedades inerentes aos ingredientes que as compõem (tais como suas composições químicas, categorização hierárquica e informações nutricionais), assim como pela análise das relações entre os ingredientes presentes em cada uma das receitas. A categorização ontológica dos ingredientes será usada na elaboração de métodos de filtragem, baseados na análise tanto das relações dos ingredientes com suas receitas, das correlações entre esses ingredientes em determinadas categorias culinárias, assim como também da análise da proporção desses ingredientes em cada receita. O sistema também agritará métodos de filtragem baseados na análise do histórico do seus usuários [Mokdara et al. 2018]. Atualmente, o sistema de geração está sendo implementado através da filtragem pela análise da probabilidade de coocorrência entre pares de ingredientes em receitas, pela análise da categoria dos ingredientes candidatos e ingrediente a ser substituído, assim como pela análise da similaridade das receitas.

4. Conclusão

Neste artigo foram descritos os processos que serão necessários para a implementação de um sistema de recomendação e geração de receitas, que irá se basear na categorização ontológica dos ingredientes. A aquisição da base de dados e o pré-processamento das receitas foram concluídos, assim como a categorização simples dos ingredientes. Alguns métodos de filtragem obtidos pela análise da literatura já foram implementados, porém a elaboração da ontologia dos ingredientes ainda está em andamento. Novos métodos de filtragem serão desenvolvidos ao término da elaboração da ontologia. Com isso, espera-se uma geração de receitas mais precisa e compatível com as expectativas dos possíveis usuários do sistema a ser proposto.

Referências

- Elsweiler, D., Trattner, C., and Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 575–584. ACM.
- Ge, M., Ricci, F., and Massimo, D. (2015). Health-aware food recommender system. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 333–334. ACM.
- Gorbonos, E., Liu, Y., and Hoàng, C. T. (2018). Nutrec: Nutrition oriented online recipe recommender. pages 25–32.
- Harvey, M. and Elsweiler, D. (2015). Automated recommendation of healthy, personalised meal plans. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 327–328. ACM.
- Jayaraman, S., Choudhury, T., and Kumar, P. (2017). Analysis of classification models based on cuisine prediction using machine learning. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 1485–1490.
- Lo, Y.-W., Zhao, Q., Ting, Y.-H., and Chen, R.-C. (2015). Automatic generation and recommendation of recipes based on outlier analysis. In *2015 IEEE 7th International Conference on Awareness Science and Technology (iCAST)*, pages 216–221. IEEE.
- Mokdara, T., Pusawiro, P., and Harnsomburana, J. (2018). Personalized food recommendation using deep neural network. In *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*, pages 1–4. IEEE.
- Nirmal, I., Caldera, A., and Bandara, R. D. (2018). Optimization framework for flavour and nutrition balanced recipe: A data driven approach. In *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–9. IEEE.
- Ooi, A., Iiba, T., and Takano, K. (2015). Ingredient substitute recommendation for allergy-safe cooking based on food context. In *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 444–449. IEEE.
- Shino, N., Yamanishi, R., and Fukumoto, J. (2016). Recommendation system for alternative-ingredients based on co-occurrence relation on recipe database and the ingredient category. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 173–178.
- Ting, Y.-H., Zhao, Q., and Chen, R.-C. (2014). Dietary recommendation based on recipe ontology. In *2014 IEEE 6th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6. IEEE.

Um Benchmark para Sistemas de Extração de Informação Aberta em Português

**Florencia Mara Malenchini, Daniel Vitor Oliveira Rodrigues,
Rafael Glauber, Daniela Barreiro Claro, Marlo Souza**

¹Formalismos e Aplicações Semânticas (FORMAS), LaSiD,
Departamento de Ciência da Computação – IME – Universidade Federal da Bahia,
Av. Adhemar de Barros, s/n, Ondina, Salvador - Bahia - Brasil

{florenciamalenchini, danielvitorr}@gmail.com

r glauber@dcc.ufba.br, {dclaro, msouza1}@ufba.br

Abstract. *The evaluation of open information extraction (OIE) systems is a difficult task. The intrinsic assessments performed in the fieldwork do not allow a direct comparison between the results of these studies. Also, measuring usefulness and quality of the extracted facts is an open problem in the area. In this work, a benchmark was built to evaluate this type of system for the Portuguese language. A fact-based question and answering system (QAS) was used to assess the usefulness of OIE systems. For this, a QAS was adapted to Portuguese, and a dataset was reviewed for this evaluation. The results presented in this study are in agreement with data presented in previous studies. In the end, the DptOIE system, based on dependency analysis, presented the best results in precision and recall.*

Resumo. *A avaliação de sistemas de Extração de Informação Aberta (OIE – do Inglês Open Information Extraction) é uma tarefa difícil. As avaliações intrínsecas realizadas nos trabalhos da área não permitem uma comparação direta entre os resultados destes estudos. Além disso, medir a utilidade deste tipo de sistema e a qualidade dos fatos extraídos é um problema em aberto na área. Neste trabalho foi construído um benchmark para avaliação deste tipo de sistema para a língua portuguesa. A partir de um Sistema de Pergunta e Resposta (QAS – do Inglês Question and Answering System) baseado em fatos extraídos de textos foi verificada a utilidade de sistemas de OIE. Para isso, um QAS foi adaptado para Português. Adicionalmente, foi realizada a revisão de um conjunto de dados para esta avaliação. Os resultados apresentados neste trabalho estão em acordo com dados apresentados em trabalhos anteriores. Ao final, o sistema DptOIE, baseado em Análise de Dependência, apresentou os melhores resultados de precisão e recall.*

1. Introdução

A Extração de Informação Aberta (OIE – do Inglês *Open Information Extraction*) é um paradigma que permite a descoberta de fatos em um corpus grande e heterogêneo de documentos [Banko et al. 2007]. Os sistemas de OIE não necessitam de pré-especificação das relações o que permite a execução da tarefa sem treinamento específico para algum domínio [Del Corro and Gemulla 2013]. Sistemas de OIE são úteis para diversas

aplicações, tais como: Aquisição de Conhecimento de Senso Comum [Lin et al. 2010], Reconhecimento de Inferência Textual [Berant et al. 2011, Angeli et al. 2015] e Sistema de Pergunta e Resposta (QAS – do Inglês *Question and Answering System*) [Banko et al. 2007, Yao and Van Durme 2014].

Segundo [Glauber and Claro 2018] as avaliações de sistemas de OIE se concentram na abordagem intrínseca por meio da verificação manual de uma porção de fatos extraídos pelos sistemas. Normalmente, essas avaliações são realizadas por duas métricas: número de fatos extraídos e precisão de extração [Schmitz et al. 2012]. Essa verificação manual da precisão pode estar sujeita a viés. Além disso, ainda não há um consenso em relação às métricas e o método ideal para avaliar ou comparar os resultados de diferentes sistemas de OIE. Outro problema apontado pelos autores em [Glauber and Claro 2018] é que uma parte considerável dos fatos extraídos, apesar de corretos¹, possuem baixa informatividade. Ou seja, não apresentam informações úteis para alguma solução computacional. A razão principal deste problema é a falta de contexto dos fatos extraídos, visto que muitos deles só tem significado no contexto do texto ao qual foram extraídos. Avaliar a informatividade das extrações de um sistema de OIE não é uma tarefa trivial. Como pontuado por Xavier et al. [Xavier et al. 2015], a noção de informatividade não é bem definida nos estudos sobre OIE.

Com o intuito de avaliar sistemas de OIE, alguns trabalhos na literatura preocuparam-se em criar *benchmarks* que podem ser usados para a avaliação intrínseca [Stanovsky and Dagan 2016]. Porém, a construção de tais recursos apresentam muitos desafios para esta área [Glauber et al. 2018]. Diante deste cenário, este trabalho apresenta um *benchmark* para avaliação extrínseca para sistemas de OIE em textos da língua portuguesa. Este novo recurso de avaliação foi construído sobre um QAS baseado em fatos [Léchelle and Langlais 2016] extraídos por algum sistema de OIE. No desenvolvimento deste trabalho são consideradas como as principais contribuições:

- Adaptação de um QAS em Inglês para responder perguntas na língua portuguesa.
- Revisão manual de um conjunto de dados utilizado no processo de avaliação.
- Avaliação dos sistemas de OIE do estado da arte para Português utilizando o *benchmark* desenvolvido.

O restante deste trabalho está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve o QAS adaptado para Português. A Seção 4 detalha a configuração do experimento realizado. A Seção 5 analisa os resultados dos experimentos e as conclusões são descritas na Seção 6.

2. Trabalhos Relacionados

A partir do processamento automático do conjunto de dados anotados denominado QA-SRL [He et al. 2015], os autores em [Stanovsky and Dagan 2016] construíram um *benchmark* para sistemas de OIE. O sistema gera para cada predicado anotado no QA-SRL uma tupla expressando cada elemento do produto cartesiano de respostas (excluindo os pronomes) para as perguntas sobre esse predicado. Em seguida, essas tuplas geradas são comparadas com as tuplas extraídas dos sistemas de OIE. Esta proposta apresenta uma desvantagem com origem no recurso utilizado que está restrito a predicados explícitos.

¹O fato extraído está de acordo com a sentença utilizada para realizar a extração.

Dentro da perspectiva de uma avaliação intrínseca, um possível método de avaliação pode verificar a presença dos fatos extraídos por um sistema de OIE em relação a um conjunto dourado (*golden set*). Os autores em [Glauber et al. 2018] apresentaram os desafios e as dificuldades para a construção deste tipo de recurso. Embora os resultados apresentados não indiquem a inviabilidade da construção de tais recursos, garantir a qualidade e a utilidade dos fatos extraídos durante a tarefa de anotação é um grande desafio a ser solucionado.

Os autores em [Léchelle and Langlais 2016] propõem uma nova abordagem de avaliação extrínseca de desempenho de sistemas de OIE baseada na informatividade dos fatos extraídos. A informação extraída é avaliada pela sua capacidade de responder automaticamente perguntas sobre o texto. Esta nova abordagem rejeita extrações que seriam anotadas como corretas por estarem presentes nas sentenças, mas não serem informativas, ou seja, não são úteis para responder perguntas. Ao utilizar um QAS para avaliar, o método representa uma aproximação a real necessidade dos usuários: obter informações úteis.

3. Construindo um *benchmark* para OIE em Português

O *benchmark* desenvolvido neste trabalho segue o fluxo descrito na Figura 1. A partir do banco de sentenças, um sistema de OIE é executado em uma determinada sentença para extraír fatos. Logo após, a questão relacionada a sentença é recuperada do banco de questões. Nesta fase, a correspondência lexical entre as extrações e a pergunta é realizada. O último passo é analisar se a resposta foi de fato a resposta esperada. Como parte desta implementação foi adaptado para Português o QAS desenvolvido por [Léchelle and Langlais 2016] que é limitado a textos em Inglês. Se o fato extraído contém sobreposição de mais de 60% das palavras da pergunta, excluindo-se *stopwords*², é considerado um fato correspondente e a resposta é dada pelo argumento ou descriptor da relação com a menor sobreposição com a pergunta. Uma resposta é considerada correta, se contém todas as palavras da resposta esperada no conjunto de dados.

Para utilizar a tarefa de respostas automáticas a perguntas como método de avaliação de sistemas de OIE é necessária a existência de um *golden set* de perguntas e recursos textuais que sejam aplicáveis à avaliação. Esse conjunto de dados deve satisfazer determinadas restrições, tais como a necessidade de: (i) as respostas poderem ser encontradas nos recursos textuais associados; (ii) as respostas poderem ser encontradas sem necessidade de profunda manipulação das extrações se aplicarmos um sistema de OIE usual; (iii) as perguntas possuírem diferentes naturezas (i.e. factoides, sim/não, explicação, etc.) e tratar de domínios distintos. Para atender essas necessidades foi considerado o CINTIL-QATreeBank [Gonçalves et al. 2012] como recurso base para este *benchmark*. Este *treebank* é composto por sentenças-fonte associadas a perguntas e respostas em português. O CINTIL-QATreebank é constituído de 111 perguntas de diferentes tipos gramaticais, com respostas distribuídas entre 8 classes semânticas diferentes, sendo construído de forma automatizada sobre o CINTIL-Treebank.

Em uma primeira etapa de transformação do CINTIL-QATreeBank, as perguntas foram aleatoriamente separadas em quatro grupos e distribuídas entre quatro anotadores

²Foi utilizada a lista de *stopwords* para o português da biblioteca NLTK no Python 2.7.

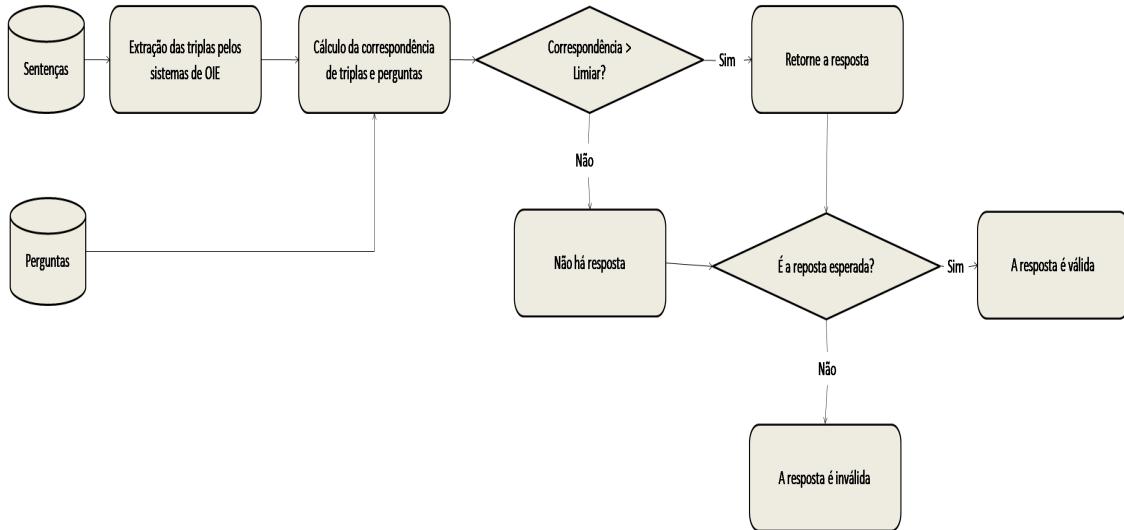


Figura 1. Fluxo do *benchmark* proposto para avaliação de sistemas de OIE adaptado de [Léchelle and Langlais 2016].

humanos. Cada anotador, dada a pergunta e uma sentença associada presente no CINTIL-QATreebank, realizou a extração de um fato em formato de tupla $t = (arg1, rel, arg2)$ [Glauber et al. 2018] que corresponde a uma informação que responde a pergunta. Após isso, foi realizada uma avaliação manual de todas as perguntas presentes no *treebank*. Do conjunto de 111 perguntas disponíveis, os anotadores identificaram que somente 102 perguntas diferiam de forma significativa em seu conteúdo, havendo variações não significativas como as perguntas 1 e 2 da Tabela 1. Das 111 perguntas foram identificadas cerca de 57 perguntas que não possuíam respostas ou a mesma não podia ser obtida por uma extração da sentença, como a pergunta 3 da Tabela 1.

Tabela 1. Exemplos de perguntas no CINTIL-QATreebank

	Sentença	Pergunta	Resposta
1	Washington acompanhou os movimentos de Saddam desde a primeira hora.	Quem acompanhou os movimentos de Saddam desde a primeira hora?	Washington
2	Washington acompanhou os movimentos de Saddam desde a primeira hora.	Quem é que acompanhou os movimentos de Saddam desde a primeira hora?	Washington
3	O consumo foi muito grande e o carro não andava.	Quanto foi o consumo?	-

O conjunto final de perguntas (denominado aqui CINTIL-QATreebank-v2) foi constituído de 45 perguntas (Q), cada uma associada a uma sentença (S) e uma extração (T) que pode ser utilizada para obter uma resposta (A), como na Figura 2.

Q: O que aplaude a Comunidade Internacional ?
S: Comunidade Internacional aplaude demissão de o líder de os sérvios de a Bósnia
T: (Comunidade Internacional, aplaude, demissão de o líder de os sérvios de a Bósnia)
A: demissão de o líder de os sérvios de a Bósnia
Q: Dentro de quanto tempo, Hong Kong voltará a ser administrada por Pequim ?
S: Dentro de um ano, Hong Kong voltará a ser administrada por Pequim
T: (Dentro de um ano, Hong Kong voltará, a ser administrada por Pequim)
A: Dentro de um ano

Figura 2. Exemplo de pergunta, resposta e tripla do *benchmark*.

4. Configuração do Experimento

Na construção deste *benchmark* a primeira avaliação é do próprio recurso em si. Ou seja, primeiro deve-se verificar o desempenho do recurso criado, antes de avaliar os sistemas de OIE do estado da arte. Nesta avaliação foram considerados somente como fonte de dados os fatos extraídos manualmente pelos anotadores. As métricas de avaliação utilizadas para avaliar o desempenho do QAS foram precisão e *recall*. A métrica precisão é calculada com base na relação de respostas corretas e expectativa de respostas corretas (Equação 1) do QAS. A métrica *recall* é calculada com base na relação de respostas corretas e respostas esperadas (Equação 2) do QAS.

$$\text{precisão} = \frac{\#(\text{respostas corretas})}{\#(\text{respostas dadas})} \quad (1)$$

$$\text{recall} = \frac{\#(\text{respostas corretas})}{\#(\text{perguntas com respostas})} \quad (2)$$

Ao usar as extrações manuais feitas pelos anotadores, o sistema obteve uma precisão de 79% e *recall* de 76%. Estes valores indicam o grau de confiança do *benchmark* em relação aos resultados que foram obtidos dos sistemas de OIE posteriormente avaliados. Ao final, foram selecionados os principais sistemas de OIE com suporte a textos escritos na língua portuguesa. Ao melhor de nosso conhecimento, os sistemas selecionados são:

- *ArgOE* [Gamallo and Garcia 2015] – 2015
- *DependentIE* [Oliveira et al. 2017] – 2017
- *DptOIE* [de Oliveira and Claro 2018] – 2018
- *PragmaticOIE* [Sena and Claro 2018] – 2018
- *InferPortOIE* [Sena and Claro 2019] – 2019

Cada sistema foi utilizado individualmente como fonte de dados para as respostas o que permitiu avaliar precisão e *recall* para cada um dos representados do estado da arte.

5. Resultados

Os sistemas de OIE foram avaliados utilizando a versão revisada do *treebank*, aqui denominado CINTIL-QATreebank-v2. Na Figura 3 estão apresentados os valores de precisão e *recall* obtidos por cada um dos sistemas.

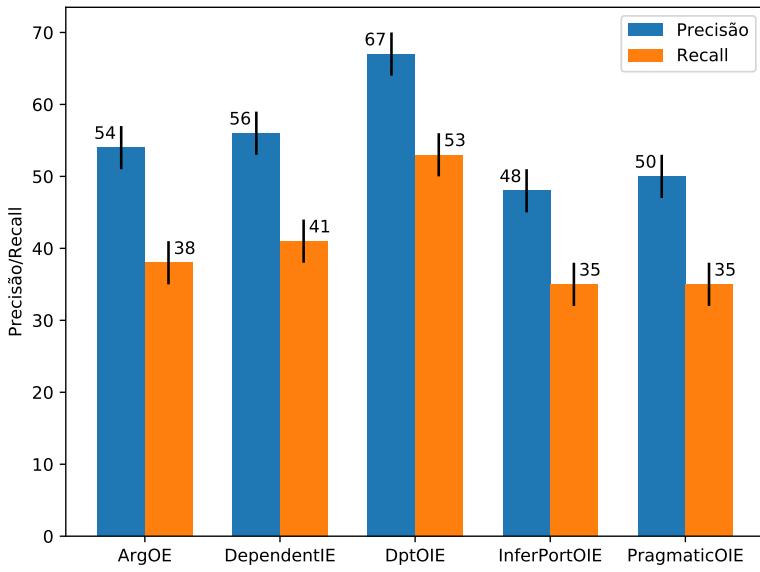


Figura 3. Resultados de precisão e recall (em percentual) para os sistemas de OIE avaliados utilizando o conjunto de dados CINTIL-QATreebank-v2.

Os três maiores valores de precisão e *recall* estão atribuídos para os sistemas *ArgOE*, *DependentIE* e *DptOIE*. Estes três sistemas se caracterizam por utilizarem Análise de Dependência (AD) [Marcheggiani and Titov 2017] na identificação de cláusulas³ e regras manuais para extrair os fatos das sentenças. Os dois sistemas pior avaliados no cenário proposto também pertencem a uma categoria específica: utilizam Análise Rasa⁴ e regras manuais para extrair fatos das sentenças. Apesar do conjunto de dados utilizado representar uma pequena amostra, estes dois grupos de resultados sugerem uma maior utilidade dos sistemas baseados em AD. A diferença percentual de 20% entre precisão e *recall* do sistema melhor avaliado (*DptOIE*) e o pior (*InferPortOIE*) caracteriza o melhor desempenho dos representantes das abordagens utilizadas nos sistemas de OIE na avaliação proposta. Porém, quando se compara os resultados do *ArgOE* (um sistema mais velho baseado em AD) e o *PragmaticOIE* esta conclusão perde força pela proximidade dos valores de precisão e *recall*.

6. Conclusão e Trabalhos Futuros

Neste trabalho, foi proposto um *benchmark* de avaliação extrínseca de sistemas de OIE do estado da arte para língua portuguesa a partir de um QAS. Apesar do método de seleção de respostas do QAS utilizado ser simplório e não tratar fenômenos complexos como compreensão de questões, o uso de tal método não se apresentou como uma limitação. Uma vez que OIE se propõe a extrair todas as possíveis informações da sentença, quanto menor a quantidade de manipulações da pergunta e das extrações realizadas pelo sistema for necessária, menor o erro do método, o que resulta num maior grau de confiança para a avaliação dos sistemas de OIE. Assim, o método desenvolvido de identificação de resposta por sobreposição lexical pareceu bem justificado para a aplicação deste trabalho.

³Os trabalhos na área como [Del Corro and Gemulla 2013] utilizam este termo para definir pedaços úteis da sentença que são utilizados como argumentos dos fatos extraídos.

⁴Algoritmos de *Part-of-speech tagger* e *Noun Phrase/Verb Phrase Chunker*.

Os experimentos iniciais de avaliação de sistemas de OIE apresentam resultados coerentes com outras avaliações para tais sistemas na literatura, a exemplo de [de Oliveira and Claro 2018] que apresenta os melhores resultados contra os demais sistemas por meio de uma avaliação intrínseca. Embora aqui seja importante salientar que o conjunto de dados criado neste trabalho é pequeno e isso pode ter influenciado nos resultados. Outra questão ainda sobre a coerência deste resultado é o fato de sistemas baseados em AD obterem resultados melhores que aqueles baseados em AR. Apesar do senso comum indicar que uma abordagem pode gerar sistemas melhores que a outra, o tamanho do conjunto de dados e a proximidade dos resultados obtidos por alguns dos representantes de cada uma das abordagens, fragilizam a confirmação desta ideia.

6.1. Trabalhos Futuros

Como trabalhos futuros, pretende-se aumentar o conjunto de dados utilizado com o intuito de aumentar a representatividade, adicionando perguntas para sentenças retiradas de textos de estilos variados (jornalísticos, artigos científicos, encyclopédias etc).

Adicionalmente, pretende-se também adaptar o *benchmark* para utilizar outros QAS. A proposta é diversificar as abordagens entre QAS de modo a analisar os sistemas de OIE em diferentes competências. Na versão atual do *benchmark*, por exemplo, não foram consideradas questões passíveis de resposta por meio de inferência, já que o QAS utilizado não suporta este recurso.

Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

- Angeli, G., Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of ACL-IJCNLP*, volume 1, pages 344–354.
- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of IJCAI*, volume 7, pages 2670–2676.
- Berant, J., Dagan, I., and Goldberger, J. (2011). Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 610–619. Association for Computational Linguistics.
- de Oliveira, L. S. and Claro, D. B. (2018). Dptoie: Um método para extração de informação anerta na língua portuguesa baseado em análise de dependência. Master's thesis, Universidade Federal da Bahia.
- Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.
- Gamallo, P. and Garcia, M. (2015). Multilingual open information extraction. In *Portuguese Conference on Artificial Intelligence*, pages 711–722. Springer.

- Glauber, R. and Claro, D. B. (2018). A systematic mapping study on open information extraction. *Expert Systems with Applications*, 112:372–387.
- Glauber, R., de Oliveira, L. S., Sena, C. F. L., Claro, D. B., and Souza, M. (2018). Challenges of an annotation task for open information extraction in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 66–76. Springer.
- Gonçalves, P., Santos, R., and Branco, A. (2012). Treebanking by sentence and tree transformation: Building a treebank to support question answering in Portuguese. In *Proceedings of LREC*, pages 1895–1901. ELRA.
- He, L., Lewis, M., and Zettlemoyer, L. (2015). Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics”.
- Léchelle, W. and Langlais, P. (2016). An informativeness approach to open ie evaluation. In *Proceedings of CICLING*. Springer, Springer.
- Lin, T., Etzioni, O., et al. (2010). Identifying functional relations in web text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1276. Association for Computational Linguistics.
- Marcheggiani, D. and Titov, I. (2017). Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of EMNLP*, pages 1506–1515. ACL.
- Oliveira, L., Glauber, R., and Claro, D. B. (2017). Dependente: An open information extraction system on portuguese by a dependence analysis. In *ENIAC - 2017 XIV Encontro Nacional de Inteligência Artificial e Computacional*.
- Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.
- Sena, C. F. L. and Claro, D. B. (2018). Pragmatic information extraction in brazilian portuguese documents. In *International Conference on Computational Processing of the Portuguese Language*, pages 46–56. Springer.
- Sena, C. F. L. and Claro, D. B. (2019). Inferportoie: A portuguese open information extraction system with inferences. *Natural Language Engineering*, 25(2):287—306.
- Stanovsky, G. and Dagan, I. (2016). Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305.
- Xavier, C. C., Lima, V. L. S., and Souza, M. (2015). Open information extraction based on lexical semantics. *Journal of the Brazilian Computer Society*, 21(4).
- Yao, X. and Van Durme, B. (2014). Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 956–966.

Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural

Sidney E. Leal¹, Vanessa M. A. Magalhães², Magali S. Duran¹, Sandra M. Aluísio¹

¹Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP

²Núcleo de Gestão da Informação e Conhecimento, Embrapa Gado de Leite
Juiz de Fora - MG

¹sidleal@gmail.com, magali.duran@uol.com.br, sandra@icmc.usp.br

²vanessa.magalhaes@embrapa.br

Abstract. Low literacy is a common problem in the Brazilian dairy sector that may undermine productivity. Hence the importance of simplifying newsletters, technical texts and instructions addressed to this public. The task of automatic evaluation of sentential complexity is new to Portuguese and allows us, for example, to identify which sentences in a text should be simplified. This paper presents a 3-step method for this task, using classical machine learning with MLP neural networks for ranking and regression. The model was trained in a public corpus of sentences collected from journalistic texts and its generalization to other scenarios was evaluated for the rural domain. We obtained accuracy of 87.80% in the ranking, root-mean-square error (RMSE) of 0.06 in the regressor and F-measure of 88.4% in the robustness test.

Resumo. A maioria dos produtores de leite possuem baixo letramento, o que prejudica seu acesso às tecnologias com vistas à melhoria das condições de trabalho, bem como ao aumento na produção e na renda. Esse é o motivo da importância de simplificar informativos e textos técnicos dirigidos a esse público. A tarefa de avaliação automática da complexidade sentencial é nova para o português e permite identificar, por exemplo, quais sentenças em um texto devem ser alvo de simplificação. Este artigo apresenta um método de 3 passos para essa tarefa, utilizando redes neurais do tipo MLP para ranqueamento e regressão. O modelo foi treinado em um córpus público de sentenças do gênero jornalístico e sua generalização para outros cenários foi avaliada para o domínio rural. Foram obtidas uma acurácia de 87,80% no ranqueamento, raiz do erro quadrático médio (ou RMSE, em inglês) de 0,06 no regressor e F-measure de 88,4% no teste de robustez.

1. Introdução

Segundo o INAF Brasil (Indicador de Alfabetismo Funcional), apenas um em cada dez brasileiros adultos é considerado letrado de forma proficiente [IPM 2018]. Esse indicador é bastante alarmante e explicita um dos grandes desafios brasileiros: o acesso à evolução econômica e tecnológica pela população. Percebe-se que a situação é ainda mais crítica quando isolamos certos setores da economia, como o da agropecuária, em que apenas 1% dos entrevistados foram considerados proficientes. Isso significa que a quase totalidade dos produtores rurais pode não ser capaz de usufruir das novas tecnologias desenvolvidas

e publicadas por entidades de pesquisa. A falta de acesso ao conhecimento prejudica bastante esse setor, um dos mais importantes do Brasil, que é responsável por 23% do Produto Interno Bruto (PIB)¹ e 40% da renda da população economicamente ativa².

Uma alternativa viável na atualidade é simplificar essas publicações, como fez a Embrapa Gado de Leite, no projeto APP@Rural [Magalhães et al. 2017], que simplificou os comunicados técnicos e partes do Manual de Bovinocultura de Leite, tornando-os mais acessíveis aos produtores, estudantes e extensionistas. A Embrapa utiliza os métodos de classificação textual e de simplificação lexical e sintática do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) [Aluisio and Gasperin 2010], com adaptação da simplificação lexical para atender aos termos técnicos do domínio rural. O PorSimples teve como objetivo promover o acesso a textos em português brasileiro (PB) por analfabetos funcionais e crianças ou adultos em fase de alfabetização e criou os modelos automáticos com base em textos jornalísticos.

O trabalho presente propõe uma evolução nesses métodos ao criar um método para indicar automaticamente as sentenças alvos de simplificação, permitindo a sua classificação nos quatro níveis indicados pelo relatório de 2018 do INAF: Proficiente, Intermediário, Elementar e Rudimentar.

Na Seção 2, são apresentados a tarefa e os principais trabalhos da área de complexidade sentencial. Na Seção 3, é apresentado o método de avaliação da complexidade proposto neste trabalho, juntamente com o córpus de sentenças alinhadas PorSimplesSENT, um resumo das *features* e o modelo de aprendizado de máquina escolhido. Finalmente, a Seção 4 mostra o teste de robustez feito para o modelo treinado com textos jornalísticos e avaliado em textos produzidos pela Embrapa Gado de Leite, chamados aqui de textos do domínio rural.

2. Avaliação da Complexidade Sentencial

A inteligibilidade³ de um texto, do inglês *text readability* é, segundo [Dubay 2007, pg.6], *a facilidade de leitura de um texto criada pela escolha de conteúdo, estilo, estruturação e organização que atende ao conhecimento prévio, habilidade de leitura, interesse e motivação da audiência*. As primeiras fórmulas de inteligibilidade foram criadas há quase um século, na década de 1920, nos Estados Unidos e consideravam que a complexidade poderia ser inferida por métricas de palavras e sentenças, baseadas na frequência e tamanho (quantidade de letras) das palavras e na média da quantidade de palavras por sentença. Desde então, a Inteligibilidade Textual tornou-se uma grande área de pesquisa multidisciplinar, com uma vasta bibliografia, e ganhou novas abordagens neste século com o uso de métodos de PLN e AM.

Tradicionalmente, a tarefa tem sido aplicada no nível textual, atribuindo uma nota (ou nível de *ranking*, de proficiência) para um documento inteiro. Porém, em um documento classificado como simples, podem ocorrer sentenças complexas, assim como existem sentenças simples em um documento complexo. Uma sentença é uma unidade

¹<http://www.agricultura.gov.br/noticias/agropecuaria-puxa-o-pib-de-2017>

²<http://www.mda.gov.br/sitemda/noticias/agricultura-familiar-do-brasil-%C3%A9-8%C2%AA-maior-produtora-de-alimentos-do-mundo>

³Neste trabalho, usamos os termos complexidade e inteligibilidade (seja no nível sentencial ou textual) como sinônimos.

importante que traz, na maioria das vezes, informação suficiente para inferência e análise da sua complexidade. Embora seja possível usar a mesma abordagem de avaliação da complexidade dos textos para o nível das sentenças, [Dell’Orletta et al. 2014] demonstraram que é necessário um número maior de *features* para a segunda tarefa.

A Tabela 1 mostra: a) uma simplificação por meio da substituição lexical, em que um termo técnico é substituído por outro mais frequente; b) uma sentença simplificada no nível sintático por meio da sua divisão em duas sentenças e c) um caso de elaboração textual, incluindo uma breve explicação de um termo técnico.

Tabela 1. Exemplos de sentenças simplificadas

a) Lexicalmente	Original Simplificada	Se acentuada e prolongada, a hipertermia pode causar a morte do animal. Se acentuada e prolongada, a febre pode causar a morte do animal.
b) Sintaticamente	Original Simplificada	O uso de forragem conservada, cujas formas mais comuns são: ensilagem e fenação, é uma solução para alimentar o rebanho. O uso de forragem conservada é uma solução para alimentar o rebanho. As formas mais comuns para conservar forragens são: ensilagem e fenação.
c) Elaboração textual	Original Simplificada	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia. A ensilagem é o processo de conservação do alimento por fermentação anaeróbia (sem a presença de ar).

A avaliação do nível de inteligibilidade de sentenças é uma tarefa de pesquisa recente e visa analisar e avaliar individualmente as sentenças de um texto, permitindo uma informação mais acurada dos pontos complexos de um texto. Essa abordagem é relevante, pois, como afirmam [Dell’Orletta et al. 2014] as abordagens de classificação de inteligibilidade que levam em consideração os textos inteiros não trazem grandes vantagens para a posterior aplicação de métodos automáticos de simplificação. Além disso, considerar como complexas todas as sentenças de um texto classificado como complexo, pode prejudicar o treinamento dos métodos, principalmente quando essas sentenças são utilizadas para avaliar a tarefa de predição da complexidade sentencial. Isso foi demonstrado por [Vajjala and Meurers 2014] durante a investigação dos motivos da baixa acurácia que obtiveram ao utilizar o córpus Wikipedia-SimpleWikipedia (sem alinhamento sentencial).

O primeiro trabalho a considerar a tarefa de complexidade especificamente para o nível sentencial foi [Dell’Orletta et al. 2011], comparando a sua dificuldade em relação ao nível textual. Porém, a definição da forma de avaliação da tarefa só foi consolidada por [Vajjala and Meurers 2016] e permitiu que os trabalhos posteriores aperfeiçoassem os resultados comparativamente. [Ambati et al. 2016] conseguiram melhorar significativamente os resultados utilizando um parser do tipo *Combinatory Categorial Grammar* (CCG), e [Gonzalez-Garduño and Søgaard 2018] chegaram no estado da arte para o inglês, utilizando métricas de rastreamento ocular aliadas às linguísticas e psico-linguísticas. [Howcroft and Demberg 2017] e [Singh et al. 2016] também publicaram trabalhos explorando a tarefa com novas métricas; o primeiro trabalho exclusivamente

com métricas psicolinguísticas e o segundo com métricas de rastreamento ocular. Uma comparação dos resultados obtidos para a tarefa de avaliação da complexidade sentencial na língua inglesa é mostrada na Tabela 2.

Tabela 2. Avaliação da tarefa na Wikipedia-SimpleWikipedia

Trabalho	Método	Acurácia
Flesch-Kincaid	Método Clássico	72,30
[Vajjala and Meurers 2016]	RankSVM	74,58
[Ambati et al. 2016]	SMO/Incr CCG	78,87
[Singh et al. 2016]	Regressão Log.	75,21
[Howcroft and Demberg 2017]	RankAsClass.	73,22
[Gonzalez-Garduño and Søgaard 2018]	MultiTask MLP	86,62

Mais recentemente, [Stajner et al. 2017] e [Scarton et al. 2018] contribuíram para a tarefa, avaliando a complexidade com o apoio do córpus Newsela (que possui 550 mil sentenças, três vezes maior que o Wikipedia-SimpleWikipedia) e [Bosco et al. 2018] obtiveram bons resultados para o italiano, utilizando Redes Neurais Recorrentes do tipo *Long Short Term Memory* (LSTM). Finalmente, [Brunato et al. 2018] contribuíram com um trabalho sobre a percepção da complexidade e concordância entre anotadores, enquanto [Timm 2018] investigou simplificações sentenciais automáticas, utilizando rastreamento ocular. Para o português (até onde foi possível verificar), foi encontrado apenas o trabalho de [Leal et al. 2018], em que foi publicado o córpus PorSimpleSent, com foco nesta tarefa.

3. Método de Avaliação da Complexidade Sentencial para o Português

3.1. Córpus

O PorSimpleSent [Leal et al. 2018] é um córpus de sentenças alinhadas, disponível publicamente, que foi compilado a partir do PorSimple [Caseli et al. 2009] e organizado a partir do alinhamento sentencial dos textos, em três níveis: a) **Original**: Sentenças originais; b) **Simplificação Natural**: Textos simplificados de forma livre pelos anotadores e c) **Simplificação Forte**: Textos simplificados seguindo as regras do manual desenvolvido no projeto.

O córpus PorSimpleSent possui três versões, com diferentes abordagens para as sentenças que sofreram operação de divisão. O PSS1 repete a sentença original para cada sentença resultante da divisão; o PSS2 seleciona apenas a maior sentença resultante, que também tenha maior sobreposição de palavras, e o PSS3 contém apenas sentenças que não sofreram divisão, sendo, portanto, o menor dos três. Para este trabalho foi escolhida a versão **PSS2**, que possui 4.962 pares de sentenças, com alinhamentos Original-Natural, Natural-Forte e Original-Forte, obtida no formato TSV (*Tab Separated Values*)⁴.

3.2. Features

Este trabalho utiliza como *features* as métricas disponibilizadas pelas ferramentas Coh-Metrix-Port, Coh-Metrix-Dementia, LIWC, AIC e as métricas psicolinguísticas disponibilizadas por [dos Santos et al. 2017], que anotaram automaticamente um banco⁵

⁴<http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources>

⁵A base está disponível em: <http://143.107.183.175:21380/portlex/index.php/en/component/content/article/2-uncategorised/23-psycholinguistic>

de 26.874 palavras do PB com Imageabilidade, Concretude, Familiaridade e Idade de Aquisição. O conjunto utilizado totaliza 189 features.

O Coh-Metrix-Port⁶ [Scarton and Aluísio 2010] é uma adaptação para o PB do Coh-Metrix, desenvolvida dentro do projeto PorSimples, e implementa 48 métricas, divididas nas categorias: contagens básicas, operadores lógicos, frequências, hiperônimos, tokens, constituintes, conectivos, ambiguidade, co-referência e anáforas. O Coh-Metrix-Dementia⁷ [Cunha 2015] é uma adaptação do Coh-Metrix-Port para análise automática de distúrbios de linguagem nas demências (como Doença de Alzheimer) ou no Comprometimento Cognitivo Leve (CCL). Ele adiciona 25 novas métricas às 48 do Coh-Metrix-Port, nas categorias: disfluências, análise de semântica latente, diversidade lexical, complexidade sintática e densidade semântica. LIWC (*Linguistic Inquiry and Word Count* - liwc.wengine.com) é uma ferramenta baseada em dicionários para análise dos vários componentes emocionais, cognitivos e linguísticos em amostras de textos, com categorias como: estatísticas comuns do texto, dimensão linguística, processos psicológicos, relatividade, assuntos pessoais e miscelânea, totalizando aproximadamente 100 métricas [Cunha 2015]. A tradução e adaptação do dicionário para o PB foi realizada em uma colaboração entre NILC, Checon Pesquisa e Unisinos no período de 2010 a 2012 e está disponível no site do projeto PortLex⁸. Também criada dentro do contexto do PorSimples [Maziero et al. 2008], a ferramenta AIC (Análise Automática de Inteligibilidade de Córpus) traz 39 métricas, com o principal diferencial de utilizar o analisador sintático PALAVRAS [Bick 2000] para o cálculo delas. Elas estão organizadas em seis classes: estatísticas do texto, voz passiva, características das orações, densidade, personalização e marcadores discursivos [Cunha 2015].

3.3. Modelo Proposto

O treinamento do modelo foi feito em três fases, ilustradas na Figura 1, na qual é feita uma analogia entre a complexidade e o espectro de cores (vermelho = mais complexo, azul = mais simples). Neste exemplo, uma sentença nunca vista antes (F3), de cor verde mais intensa que o verde da posição 4, tem o valor estimado de complexidade 4,5 (entre 4-verde e 5-amarelo).

Fase 1 - Pairwise Ranking

A primeira tarefa consistiu em treinar uma rede neural MLP (*Multi Layer Perceptron*) com 3 camadas, utilizando os 4.962 pares de sentenças do córpus PSS2 e todas as features disponíveis. Metade dos pares foi invertida, de forma a balancear melhor as duas classes: o lado complexo foi anotado com 1 e o lado simples com 0.

A camada de entrada da MLP contou com 378 neurônios (189 features x 2 sentenças); a camada oculta foi configurada com 30 neurônios, utilizando a função de ativação *ReLU*, e a saída possui apenas um neurônio, com a função *sigmoid*, predizendo 1 quando a sentença A é mais complexa que a B e 0 no inverso. O ranqueador conseguiu uma acurácia de **87,8%**, utilizando *10 fold cross validation*, um pouco superior ao estado da arte da tarefa para a língua inglesa no córpus Wikipedia-SimpleWikipedia. Utilizamos como *baseline* os resultados reportados por [Leal et al. 2018], cujo melhor modelo atin-

⁶<http://fw.nilc.icmc.usp.br:22680/>

⁷<http://fw.nilc.icmc.usp.br:22380/>

⁸<http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc>

Figura 1. Diagrama da sequência de passos do modelo

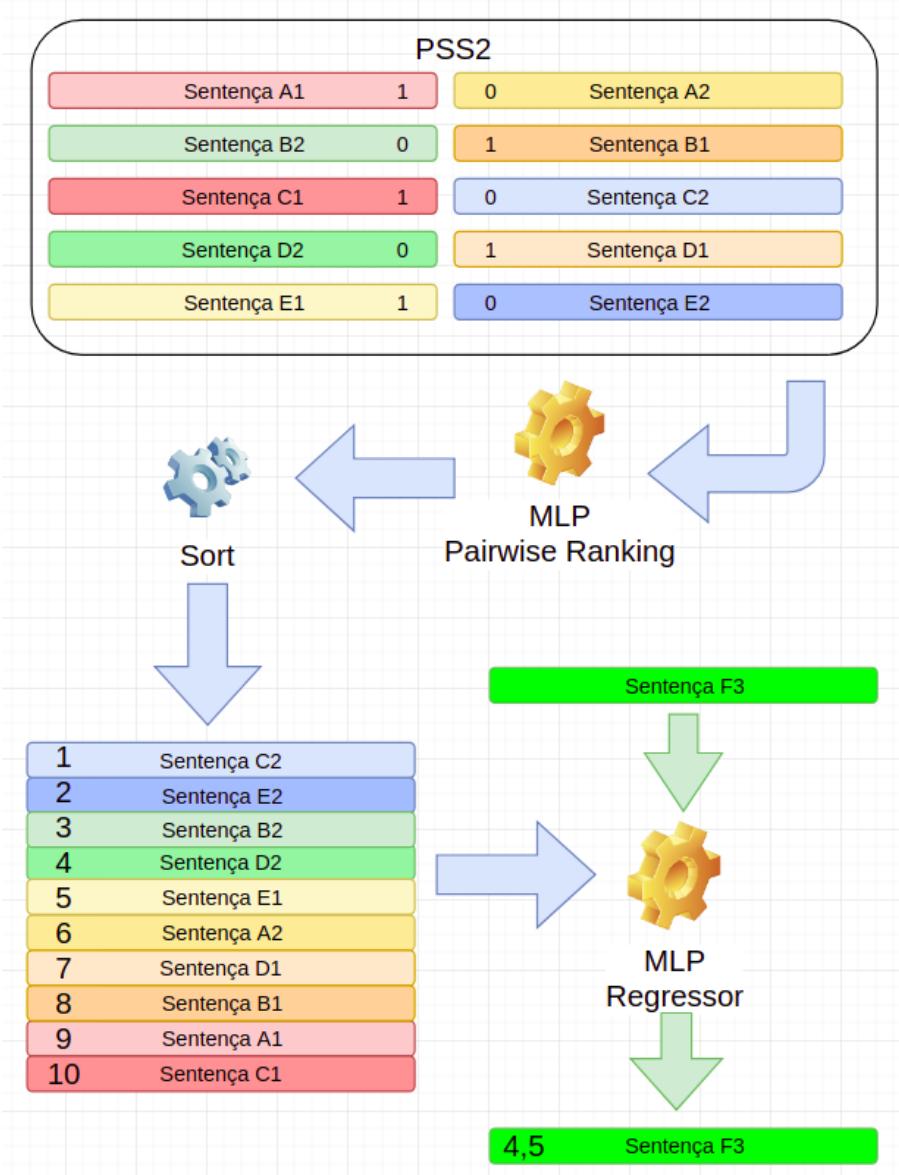


Tabela 3. Baselines e resultados do ranqueamento no PSS2

Modelo	Acurácia
Número de orações por sentença	41,28%
Média de sílabas por palavra de conteúdo	50,90%
Total de <i>tokens</i> por sentença	69,35%
SVMRank [Leal et al. 2018]	74,20%
MLP Pairwise Ranking	87,80%

giu 74,20% no mesmo córpus, conforme tabela 3, que também mostra outras 3 *baselines* simples nas primeiras linhas da tabela.

Fase 2 - *Ranking global de sentenças*

Uma vez obtido um modelo que conseguiu julgar razoavelmente bem a comple-

xidade das sentenças apresentadas em pares, ele foi utilizado para comparar todas as sentenças do PSS2, resultando em um *ranking* ordenado de 1 a 9.924. Algumas sentenças estão repetidas, quando aparecem em lados opostos do par, uma vez como a mais simples e outra vez como a mais complexa. Essas sentenças foram mantidas para uma validação adicional do resultado da ordenação, pois mesmo em lados diferentes, enquanto pares, no *ranking* global elas precisam estar próximas.

Conforme esperado, a maioria das sentenças do nível Original ficaram nas últimas posições do *ranking*, enquanto as primeiras posições foram preenchidas pelos níveis Natural e Forte. O primeiro terço do ranking ficou com 16% das sentenças originais, 30% das naturais e 55% das fortes. O último terço ficou com 52% de originais, 35% de naturais e 12% das fortes.

Fase 3 - Regressor

O *ranking* resultante da fase 2 foi normalizado entre 0 e 1 e utilizado para treinar uma segunda rede neural (MLP), também com 3 camadas. Porém, dessa vez, com apenas 189 neurônios na camada de entrada (apenas uma sentença) e predizendo a complexidade entre 0 e 1 no único neurônio de saída (utilizando ReLU), na forma de um regressor. O *dataset* foi dividido em 80% para treinamento e 20% para testes. Com todas as *features*, obtivemos uma raiz do erro quadrático médio (ou RMSE, em inglês) de 0,04 (MSE: 0,0017). Em seguida, foi aplicado o método de seleção de *features* **Permutation Importance** implementado no *eli5.sklearn*⁹ para escolher as 50 mais importantes¹⁰ e o regressor foi retreinado, obtendo uma RMSE de **0,06** (MSE: 0,0033). Foi implementada uma interface simples para validação no portal *open source* Simpligo (<https://simpligo.sidle.al>), na qual é possível entrar com um texto e conferir os valores preditos para cada sentença, numa escala de complexidade entre 0 e 100.

4. Teste de Robustez: generalização para outros gêneros de texto

O teste de robustez foi projetado para avaliar o desempenho do preditor de complexidade (em termos de *F-Measure*) em sentenças de outros gêneros diferentes do jornalístico no qual o modelo foi treinado. Foram escolhidas 500 sentenças do domínio rural conforme Tabela 4, de materiais selecionados com ajuda de uma pesquisadora da Embrapa Gado de Leite. Os materiais vieram dos gêneros instrucionais/procedimentais, técnicos e administrativos e foram agrupados para atender os quatro níveis de letramento do INAF 2018, sendo eles: rudimentar¹¹; elementar¹²; intermediário¹³ e proficiente¹⁴.

⁹https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html

¹⁰A lista completa das *features* pode ser vista junto aos códigos fonte do trabalho em <https://github.com/sidleal/simpligo-ranker>

¹¹As cartilhas podem ser acessadas a partir dos seguintes links: <https://www.infoteca.cnptia.embrapa.br/handle/doc/1035506> e <https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1055203>.

¹²Os materiais de EaD vieram do espaço e-Campo (<https://www.embrapa.br/e-campo/>)

¹³Os comunicados técnicos (COT) podem ser acessados a partir dos links: <https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1034878/1/COT77Teormatseca.pdf> e <https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/594901/1/COT28Ensilagemdomilhoedosorgo.pdf>

¹⁴O Plano Diretor: <http://www.cnpms.embrapa.br/edital/PDU.pdf>.

As sentenças foram passadas pelo regressor e classificadas em 4 níveis, de acordo com a complexidade predita, sendo 1 o mais simples e 4 o mais complexo: Nível 1 (1-25), Nível 2 (26-50), Nível 3 (51-75) e Nível 4 (76-100). Posteriormente, elas foram avaliadas pela pesquisadora da Embrapa que anotou aquelas em que discordou do nível atribuído pelo regressor, atribuindo o nível que considerava apropriado. A Tabela 5 traz os resultados, sendo a *F-Measure* média obtida de **88,4%**.

Tabela 4. Sentenças selecionadas para teste de robustez do piloto

Publicação	Quantidade de Sentenças
Cartilhas de Ensilagem Milho e Sorgo	98
Curso de EaD de Silagem Capim	97
Curso de EaD de Silagens de milho e sorgo para produção de leite	95
Comunicado Técnico sobre Matéria Seca	61
Comunicado Técnico sobre Milho e Sorgo	91
Plano Diretor para Milho e Sorgo	58
Total	500

Quanto aos erros, 59,6% foram contíguos (nos quais o nível correto é imediatamente acima ou inferior ao predito) e 40,3% foram erros distantes. Os erros mais comuns aconteceram pela presença de termos que são simples no domínio rural, mas pouco frequentes nos demais domínios, e nas sentenças com pontuação diferente (por exemplo, as terminadas em dois pontos), o que vai exigir reavaliação das métricas utilizadas, incluindo novas métricas de natureza lexical para o domínio rural. Importante salientar que esses erros não são necessariamente do regressor, mas podem ter sido introduzidos por deficiências das ferramentas das etapas anteriores.

Tabela 5. Resultados do teste robustez

Nível	Precisão (%)	Recall (%)	<i>F-Measure</i> (%)
Nível 1	98,0	76,2	85,7
Nível 2	86,0	83,2	84,6
Nível 3	89,1	94,6	91,8
Nível 4	84,6	100,0	91,6
Média	89,4	88,5	88,4

5. Considerações Finais

Este trabalho apresentou uma evolução significativa para a abordagem da tarefa de Avaliação da Complexidade Sentencial para o português brasileiro, com um incremento de mais de 10% na acurácia sobre o melhor resultado anterior reportado em [Leal et al. 2018]. Também disponibilizou uma aplicação prática para o modelo, permitindo a avaliação das sentenças de um texto, além de provar sua capacidade de generalização para outros domínios. O teste de robustez demonstrou que o modelo desenvolvido pode ser útil no apoio da avaliação e simplificação dos materiais usados pela Embrapa, mesmo esses sendo de outros gêneros textuais. A análise de erros mostrou que novas métricas simples podem ajudar a aumentar o desempenho da tarefa. Como trabalhos futuros, pretendemos investigar a tarefa utilizando métodos de *Deep Learning* e *Transfer Learning*, além da inclusão de mais *features* no modelo, em especial as de rastreamento ocular reportadas no trabalho detentor do estado da arte da tarefa para a língua inglesa.

Referências

- Aluisio, S. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: the Porsimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics*, pages 46–53.
- Ambati, B. R., Reddy, S., and Steedman, M. (2016). Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1051–1057.
- Bick, E. (2000). *The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.
- Bosco, G. L., Pilato, G., and Schicchia, D. (2018). A neural network model for the evaluation of text complexity in Italian language: a representation point of view. In *Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2018)*, pages 464–470.
- Brunato, D., Mattei, L. D., Dell’Orletta, F., Iavarone, B., and Venturi, G. (2018). Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.
- Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and Aluísio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)*, vol. 41:59–70.
- Cunha, A. L. V. (2015). *Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais*. Master’s thesis, ICMC - USP, São Carlos - SP - Brasil.
- Dell’Orletta, F., Wieling, M., Cimino, A., Venturi, G., and Montemagni, S. (2014). Assessing the readability of sentences: Which corpora and features? *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.
- Dell’Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of Italian texts with a view to text simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.
- dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., and Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In Ekštein, K. and Matoušek, V., editors, *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, volume 10415, pages 281–289. Springer, Cham.
- Dubay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, Costa Mesa, CA. ISBN: 1-4196-5439-X.
- Gonzalez-Garduño, A. V. and Søgaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5118–5124.

- Howcroft, D. M. and Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 958–968.
- IPM (2018). *INAF Brasil 2018: Indicador de Alfabetismo Funcional - Resultados Preliminares*. Instituto Paulo Montenegro. Disponível em http://acaoeducativa.org.br/wp-content/uploads/2018/08/Inaf2018_Relat%C3%B3rio-Resultados-Preliminares_v08Ago2018.pdf.
- Leal, S. E., Duran, M. S., and Aluísio, S. M. (2018). A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413.
- Magalhães, V. M. A., Bernardo, W. F., Diniz, F. H., dos Santos, K. C. L., Fonseca, L. M. G., Aluisio, S. M., and Leal, S. E. (2017). E-rural methodology: Contents elaborated according to the literacy level of the target audience. In *Twelfth Latin American Conference on Learning Technologies (LACLO)*, pages 1–9.
- Maziero, E. G., Pardo, T. A. S., and Aluísio, S. M. (2008). *Ferramenta de Análise Automática de Intelligibilidade de Córpus (AIC)*. NILC - ICMC-USP. Disponível em <http://www.nilc.icmc.usp.br-nilc/download/NILCTR0808-MazieroPardo.pdf>.
- Scarton, C. and Aluísio, S. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, pages 45–62.
- Scarton, C., Paetzold, G. H., and Specia, L. (2018). Text simplification from professionally produced corpora. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3504–3510.
- Singh, A. D., Mehta, P., Husain, S., and Rajkumar, R. (2016). Quantifying sentence complexity based on eye-tracking measures. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 202–212.
- Stajner, S., Ponzetto, S. P., and Stuckenschmidt, H. (2017). Automatic assessment of absolute sentence complexity. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4096–4102.
- Timm, L. B. (2018). *Looking at text simplification: Using eye tracking to evaluate the readability of automatically simplified sentences*. PhD thesis, Linköping University, Department of Computer and Information Science, Human-Centered systems, Linköping, Sweden.
- Vajjala, S. and Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.
- Vajjala, S. and Meurers, D. (2016). Readability-based sentence ranking for evaluating text simplification. *CoRR - Computer Research Repository*, Disponível em <http://arxiv.org/abs/1603.06009>.

Generating Sense Embeddings for Syntactic and Semantic Analogy for Portuguese

Jéssica Rodrigues da Silva¹, Helena de Medeiros Caseli¹

¹Federal University of São Carlos, Department of Computer Science

jsc.rodrigues@gmail.com, helenacaseli@ufscar.br

Abstract. Word embeddings are numerical vectors which can represent words or concepts in a low-dimensional continuous space. These vectors are able to capture useful syntactic and semantic information. The traditional approaches like Word2Vec, GloVe and FastText have a strict drawback: they produce a single vector representation per word ignoring the fact that ambiguous words can assume different meanings. In this paper we use techniques to generate sense embeddings and present the first experiments carried out for Portuguese. Our experiments show that sense vectors outperform traditional word vectors in syntactic and semantic analogy tasks, proving that the language resource generated here can improve the performance of NLP tasks in Portuguese.

1. Introduction

Any natural language (Portuguese, English, German, etc.) has ambiguities. Due to ambiguity, the same word surface form can have two or more different meanings. For example, the Portuguese word *banco* can be used to express the financial institution but also the place where we can rest our legs (a seat). Lexical ambiguities, which occur when a word has more than one possible meaning, directly impact tasks at the semantic level and solving them automatically is still a challenge in natural language processing (NLP) applications. One way to do this is through word embeddings.

Word embeddings are numerical vectors which can represent words or concepts in a low-dimensional continuous space, reducing the inherent sparsity of traditional vector-space representations [Salton et al. 1975]. These vectors are able to capture useful syntactic and semantic information, such as regularities in natural language. They are based on the distributional hypothesis, which establishes that the meaning of a word is given by its context of occurrence [Bruni et al. 2014]. The ability of embeddings to capture knowledge has been exploited in several tasks, such as Machine Translation [Mikolov et al. 2013b], Sentiment Analysis [Socher et al. 2013], Word Sense Disambiguation [Chen et al. 2014] and Language Understanding [Mesnil et al. 2013].

Although very useful in many applications, the word embeddings (word vectors), like those generated by Word2Vec [Mikolov et al. 2013a], GloVe [Pennington et al. 2014] and FastText [Bojanowski et al. 2016] have an important limitation: the Meaning Conflation Deficiency, which is the inability to discriminate among different meanings of a word. In any natural language, there are words with only one meaning (monosemous) and words with multiple meanings (ambiguous) [Camacho-Collados and Pilehvar 2018]. In word embeddings, each word is associated with only one vector representation ignoring the fact that ambiguous words can assume different meanings for which different vectors should be generated. Thus, there is a loss

of information by representing a lexical ambiguity in a single vector, since it will only contain the most commonly used meaning for the word (or that which occurs in the corpus from which the word vectors were generated).

Several works [Pina and Johansson 2014, Neelakantan et al. 2014, Wu and Giles 2015, Liu et al. 2015, Huang et al. 2012, Reisinger and Mooney 2010, Iacobacci et al. 2015] have investigated the representation of word senses instead of word occurrences in what has been called *sense embeddings* (sense vectors).

In this paper, we present the first experiments carried out to evaluate sense vectors for Portuguese. In section 2 we describe some of the approaches for generating sense vectors proposed in the literature. The approaches investigated in this paper are described in section 3. The experiments carried out for evaluating sense vectors for Portuguese are described in section 4. Section 5 finishes this paper with some conclusions and proposals for future work.

2. Related Work

[Schütze 1998] was one of the first works to identify the meaning conflation deficiency of word vectors and to propose the induction of meanings through the clustering of contexts in which an ambiguous word occurs. Then, many other works followed these ideas.

One of the first works using neural network to investigate the generation of sense vectors was [Reisinger and Mooney 2010]. The approach proposed there is divided in two phases: pre-processing and training. In the pre-processing, firstly, the context of each target word is defined as the words to the left and to the right of that target word. Then, each possible context is represented by the weighted average of the vectors of the words that compose it. These context vectors are grouped and each centroid is selected to represent the sense of the cluster. Finally, each word of the corpus is labeled with the cluster with the closest meaning to its context. After this pre-processing phase, a neural network is trained from the labeled corpus, generating the sense vectors. The model was trained in two corpora, a Wikipedia dump in English and the third English edition of Gigaword corpus. The authors obtained a correlation of Spearman of around 62.5% in WordSim-353 [Finkelstein et al. 2001]¹, for the Wikipedia and Gigaword corpus.

Another approach for generating sense vectors was [Huang et al. 2012], which extends the [Reisinger and Mooney 2010]'s approach by incorporating a global context into the generation of word vectors. According to them, aggregating information from a larger context improves the quality of vector representations of ambiguous words that have more than one possible local context. To provide the vector representation of the global context, the proposed model uses all words in the document in which the target word occurs, incorporating this representation into the local context. The authors trained the model in a Wikipedia dump (from April 2010) in English with 2 million articles and 990 million tokens. The authors obtained a Spearman correlation of 65.7% in the Stanford's Contextual Word Similarities (SCWS)², surpassing the baselines.

Based on [Huang et al. 2012], [Neelakantan et al. 2014] proposed the generation

¹WordSim-353 is a dataset with 353 pairs of English words for which similarity scores were set by humans on a scale of 1 to 10.

²The SCWS is a dataset with 2,003 word pairs in sentential contexts.

of sense vectors by performing a Skip-Gram adaptation of [Mikolov et al. 2013a]. In this approach, the identification of the senses occurs together with the training to generate the vectors, making the process efficient and scalable. This approach was the one chosen to be used in this paper and is explained in detail in the next section. The authors used the same corpus as [Huang et al. 2012] for training the sense vectors and obtained a Spearman correlation of 67.3 % also in the SCWS, surpassing the baselines.

[Trask et al. 2015] propose a different approach that uses a tagged corpus rather than a raw corpus for sense vectors generation. The authors annotated the corpus with part of speech (PoS) tags and that allowed the identification of ambiguous words from different classes. For example, this approach allow to distinguish between the noun *livro* (book) and the verb *livro* (free). After that they trained a word2vec (CBOW or Skip-Gram) model [Mikolov et al. 2013a] with the tagged corpus. The authors did not report results comparing their approach with baselines. In addition to the PoS tags, the authors also tested the ability of the method to disambiguate named entities and feelings, also labeling the corpus with these tags, before generating word embeddings. This approach was one of the chosen to be investigated in this paper and it will be explained in detail in the next section.

More recently, new proposals for language model generation like ELMo [Peters et al. 2018], OpenAI GPT [Radford et al. 2018] and BERT [Devlin et al. 2018] have begun to use more complex architectures to model context and capture the meanings of a word. The idea behind this language models is that each layer of the neural network is able to capture a different sense of the input word and generate dynamic vector representations, according to each input context. This idea of dynamic embeddings facilitates the use of these representations in downstream tasks. These architectures are complex and require very powerful hardware resources for training. The difference between sense vectors and language models like those lies in the architecture and in the way the trained model is used. Sense vectors are features that will be used for specific NLP tasks. On the other hand, the complex architecture of language models has both the neural networks that will create the language model and the NLP tasks, which can even share the same hyper-parameters (fine-tuning approach).

3. Sense embeddings

In this paper, two approaches were used for sense vectors generation: the MSSG [Neelakantan et al. 2014] and the Sense2Vec [Trask et al. 2015]. Each one is explained in the next sections.

3.1. Multiple-Sense Skip-Gram (MSSG)

In [Neelakantan et al. 2014], two methods were proposed for generating sense vectors based on the original Skip-Gram model [Mikolov et al. 2013a]: MSSG (Multiple-Sense Skip-Gram) and NP-MSSG (Non-Parametric Multiple-Sense Skip-Gram). The main difference between them is that MSSG implements a fixed amount of possible meanings for each word while NP-MSSG does this as part of its learning process.

In both methods, the vector of the context is given by the weighted average of the vectors of the words that compose it. The context vectors are grouped and associated to the words of the corpus by approximation to their context. After predicting the sense, the

gradient update is performed on the centroid of the cluster and the training continues. The training stops when vector representations have already been generated for all the words.

Different from the original skip-gram, its extensions, MSSG and NP-MSSG, learn multiple vectors for a given word. They were based on works such as [Huang et al. 2012] and [Reisinger and Mooney 2010]. In the MSSG model, each word $w \in W$ is associated to a global vector $v_g(w)$ and each sense of the word has a sense vector $v_s(w, k)$ ($k = 1, 2, \dots, K$) and a context cluster with centroid $u(w, k)$ ($k = 1, 2, \dots, K$). The K sense vectors and the global vectors are of dimension d and K is a hyperparameter.

Considering the word w_t , its context $c_t = \{w_{t-R_t}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+R_t}\}$ and the window size R_t , vector representation of the context is defined as the mean of the global vector representation of the words in the context. Let $v_{context}(c_t) = \frac{1}{2*R_t} \sum_{c \in c_t} v_g(c)$ be the vector representation of the context c_t . The global vectors of context words are used instead of their sense vectors to avoid the computational complexity associated with predicting the meanings of words in the context. It is possible, then, to predict the meaning of the word w_t , s_t , when it appears in the context c_t .

The algorithm used for building the clusters is similar to k-means. The centroid of a cluster is the mean of the vector representations of all contexts that belong to this cluster and the cosine similarity is used to measure the similarity.

In MSSG, the probability (P) that the word c is observed in the context of the word w_t ($D = 1$), given the sense and the probability that it is not observed ($D = 0$), has the addition of s_t (sense of w_t) in the formulas of the original Skip-gram (formula 1 and 2). The objective function (J) also considers (w_t, s_t) instead of just (w_t) (formula 3).

$$P(D = 1|v_s(w_t, s_t), v_g(c)) = \frac{1}{1 + e^{-v_s(w_t, s_t)^T v_g(c)}} \quad (1)$$

$$P(D = 0|v_s(w_t, s_t), v_g(c)) = 1 - P(D = 1|v_s(w_t, s_t), v_g(c)) \quad (2)$$

$$\begin{aligned} J = & \sum_{(w_t, c_t) \in D_+} \sum_{c \in c_t} \log P(D = 1|v_s(w_t, s_t), v_g(c)) + \\ & \sum_{(w_t, c'_t) \in D_-} \sum_{c' \in c'_t} \log P(D = 0|v_s(w_t, s_t), v_g(c')) \end{aligned} \quad (3)$$

After predicting the meaning of the word w_t , MSSG updates the sense vector generated for the word $w_t(v_s(w_t, s_t))$, the global vector of context words and the global vector of noisy context words selected by chance. The centroid of the context cluster s_t for the word $w_t(u(w_t, s_t))$ is updated when the context c_t is added to the cluster s_t .

In this paper, we choose to work with the MSSG fixing the amount of senses for each target word. We did that to allow a fair comparison with the second approach investigated here which a limited amount of meanings.

3.2. Sense2Vec

[Trask et al. 2015] propose the generation of sense vectors from a corpus annotated with part-of-speech (PoS) tags, making it possible to identify ambiguous words from the amount of PoS tags they receive (for example, the noun *livro* (book) in contrast with the verb *livro* (free)).

The authors suggest that annotating the corpus with PoS tags is a costless approach to identify the different context of ambiguous words with different PoS tags in each context. This approach makes it possible to create a meaningful representation for each use. The final step is to train a word2vec model (CBOW or Skip-Gram) [Mikolov et al. 2013a] with the tagged corpus, so that instead of predicting a word given neighboring words, it predicts a sense given the surrounding senses.

[Trask et al. 2015] presents experiments demonstrating the effectiveness of the method for sentiment analysis and named entity recognition (NER). For sentiment analysis, sense2vec was trained with a corpus annotated with PoS tags and adjectives with feeling tags. The word “bad” was disambiguated between positive and negative sentiment. For the negative meaning, words like “terrible”, “horrible” and “awful” appeared, while in the positive meaning there was present words like “good”, “wrong” and “funny”, indicating a more sarcastic sense of “bad”.

In the NER task, sense2vec was trained with a corpus annotated with PoS and NER tags. For example, the NE “Washington” was disambiguated between the entity categories PERSON-NOME (person’s name) and GPE (geolocation). In the PERSON-NOME category it was associated with words like “George-Washington”, “Henry-Knox” and “Philip-Schuyler” while in the GPE category the word was associated with “Washington-DC”, “Seattle” and “Maryland”.

4. Experiments and Results

In this section we present the first experiments carried out to evaluate sense vectors generated for Portuguese. As follows, we first describe the corpora used to generate sense vectors, then we present the network parameters used for training the models and, finally, we show the experiments carried out to evaluate the two approaches under investigation: MSSG and sense2vec.

4.1. Training Corpora

The corpora used for the training of sense vectors were the same as [Hartmann et al. 2017] which is composed of texts written in Brazilian Portuguese (PT-BR) and European Portuguese (PT-EU). Table 1 summarizes the information about these corpora: name, amount of tokens and types and a briefly description of the genre.

The corpora were pre-processed in order to reduce the vocabulary size. For the sense2vec model, the corpora were also PoS-tagged using the nlpnet tool [Fonseca and Rosa 2013], which is considered the state-of-art in PoS-tagging for PT-BR.

It is important to say that both approaches for generating sense vectors were trained with these corpora. The only difference is that the input for the MSSG is the sentence without any PoS tag while the input for the sense2vec is the sentence annotated with PoS tags.

Table 1. Statistics of our training corpora

Corpus	Tokens	Types	Genre
LX-Corpus [Rodrigues et al. 2016]	714,286,638	2,605,393	Mixed genres
Wikipedia	219,293,003	1,758,191	Encyclopedic
GoogleNews	160,396,456	664,320	Informative
SubIMDB-PT	129,975,149	500,302	Spoken language
G1	105,341,070	392,635	Informative
PLN-Br	31,196,395	259,762	Informative
Literacy works of public domain	23,750,521	381,697	Prose
Lacio-web [Aluísio et al. 2003]	8,962,718	196,077	Mixed genres
Portuguese e-books	1,299,008	66,706	Prose
Mundo Estranho	1,047,108	55,000	Informative
CHC	941,032	36,522	Informative
FAPESP	499,008	31,746	Science
Textbooks	96,209	11,597	Didactic
Folhinha	73,575	9,207	Informative
NILC subcorpus	32,868	4,064	Informative
Para Seu Filho Ler	21,224	3,942	Informative
SARESP	13,308	3,293	Didactic
Total	1,395,926,282	3,827,725	

4.2. Network Parameters

For all training, including baselines, we generated vectors of 300 dimensions, using the Skip-Gram model, with context window of five words. The learning rate was set to 0.025 and the minimum frequency for each word was set to 10. For the MSSG approach, the maximum number of senses per word was set to 3.

4.3. Evaluation

Based on [Hartmann et al. 2017], this experiment is a task of syntactic and semantic analogies where the use of sense vectors is evaluated. Word vectors were chosen as baselines.

Dataset. The dataset of Syntactic and Semantic Analogies of [Rodrigues et al. 2016] has analogies in Brazilian (PT-BR) and European (PT-EU) Portuguese. In syntactic analogies, we have the following categories: adjective-to-adverb, opposite, comparative, superlative, present-participle, nationality-adjective, past-tense, plural, and plural-verbs. In semantic analogies, we have the following categories: capital-common-countries, capital-world, currency, city-in-state and family. In each category, we have examples of analogies with four words:

adjective-to-adverb:

- *fantástico fantasticamente aparente aparentemente (syntactic)*
fantastic fantastically apparent apparently

capital-common-countries:

- *Berlim Alemanha Lisboa Portugal (semantic)*
Berlin Germany Lisbon Portugal

Algorithm. The algorithm receives the first three words of the analogy and aims to predict the fourth. Thus, for instance considering the previous example, the algorithm would receive Berlin (a), Germany (b) and Lisbon (c) and should predict Portugal (d). Internally, the following algebraic operation is performed between vectors:

$$v(b) + v(c) - v(a) = v(d) \quad (4)$$

Evaluation metrics. The metric used in this case is accuracy, which calculates the percentage of correctly labeled words in relation to the total amount of words in the dataset.

Discussion of results. Table 2 shows the accuracy values obtained for the syntactic and semantic analogies. The Word2vec, GloVe and FastText were adopted as word vectors baselines since they performed well in [Hartmann et al. 2017] experiments. Note that the sense vectors generated by our sense2vec model outperform the baselines at the syntactic and semantic levels.

Table 2. Accuracy values for the syntactic and semantic analogies

Embedding	PT-BR			PT-EU		
	Syntactic	Semantic	All	Syntactic	Semantic	All
Word2Vec (word)	49.4	42.5	45.9	49.5	38.9	44.3
GloVe (word)	34.7	36.7	35.7	34.9	34.0	34.4
FastText (word)	39.9	8.0	24.0	39.9	7.6	23.9
MSSG (sense)	23.0	6.6	14.9	23.0	6.3	14.7
Sense2Vec (sense)	52.4	42.6	47.6	52.6	39.5	46.2

In syntactic analogies, the sense vectors generated by sense2vec outperform the word vectors generated by word2vec in opposite, nationality-adjective, past-tense, plural and plural-verbs. An example is shown in table 3. We can explain this type of success through an algebraic operation of vectors. When calculating $v(\text{aparentemente} (\text{apparently})) + v(\text{completo} (\text{complete})) - v(\text{aparente} (\text{apparent}))$ the resulting vector of word2vec is $v(\text{incompleto} (\text{incomplete}))$ when it should be $v(\text{completamente} (\text{completely}))$. The correct option appears as the second nearest neighbor.

So, we can conclude that the sense2vec's PoS tag functions as an extra feature in the training of sense vectors, generating more accurate numerical vectors, allowing the correct result to be obtained.

Table 3. Example of syntactic analogy predicted by word2vec and sense2vec

word2vec	aparente ADJ aparentemente ADV completo ADJ : completamente ADV (expected) aparente ADJ aparentemente ADV completo ADJ : incompleto ADV (predicted)
sense2vec	aparente ADJ aparentemente ADV completo ADJ : completamente ADV (expected) aparente ADJ aparentemente ADV completo ADJ : incompleto ADV (predicted)

In semantic analogies, the sense vectors generated by sense2vec outperform the word vectors generated by word2vec in capital-world, currency and city-in-state. Examples of city-in-state are shown in table 4.

Table 4. Example of semantic analogies predicted by word2vec and sense2vec

word2vec	arlington texas akron : kansas (predicted) ohio (expected)
sense2vec	arlington N texas N akron N : ohio N (predicted)(expected)
word2vec	bakersfield califórnia madison : pensilvânia (predicted) wisconsin (expected)
sense2vec	bakersfield N califórnia N madison N : wisconsin N (predicted)(expected)
word2vec	worcester massachusetts miami : seattle (predicted) flórida (expected)
sense2vec	worcester N massachusetts N miami N : flórida N (predicted)(expected)

In this case, the PoS tag is always the same for all words: N (noun). This indicates that the success of sense2vec is related to the quality of sense vectors as a whole. As all words are tagged, this feature ends up improving the inference of all vector spaces during training.

5. Conclusion and Future Work

In this paper we used techniques to generate sense embeddings (sense vectors) for Portuguese (Brazilian and European). The generated models were evaluated through the task of syntactic and semantic analogies and the accuracy values show that the sense vectors (sense2vec) outperform the baselines of traditional word vectors (word2vec, Glove, FastText) with a similar computational cost.

Our sense-vectors and the code used in all the experiments presented in this paper are available at <https://github.com/LALIC-UFSCar/sense-vectors-analogies-pt>. The application of sense vectors in NLP tasks (WSD and others) is under development. As future work we intend to experiment a combination of the two approaches (MSSG and sense2vec) and also to explore how the new approaches proposed for generating language models perform in Portuguese.

Acknowledgements

This research is part of the MMeaning project, supported by São Paulo Research Foundation (FAPESP), grant #2016/13002-0, and was also partly funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Funding Code 001.

References

- Aluísio, S. M., Pinheiro, G. M., Finger, M., Nunes, M. G. V., and Tagnin, S. E. (2003). The LacioWeb Project: Overview and Issues in Brazilian Portuguese Corpora Creation. In *Proceedings of Corpus Linguistics*, pages 14–21.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.
- Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

- Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *CoRR*, abs/1805.04032.
- Chen, X., Liu, Z., and Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web, WWW ’01*, pages 406–414, New York, NY, USA. ACM.
- Fonseca, E. and Rosa, J. (2013). A two-step convolutional neural network approach for semantic role labeling. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.
- Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.
- Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL ’12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China. Association for Computational Linguistics.
- Liu, W., Mei, T., Zhang, Y., Che, C., and Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3707–3715.
- Mesnil, G., He, X., Deng, L., and Bengio, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations Workshop (ICLR-2013)*.
- Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceed-*

- ings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*, 12:1532–1543.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Pina, L. N. and Johansson, R. (2014). A simple and efficient method to generate word sense representations. *arXiv preprint arXiv:1412.6045*.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. URL <https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language-understanding-paper.pdf>.
- Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT ’10, pages 109–117, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Rodrigues, J., António, B., Steven, N., and João, S. (2016). LX-DSemVectors: Distributional Semantics Models for Portuguese. In *Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)*. Springer International Publishing.
- Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Trask, A., Michalak, P., and Liu, J. (2015). sense2vec - a fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388.
- Wu, Z. and Giles, C. L. (2015). Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, pages 2188–2194. AAAI Press.

Part-of-Speech Embeddings for Portuguese

Paulo L. Medeiros, Bledson Bezerra, Carlos A. Prolo, Antônio C. Thomé

¹Departamento de Informática e Matemática aplicada
Universidade Federal do Rio Grande do Norte (UFRN) – Natal– RN – Brazil

{pauloaugusto99, bledson}@ufrn.edu.br, {prolo, thome}@dimap.ufrn.br

Abstract. *Training classification models on multiple datasets is a common procedure with Deep Learning. The usual approach is to merge all datasets, mapping each of the original class sets to a single one. However, for part-of-speech (POS) tagging, where dataset annotation is a quite theoretically-subjective task, there is not always an explicit correspondence between two different tag sets. Also, it strongly depends on the tokenization assumptions made in each corpus. Along with some of the most recent deep learning techniques (Bi-LSTMs stacking, word and char embeddings, residual connections), we introduce an approach to POS tagging learning that conforms to multiple tagsets with different tokenization assumptions from different training corpora. Crucial to the approach is the introduction of the concept of continuous distributed POS representations, or POS embeddings. Even without pretraining, we achieve state-of-the-art accuracy, while building a robust versatile POS tagger. We suggest that, for downstream applications, POS embeddings can be used instead of POS tags.*

1. Introduction

Word embeddings have been successfully used to represent words and word senses in Natural Language Processing (NLP) applications. They became popular after advances in neural computing allowed for their efficient training [Bengio et al. 2003, Bengio et al. 2013, Collobert et al. 2011, Mikolov et al. 2013a, Mikolov et al. 2013b, Pennington et al. 2014, Jurafsky 2000]. The use of distributed representations made possible to insert a notion of relatedness among words, which is not implicit in their morphology and is difficult to be captured using manually extracted features. Once this kind of representation became computationally feasible and popular, all tasks in NLP started being revisited, gaining accuracy by including embeddings in preexisting algorithms.

Linguistic concepts involving taxonomies may as well be able to benefit from the idea of distributed representations. In this work we focus on part-of-speech tagging. It is intuitive and appealing the idea of words in context being classified into classes that represent their syntactic distributional properties. However it is hard to define a specific finite tagset that solves all theoretical and practical concerns that have been raised over the years. Several corpora have been built with rather distinct tagsets which are hard to map into one another [Aluísio et al. 2003, Afonso et al. 2002, Petrov et al. 2012, Marcus et al. 1993, Francis and Kucera 1979]. Moreover, each of these tagsets are based on specific tokenization assumptions, such as whether contractions are split or not.

Based on well known models for POS tagging, such as in [Plank et al. 2016] and [Ling et al. 2015], using a bidirectional long short-term memory (Bi-LSTM)

architecture [Graves and Schmidhuber 2005, Hochreiter and Schmidhuber 1997], and recent concepts in neural computing, such as word and character embeddings, Bi-LSTM stacking and residual connections [Lample et al. 2016, Dos Santos and Zadrozny 2014, Mikolov et al. 2013a, Mikolov et al. 2013b, Mikolov et al. 2018, Peters et al. 2018, He et al. 2015], we built a neural architecture for POS tagging centered in the idea of continuous distributed representations for POS. The code and all related material necessary to reproduce the experiments reported here are available at <https://github.com/pauloamed/STIL2019>.

2. The architecture

2.1. Bi-LSTM architecture

Bi-LSTM is one of the most widely used deep learning components in neural systems for NLP nowadays. It is a bidirectional extension [Graves and Schmidhuber 2005] of the LSTM [Hochreiter and Schmidhuber 1997] that itself is an extension of the Elman Recurrent Neural Network (RNN) [Elman 1990].

2.1.1. Elman Recurrent Neural Networks

RNNs act as *feature extractors* of non-fixed-size sequences [Elman 1990, Goldberg 2017, Graves 2012], by mapping the entire *history* of previous time-steps to the current one. This mapping is done by keeping a *memory* vector running and being updated through the time-steps. However, one should note that the parameters used to update this memory are the same regardless of the time-step, giving generalisation power to the architecture. When dealing with natural language sentences, time steps are interpreted as positions in the input sentences, either words or characters depending on the desired granularity.

Let θ be the model's parameters, n the sequence size, and x_i , h_i and y_i the input, the memory and the output vectors, respectively, at position i . Also let f and g be non-linear mappings using θ . We can represent an RNN model as: $h_0 = 0$ (the null vector); $h_i = f(x_i, h_{i-1}; \theta)$, $1 \leq i \leq n$, and $y_i = g(h_i; \theta)$, $1 \leq i \leq n$.

2.1.2. Long short-term memory and its bidirectional extension

LSTM was designed to address the problem of vanishing/exploding gradients in RNNs. It is a more complex architecture, where units in the hidden layer are replaced by sub-nets, called memory blocks. These blocks enjoy of multiplicative gates, which, along with an efficient training algorithm, allow for a long term memory and prevent the vanishing/exploding gradient problem.

Bidirectional recurrent architectures [Graves and Schmidhuber 2005] were designed for when the current time-step output also depends on future time-steps. A second, identical network is instantiated which processes time-steps in reverse order: from t_n to t_1 . Their outputs are pairwise combined forming the output of the bidirectional network.

2.2. Residual connections

Sometimes, the optimal mapping that a layer (or consecutive layers) need to compute is the identity, or a similar mapping. Residual connections, firstly designed for Conv-

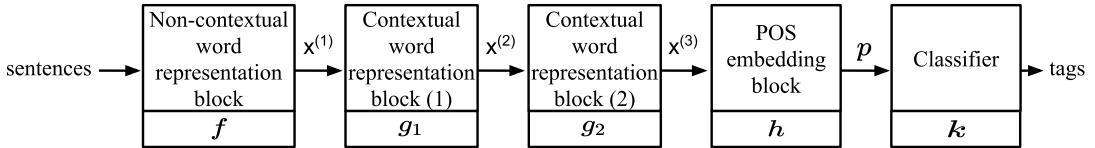


Figure 1. Architecture blocks

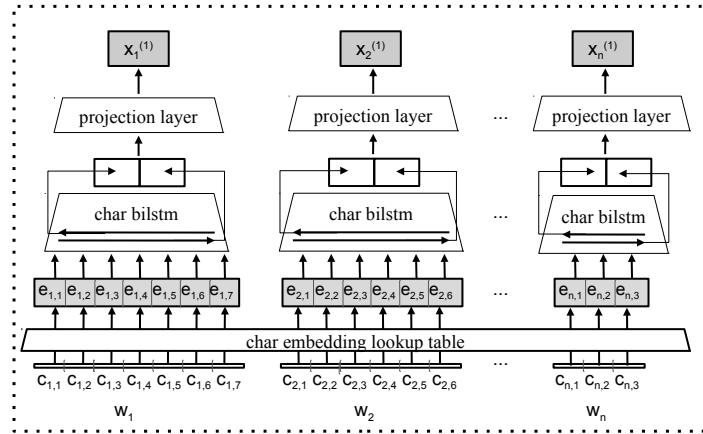


Figure 2. f : Character-based non-contextual word representation block

lutional Neural Networks [He et al. 2015], follow the hypothesis that a zero mapping is easier to approximate than an identity mapping and represents a "pass" on the model architecture from a lower to a higher layer. Thereby, it expects the affected layers to compute a null vector.

2.3. Our architecture

We organised our architecture in blocks (Figure 1). The initial blocks were designed as feature extractors for the words in context. Block f computes a non-contextual character-embedding-based word representation for every word in the sentence, while g_1 and g_2 compute contextual representations (word senses) for them. The next two blocks (h, k) are used to compute POS-refined representations for every word sense, classifying them according to the dataset from which the sentence was extracted. This is explained in detail below. Blocks f, g_1 and g_2 could be pretrained using a language modelling task.

Character embeddings have size d_c , word and word sense representations computed by f, g_1 and g_2 have size d_w , and the ones computed by h have size d_p . POS embeddings are represented as p . Sequences such as x_1, x_2, \dots, x_m will be referred to as $x_{1:m}$. We use n for the sentence size.

2.3.1. Non-contextual representations of words

Many approaches have been proposed to compute non-contextual word representations, including using subword information when dealing with morphologically rich languages, such as Portuguese. [Dos Santos and Zadrozny 2014] uses a convolutional network over character embeddings to obtain word representations. This idea is also used by [Peters et al. 2018] in a more complex manner. It has also been proposed to encode

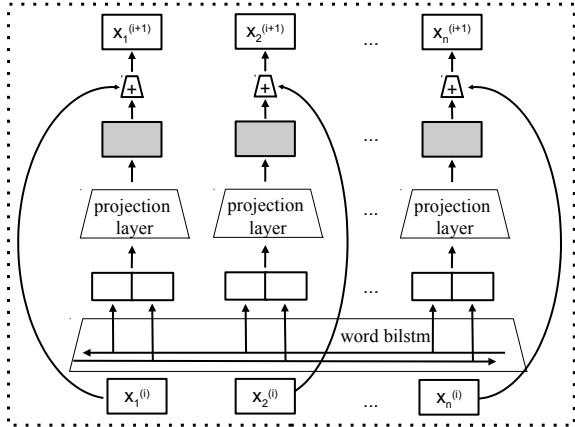


Figure 3. $g_i, i \in \{1, 2\}$: **Contextual word representation block**

character embeddings using Bi-LSTMs [Lample et al. 2016, Ling et al. 2015], and also adding all the character n-gram vectors in each word. [Mikolov et al. 2018]

As shown in Figure 2, each word is mapped into a character embedding sequence. Let $w_i, 1 \leq i \leq n$, be the i^{th} word in the sentence, and $c_{i,j}$ its j^{th} character. For every character $c_{i,j}$, an embedding $e_{i,j} \in \mathbb{R}^{d_c}$ is computed by a lookup table trained along with the whole model. For every word, their character embeddings feed a Bi-LSTM and the last output of the forward and backward passes of this Bi-LSTM are extracted and concatenated, composing an intermediate representation in \mathbb{R}^{2d_w} . These representations will then feed a linear layer (dimensionality reduction) to produce $x_{1:n}^{(1)}$, vectors in \mathbb{R}^{d_w} .

2.3.2. Contextual representations of words or word sense embeddings

Word sense embeddings take into account not only the words in isolation but also their context. In our model, these representations depend on the input from the entire sentence, unlike fixed-sized windows techniques [Mikolov et al. 2013a, Goldberg 2017]. This gives the models much more power to learn better representations.

Inspired by ELMo [Peters et al. 2018], we stack two instances (g_1 and g_2) of the same block architecture (Figure 3) to compute these representations. Block $g_i, i \in \{1, 2\}$, takes as input $x_{1:n}^{(i)}$, and passes it through a Bi-LSTM layer. For each word position, the forward and backward outputs of the Bi-LSTM are concatenated into an intermediate representation in \mathbb{R}^{2d_w} . These will then feed a linear layer in \mathbb{R}^{d_w} , computing intermediate representations which are added element-wise to $x_{1:n}^{(i)}$ to compute $x_{1:n}^{(i+1)}$.

2.4. POS-refined representations of words and classification

Finally the model computes in block h , in Figure 4, the POS embeddings $p_{1:n}$, as refinements of the word sense embeddings. It is almost the same architecture as g , but without the residual connections. The word sense representations $x_{1:n}^{(3)}$ feed the Bi-LSTM. For each word, the forward and backward outputs are concatenated resulting embeddings in \mathbb{R}^{2d_p} . A linear projection transforms them into $p_{1:n}$ in \mathbb{R}^{d_p} .

As $x_{1:n}^{(3)}$, $p_{1:n}$ are still continuous distributed representations of the word in con-

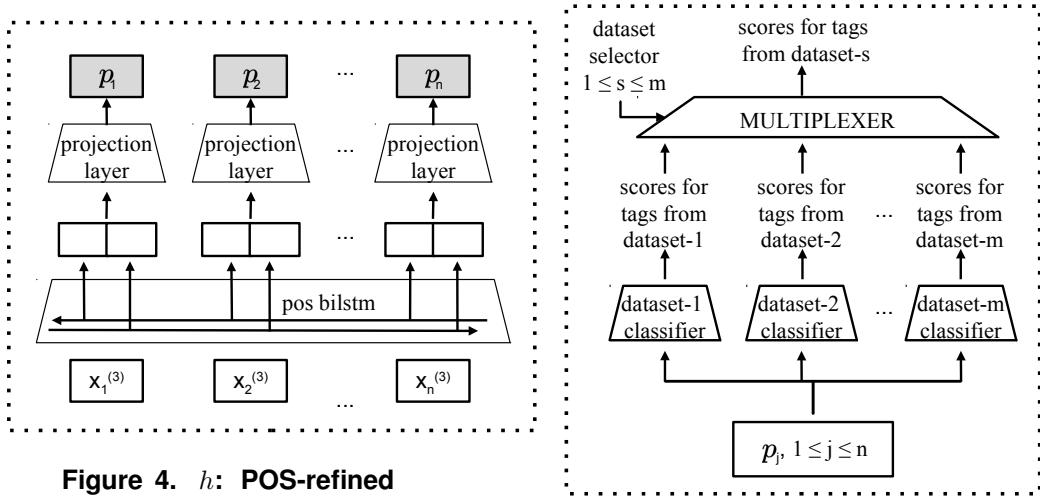


Figure 4. h : POS-refined word representation block

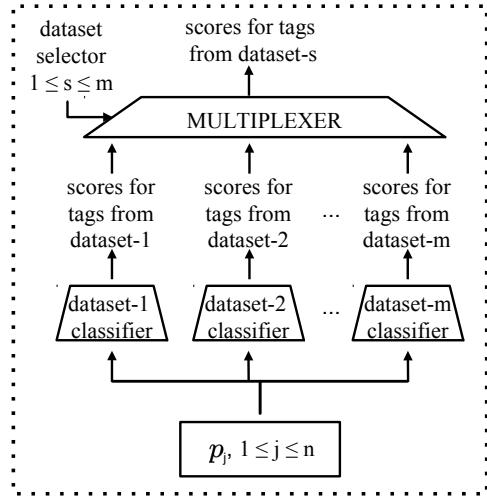


Figure 5. k : Classifier

text. However they are peculiar, as it is expected that their training stresses the syntactic information of the words in context, since a single projection layer should be able to map them into conventional POS tags. This is highly unexpected unless they already represent the POS but in a distributed continuous manner in the sense of [Harris 1954] and [Firth 1957]. Note that they cannot be thought of as "POS tag" embeddings, since they are not computed as if by a look up table with a finite tagset as input. And, on top of that, they strongly take context into account.

Figure 5 shows how this is accomplished. The $p_{1:n}$ embeddings are fed into projection layers followed by softmax normalisation, one for each dataset. Then, during training, we select the output corresponding to the tagset used in the corpora the sentence came from, and compare it to the target, computing the error used by the optimisation process. The error measure used is the cross entropy.

When used later in downstream applications, the POS-refined embeddings could be directly feed into the upper modules, instead of using the more limited POS tags.

3. Evaluation

The architecture used in this paper was implemented using *pytorch* [Paszke et al. 2017]. In this section, we depict the experiments setups, datasets details and results.

3.1. Datasets

The datasets we used are listed in Table 1.

Mac-Morpho: This is a large corpus of Brazilian Portuguese, from newspaper articles, annotated with POS tags with its own tagset, referred to in Table 1 as *MM* [Aluísio et al. 2003, Fonseca and Rosa 2013, Fonseca et al. 2015].¹ We used its third revision with 26 POS tags (22 primitive tags, plus punctuation, plus three tags for contractions such as *PREP+ART* for the word "*das*", which is a contraction of the preposition "*de*" ["of" in English] and the article "*as*" ["the"]). Training, validation and test sets were taken from the three files provided for this purpose in the site.

¹See also <http://nilc.icmc.usp.br/macmorpho/> and <http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf>.

Table 1. The datasets

Dataset name	Tagset	Train set		Validation set		Test set	
		#sent	#words	#sent	#words	#sent	#words
Mac-morpho	MM	37948	728497	1997	38881	9987	178373
GSD-UD	UD	9664	255755	1210	32129	1204	31496
Bosque-UD	UD	8328	206744	560	10851	477	10199
Bosque-LT	LT	3355	65086	419	7229	420	7995
Total	All	59295	1256082	4186	89090	12088	228063

GSD-UD: This is a corpus of Brazilian Portuguese available in <https://universaldependencies.org/#download>, converted from the Google Universal Dependencies Treebank, annotated with the 16 tags plus punctuation extended POS tagset of [Petrov et al. 2012] (UPOS). We used the training/dev/test files from version 2.4.

Bosque-UD: Bosque is part of *Floresta Sintática* a well known collection of corpora in the *Linguateca*². It contains annotated corpora of Brazilian and European Portuguese [Afonso et al. 2002]. Bosque-UD is a later version converted to UD POS tags [Rademaker et al. 2017]. We used version 2.3, available at <https://universaldependencies.org/#download>.

Bosque-LT: This is an older version of Bosque with the original tagsets in [Afonso et al. 2002]. We used the Brazilian Portuguese corpus of newspaper articles from *A Folha de São Paulo*. It has its own tagset with 23 tags plus punctuation. The sentences are not the same as those in Bosque-UD but there was a substantial overlap. For the sake of fairness, we built the training, development and test set by first moving sentences which were also in Bosque-UD to the same corresponding training/development/test sets. The remaining sentences in Bosque-LT were distributed to fit a 80-10-10 % split.³

Due to the different tokenization assumptions in each corpora, our model learns representations from all word forms involved. Recalling the previous example, it will successfully encode both the contracted word “das” and their split pair “de” and “as”. That allows for the versatile use of the POS tagging module embedded in downstream applications with any tagset assumption.

It is interesting to point out that whereas the correspondence from one tagset to another is not trivial when translating, say, from LT to UD, and even the choice of the correct POS tag by annotators is a challenging problem reported on any corpus annotator manual [Manning 2011], the continuous distributed intermediate representation for the POS is allowed to freely represent the syntactic distribution of the word in context.

3.2. Experiments and results

Each experiment consisted of a training phase of 55 epochs and a test phase. We validated the parameters at the end of each epoch by calculating the average error over some

² See <https://www.linguateca.pt/Floresta/corpus.html#bosque>

³We noticed a small overlap remained between Mac-morpho and Bosque that crosses from training to test or development files, involving no more than 10 noisy sentences. Their impact in evaluation is negligible.

Table 2. Accuracies

Dataset	Single corpus		Combined Training	
	μ	σ	μ	σ
Mac-Morpho	97.28	0.039	97.46	0.004
GSD-UD	97.27	0.024	97.87	0.034
Bosque-UD	96.24	0.097	97.18	0.039
Bosque-LT	96.40	0.102	98.32	0.226
Micro-average	97.20	-	97.53	-

development set. At the end of training we chose the parameter set giving the smallest error. Each experiment was performed three times and we report the average accuracy and standard deviation in Table 2. Column labelled "Single Corpus" reports the first set of four experiments, one for each of the original corpora, used for control. Each cell in this column reports the accuracy obtained when the model was trained, validated and tested with sentences of a specific corpus.

In the other experiment reported under "Combined Training", we trained the model using the sentences of all of the four corpora, informing, at the multiplexer (Figure 5), the corpus (and hence the tagset) the sentence came from. Validation was done over the union of the development sets as well. Each cell in that column reports the accuracy of the combined model when tested over the test set for a specific corpus. The bottom row has the micro-average accuracy.

For the sake of reproducibility, we inform that: batches of 32 sentences were defined at the beginning of each experiment, each batch contained sentences from the same corpus (even in the combined training). We shuffled the order of the batches for each epoch. We used Adadelta optimizer [Zeiler 2012], with $\rho = 0.9$, and $\epsilon = 1e-6$. For all experiments, $d_c = 70$, $d_w = 350$ and $d_p = 150$. Dropout [Srivastava et al. 2014] was used, and the layers in which it was applied are the ones filled gray in Figures 2, 3, 4 and 5. Drop ration was set to 0.1 in f , 0.2 in g_1 and g_2 , and 0.4 in h . We used begin/end delimiters for words (BOW, EOW) and sentences (BOS, EOS). We did not pretrain the word representation level.

Crucially, the results obtained when training on all corpora, mixing different tagsets and different tokenization assumptions, not only give higher micro-average accuracy, but surpasses the accuracy on all test sets.

More importantly, we obtain an instance of representation of the parts of speech, the POS embeddings p_i , which captures the syntactic notion of lexical category of the words in context, which are not bound to specific finite tagsets and are more likely to represent a true word class distribution.

4. Related Word

[Bick 2000] is a handwritten rule-based parser that for long time was probably the best choice for reasonably accurate POS tagging for Portuguese, with the tagset from the VISL Project used in the Linguateca (Bosque-LT in Table 1). With the advent of the neural approaches revisiting POS tagging, [Dos Santos and Zadrozny 2014] achieved accuracy of 97.47% on the Mac-Morpho corpus revision 1, with a character-based convolutional

neural network (CNN) approach. [Fonseca et al. 2015] obtained higher accuracy scores of 97.57% in revision 1 (the one used in [Dos Santos and Zadrozny 2014]), 97.48% in revision 2, and 97.33% in revision 3 which was the one used in our experiments. The decreasing scores are explained by the fact that contractions were split up in earlier version and hence somewhat easier to tag. Our architecture scored 97.46% in revision 3 slightly topping the state-of-the-art for this dataset.

Multilingual architectures that used Bosque-UD v1.2 (an earlier version than the one we have used), achieved accuracies of 97.94% [Plank et al. 2016] and 98.20% [Heinzerling and Strube 2019].

5. Conclusions and future work

In this paper we present a multi-objective neural architecture for POS tagging for Portuguese that is trained over multiple heterogeneous corpora, with different tagsets and different tokenization assumptions. We propose and generate a new continuous distributed representation for parts of speech, the POS embeddings which accurately adapt to different tagsets with a single neural output layer, still being allowed to preserve the true distribution of the word classes in context in the sense of [Harris 1954] and [Firth 1957]. They can be used in downstream applications, instead of the discrete, finite tagsets traditionally found in the literature and used to annotate corpora. Yet we achieve state-of-the-art accuracy even without pretraining our model.

As future work, we intend to investigate relatedness and similarity characteristics in the POS embeddings $p_{1:n}$ compared to other, contextual and non-contextual, word and word sense representations instances in our architecture $(x_{1:n}^{(1)}, x_{1:n}^{(2)}, x_{1:n}^{(3)})$ to measure to what extent the claim is valid, that the POS embeddings indeed factor out semantic content in favour of syntactic context, and generate a more appropriate POS representation than the usual finite ones.

We would also like to do extrinsic evaluation, to confirm expected accuracy gains of replacing traditional POS tags with the POS embeddings in downstream applications.

Still, we have to see whether pretraining improves accuracy even more.

References

- Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá(c)tica: A treebank for portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*.
- Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., and Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In Mamede, N. J., Trancoso, I., Baptista, J., and das Graças Volpe Nunes, M., editors, *Computational Processing of the Portuguese Language*, pages 110–117, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Bengio, Y., Courville, A. C., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

- Bick, E. (2000). *The Parsing System “Palavras”. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Århus.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.
- Dos Santos, C. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *31st International Conference on Machine Learning, ICML 2014*, volume 5.
- Elman, J. L. (1990). Finding structure in time. *COGNITIVE SCIENCE*, 14(2):179–211.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.
- Fonseca, E. R. and Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, STIL 2013, Fortaleza, Brazil, October 21-23, 2013*.
- Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *J. Braz. Comp. Soc.*, 21(1):2:1–2:14.
- Francis, W. N. and Kucera, H. (1979). Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.
- Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, 1th edition.
- Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385.
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *NEURAL NETWORKS*, pages 5–6.
- Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.
- Heinzerling, B. and Strube, M. (2019). Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. *arXiv preprint arXiv:1906.01569*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.
- Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.
- Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fernandez, R., Amir, S., Marujo, L., and Luis, T. (2015). Finding function in form: Compositional character models for open

- vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *LREC*. European Language Resources Association (ELRA).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, pages 3111–3119, USA. Curran Associates Inc.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.
- Petrov, S., Das, D., and McDonald, R. T. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2089–2096.
- Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *ACL (2)*. The Association for Computer Linguistics.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing)*, pages 197–206, Pisa, Italy.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.
- Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Um Modelo para Sistema de Diálogo Fim-a-Fim usando Conhecimento de Senso Comum.

Cecília S. Carvalho¹, Vládia C. Pinheiro¹, Lívio Freire²

¹Programa de Pós Graduação em Informática Aplicada – Universidade de Fortaleza (Unifor)
Fortaleza – CE – Brasil

²Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil

ceciliacarvalhoo@gmail.com, vladiacelia@gmail.com, livio.amf@gmail.com

Abstract. *In this paper we present a end-to-end dialogue system that use common sense knowledge to train a neural network model to learn information not explicit in a conversation. The results obtained in the tests show that using common sense knowledge to train dialogue systems with a Portuguese corpus improve the classification of pairs of dialogues.*

Resumo. *Neste trabalho, apresentamos um sistema de diálogo fim-a-fim, para língua portuguesa, que utiliza conhecimento de senso comum como forma de melhorar o processo de aprendizagem com informações implícitas em uma conversa. Os resultados dos experimentos realizados mostram melhoria significativa na cobertura das respostas associadas pelo Chatbot, mantendo-se o mesmo nível de corretude.*

1. Introdução

Trabalhos recentes mostram que modelos de sistemas de diálogos ou *Chatbots* podem ser treinados de maneira orientada a dados e de forma fim-a-fim, sem o uso de codificação de regras manuais [Ghazvininejad et al. 2018]. [Young et al. 2018] argumentam que, para maximizar o desempenho de sistemas de diálogos, deve-se utilizar informações inicialmente subentendidas no texto, na forma de conhecimento de mundo. Segundo os autores, pessoas respondem umas às outras de maneira significativa, não somente analisando a última expressão falada, mas relembrando informações importantes mencionadas na conversa e integrando-as a novas falas.

Atualmente, os métodos para sistemas de diálogos são categorizados em *retrieval-based* e *generation-based*. O método *retrieval-based* utiliza um conjunto de respostas e, dentre elas, escolhe a mais adequada. Já o método *generation-based* gera uma resposta do sistema para um diálogo com um usuário, ou seja, a resposta não é escolhida a partir de um conjunto de sentenças pré-definidas, pois o método foca em construir uma nova resposta à pergunta do usuário. O trabalho de [Young et al. 2018] consiste em um sistema de diálogo fim-a-fim do tipo *retrieval-based*, para língua inglesa, que utiliza conhecimento de senso comum. [Maeda and Moraes 2017], por sua vez, propõem um *Chatbot* para língua portuguesa com foco em conversas de domínio aberto, o qual utiliza a arquitetura de redes neurais *Sequence to Sequence*. Para língua portuguesa não foram encontrados trabalhos que considerem o aprendizado de conhecimento de senso comum na geração de *Chatbots*. Um dos desafios da pesquisa é justamente a falta de disponibilidade de dados conversacionais para treinamentos dos modelos fim-a-fim.

Neste trabalho, propomos um modelo de Rede Neural baseado em *long short-term memory* (LSTM) [Hochreiter and Schmidhuber 1997] para sistemas de diálogo fim-a-fim em português, que utiliza conhecimento de senso comum. Foram realizados experimentos com uma base de 42000 diálogos reais entre usuários e um *Chatbot* em uso em uma rede de clínicas médicas populares. O *córpus* foi analisado e revisado de forma a conter somente diálogos significativos e coerentes. Os resultados indicam a viabilidade do modelo devido ao aumento do desempenho na cobertura das respostas, da ordem de 13%, quando enriquecido com conhecimento de senso comum.

2. Trabalhos Relacionados

No contexto de sistemas de diálogos e processamento de linguagem natural é possível selecionar alguns trabalhos que colaboraram com o avanço da área. Por exemplo, [Bordes et al. 2016] propuseram um sistema de dialogo baseado em *Memory Network* no contexto de reserva de restaurante. [Sukhbaatar et al. 2015] apresentam um modelo de rede neural com atenção recorrente. A arquitetura utilizada é semelhante a uma *Memory network*, entretanto a arquitetura é treinada de forma fim-a-fim, tornando possível sua aplicação em diversas tarefas como *question answering*.

Atualmente, é possível relatar algumas pesquisas que envolvem conhecimento de senso comum aplicado em sistemas de diálogos, como por exemplo o trabalho de [Young et al. 2018], que propuseram um sistema de conversação de temas gerais, explorando a utilização de conhecimento de senso comum como memória externa. A arquitetura de redes neurais LSTM foi utilizada para processar o conhecimento de senso comum. Outro trabalho que utiliza uma fonte externa de informação é o proposto em [Lowe et al. 2015]. Os autores trabalharam em um sistema que seleciona respostas de um diálogo dentre uma lista de respostas candidatas. Na arquitetura deste sistema, eles incorporaram fontes de conhecimentos textuais desestruturados. A arquitetura é responsável por identificar as partes mais relevantes do texto externo e três redes neurais recorrentes são responsáveis por processar as entradas. Ambos os trabalhos anteriores aplicaram suas técnicas em corpus de língua inglesa.

Em [Young et al. 2018], o *Dataset* utilizado é composto por 1.4M de pares de diálogos do Twitter, e para obtenção do conhecimento de senso comum a base utilizada foi a *ConceptNet* [Speer et al. 2017]. A estratégia utilizada neste trabalho para a utilização de senso comum foi uma arquitetura chamada *Tri-LSTM Encoder* que utiliza uma terceira LSTM para codificar as declarações de senso comum. De acordo com os autores, o modelo é responsável por classificar N respostas, onde uma destas respostas era positiva e as restantes eram negativas. O valor de N=10 foi usado na fase de testes. Caso a classificação da resposta positiva fosse menor que k, a métrica *Recall@K* é positiva para aquela instância. Os resultados dos experimentos para as métricas *Recall@1*, *Recall@2* e *Recall@3* foram de 77.5%, 88.0% e 96.6%.

Para língua Portuguesa não foram encontrados trabalhos que exploram conhecimento de senso comum no desenvolvimento de *Chatbots*. No momento, boa parte dos trabalhos que envolvem *Chatbots* em língua Portuguesa não utilizam arquiteturas de redes neurais. Uma exceção é o trabalho de [Maeda and Moraes 2017], o qual apresenta um sistema de *Chatbot* de domínio aberto utilizando um modelo de rede neural conhecido como *Sequence to Sequence*. Este modelo utiliza LSTM e tem uma arquitetura composta por um

Encoder que recebe a entrada e um *Decoder* que gera a sequencia de saída. O *córpus* utilizado neste trabalho foi de legendas de filmes totalizando 12.775 sentenças, bem como um *córpus* de mensagens de um bate papo em um aplicativo de celular, totalizando 30.684 sentenças. A etapa de avaliação para o *córpus* de legendas de filmes foi realizada com 90 usuários através de uma interface web. Em seguida foram feitos os testes com o *córpus* de bate papo, onde concluiu-se que, utilizando um corpus com identificação das falas dos participantes, o sistema gera respostas mais coerentes.

Outro projeto em língua Portuguesa é o de [Moreno et al. 2015], o qual apresenta um *Chatbot* para divulgação do Atlas Linguístico do Brasil através do Whatsapp. O *chatbot* funciona operando dois arquivos XML sendo um para perguntas e respostas e outro para sinônimos. Este *Chatbot* é orientado para uma tarefa específica, não sendo trivial sua adaptação para outros domínios.

2.1. Conhecimento de Senso Comum

Atualmente, existem várias bases de senso comum disponíveis na internet, tais como, *ConceptNet* [Speer et al. 2017], *InferenceNet* [Pinheiro et al. 2010] e *SenticNet* [Cambria et al. 2018]. De acordo com [Young et al. 2018], o objetivo de se representar conhecimento de senso comum é prover uma base de conhecimento de mundo a uma variedade de aplicações em Inteligencia Artificial.

Uma forma de visualizar o conteúdo destas bases de conhecimento de senso comum é através de um grafo, onde os nós são conceitos e as arestas são a relação entre estes conceitos. Por exemplo, "Um(a) [[computador]] é usado(a) para [[trabalhar]].", sendo 'computador' e 'trabalhar' conceitos e a relação entre eles é 'é usado(a) para'.

3. Um Modelo para *Chatbot* usando Conhecimento de Senso Comum

Neste trabalho, propõe-se um sistema de diálogo que faz uso de conhecimento de senso comum, extraídos de redes semânticas, tais como, *ConceptNet*, *InferenceNet* e *SenticNet*. O *Chatbot* foi criado utilizando um modelo de redes neurais com arquitetura LSTM que recebe como entrada a frase do usuário, a resposta do sistema à frase do usuário e as sentenças que expressam conhecimento de senso comum.

Para cada tripla de entrada, da forma (entrada do usuário, resposta do sistema e sentenças de senso comum) é atribuído um valor binário que representam a coerência do diálogo. Ou seja, se a resposta do sistema corresponde a entrada do usuário o valor 1 é atribuído à tripla de entrada, indicando que estas entradas são coerentes. Já o valor 0 é atribuído quando a resposta do sistema não é coerente com a entrada do usuário.

O conhecimento de senso comum é utilizado para ensinar, ao modelo de rede neural, informações que não estão explícitas nas conversas. O conhecimento de senso comum é relacionado a uma entrada de usuário, se em ambos houver uma palavra em comum. Por exemplo, se o usuário pergunta "Onde tem uma clínica mais próxima?", o conhecimento de senso comum 'Você geralmente encontra um(a) [[doutor]] em um(a)[[clínica]].' será relacionado à esta entrada de usuário devido a palavra **clínica** se encontrar em ambas as sentenças. Entretanto, pode ocorrer que nem todos os conhecimentos associados a uma entrada de usuário sejam úteis na classificação da melhor resposta do sistema, sendo necessário um modelo de aprendizagem profunda que também aprenda qual conhecimento tem maiores chances de auxiliar na classificação do diálogo.

A Figura 1 apresenta a arquitetura do modelo de *Chatbot* que faz uso de conhecimento de senso comum. Nesta arquitetura estão presentes três entradas que são passadas para o sistema: *eu* - representando a entrada do usuário; *es* - representando a entrada do sistema; e *esc* - representando as sentenças com conhecimento de senso comum. Inicialmente, as entradas do *Chatbot* passam por uma etapa de pré-processamento, onde são removidas *stopwords* e, em seguida, cada palavra é representada por vetores pré treinados de *Word Embeddings* - *GloVe* [Pennington et al. 2014], *Word2Vec* [Mikolov et al. 2013] e *FastText* [Bojanowski et al. 2017]. No exemplo da figura 1 na frase “Oi layse Boa noite”, as palavras “*Oi*”, “*layse*”, “*Boa*” e “*noite*” serão substituídas por vetores numéricos. Da mesma forma, as sentenças que representam conhecimento de senso comum serão substituídas por *Word Embeddings*: na sentença “Um(a) noite é usado(a) para dormir”, as palavras “*noite*” e “*dormir*” serão substituídas por seus *Word Embeddings* correspondentes.

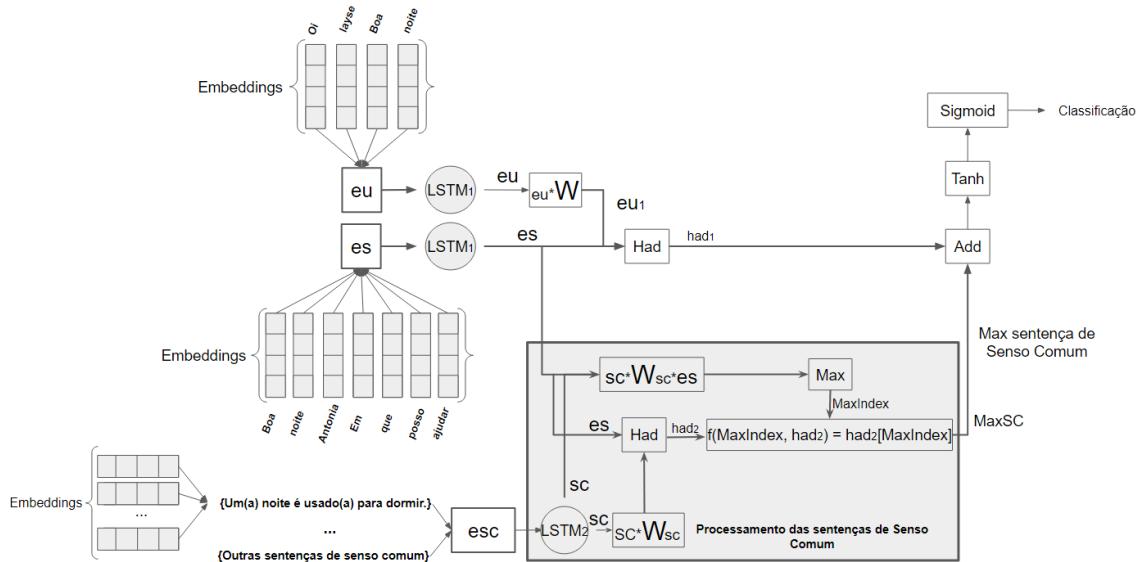


Figura 1. Arquitetura do modelo de aprendizagem de *Chatbot* fim-a-fim, que utiliza conhecimento de senso comum

Após passar pela LSTM, *eu* é multiplicada por uma matriz *W* que tem seus valores aprendidos pela rede neural. Com o resultado desta multiplicação, é aplicado um produto de Hadamard [Horn 1990] com *es* que faz uma multiplicação elemento a elemento, como mostra a equação 1. O resultado é nomeado de *had*₁ e será usado mais a frente.

$$f(eu_1, es) = eu_1 \circ es \quad (1)$$

Já as sentenças de senso comum, agrupadas em uma lista e ligadas à entrada do usuário, são primeiramente processadas por uma LSTM gerando *sc* e em seguida dois cálculos são realizados. Inicialmente, é calculada uma métrica de similaridade entre o conhecimento de senso comum *sc* e a resposta do sistema, como mostra equação 2. Nesta equação *W_{sc}* é uma matriz cujos valores são aprendidos pela rede neural.

$$f(sc, es) = sc * W_{sc} * es \quad (2)$$

Com o resultado da equação 2, é possível retornar o índice do senso comum de maior correspondência com a entrada do sistema, chamado de *MaxIndex*. Este índice será

usado para buscar a sentença de conhecimento de senso comum com o resultado da equação 3, chamado de had_2 .

$$f(sc, es) = (sc * W_{sc}) \circ es \quad (3)$$

Em outras palavras, tendo o resultado had_2 da equação 3, é utilizado o *MaxIndex* para recuperar o conhecimento de senso comum que mais ajudará na classificação. Este conhecimento é chamado de *MaxSC* e é adicionado ao resultado had_1 da equação 1. Em seguida, é aplicado uma função de ativação tangente hiperbólica. Estes cálculos são apresentados na equação 4. Por fim, no resultado da equação 4 é aplicado uma função *Sigmoid* que retornará a classificação.

$$f(had_1, MaxSC) = \tanh(had_1 + MaxSC) \quad (4)$$

Argumentamos que, como o conhecimento de senso comum foi selecionado tendo como referência a entrada do usuário para o sistema, ao calcular a similaridade entre este conhecimento de senso comum e a saída do sistema, é possível identificar a resposta adequada com maior acurácia. Em outras palavras, nosso argumento é que a adição de conhecimento externo ao modelo de aprendizagem, conhecimento este diretamente relacionado à entrada do usuário, melhora o entendimento da resposta correta.

4. Experimentos e Resultados

4.1. Dataset

O conjunto de dados utilizado foi a base de diálogos do *Chatbot* da Clínica SIM, uma rede de Clínicas que atua no Nordeste Brasileiro, com foco em oferecer atendimento médico e odontológico a pessoas que não possuem plano de saúde. Inicialmente, o conjunto de dados foi organizado em pares de diálogos da forma (entrada do usuário, resposta dada pelo sistema ao usuário). A Figura 2 apresenta dois exemplos de pares de diálogos extraídos do conjunto de dados disponibilizado. O par de diálogo A apresenta uma conversa considerada coerente, e o par de diálogo B corresponde a uma conversa incoerente, pois a resposta do sistema não atende à necessidade do usuário.

Primeiramente, foram selecionados os pares de diálogos onde a entrada do usuário e a resposta do sistema continham mais de três palavras e pelo menos uma palavra não *stopword*.

A	Entrada do usuário: Boa tarde Tem cardiologista Agora a tarde	Resposta do sistema: Boa tarde Vou te enviar o link para você verificar os dias e horários disponíveis Só um momento
B	Entrada do usuário: Quais serviços você oferece	Resposta do sistema: Obrigado pelo contato sempre que precisar pode contar com a gente

Figura 2. Exemplos de pares de Diálogos

No final, 19361 pares de diálogos compõem a base de diálogos coerentes. Uma dificuldade inicial foi construir o conjunto de exemplos negativos, ou seja, pares de diálogos incoerentes. Para construir os exemplos negativos, a partir do atual conjunto de dados,

uma entrada de usuário foi escolhida aleatoriamente e agrupada a uma resposta do sistema, também escolhida aleatoriamente. Para diminuir a probabilidade de ocorrência de diálogos coerentes classificados erradamente como “incoerentes”, devido a aleatoriedade, foram gerados somente 12000 diálogos incoerentes. No total, o *córpus* disponível para este trabalho é composto por 31361 pares de diálogos.

Por questões de performance, foram selecionados aleatoriamente 15.680 pares de diálogos que equivale aproximadamente a 50% do corpus inicial. O conjunto de dados de treinamento é composto por 12.544 pares de diálogos, igualmente divididos em pares coerentes e incoerentes. O conjunto de teste contém 3136 pares de diálogos e também está igualmente dividido em pares coerentes e incoerentes. Tabela 1 apresenta os quantitativos do *córpus* de treinamento e teste.

Tabela 1. Estatística do Conjunto de Pares de diálogos, para a fase de treinamento e teste do modelo

	Treinamento	Teste	Total
Pares de Diálogos Coerentes	6272	1568	7840
Pares de Diálogos Incoerentes	6272	1568	7840

4.2. ConceptNet

Neste trabalho foi utilizado como base de conhecimento de senso comum a rede semântica *ConceptNet*[Speer et al. 2017]. As triplas que descrevem o conhecimento de senso comum são recuperadas utilizando o vocabulário criado a partir dos diálogos que compõem o *córpus*, ou seja, para cada palavra presente no vocabulário, se esta palavra estiver em uma tripla de conhecimento de senso comum, este conhecimento é recuperado para uso. Por exemplo, para a palavra de busca “clínica” as seguintes triplas de conhecimento de senso comum são recuperadas: ‘Você geralmente encontra um(a) [[doutor]] em um(a) [[clínica]].’ e ‘Uma coisa que você pode encontrar em um(a) [[clínica]] é um(a) [[médico]].’

Com um vocabulário inicial de 11582 palavras, foram recuperadas da *ConceptNet* 14363 relações de conhecimento de senso comum (triplas). Entretanto, algumas das relações estavam em inglês e eram a tradução de alguma palavra, como por exemplo, “[especialista] is a translation of [[specialist]]”. Estas relações foram descartadas, restando 7642 relações que expressam conhecimento de senso comum. Deste conjunto foram removidas as *stopwords*. O vocabulário inicial foi então acrescido com os termos das relações de conhecimento de senso comum e, ao final, totalizou em 14164 palavras.

Em seguida, para cada entrada do usuário (*eu*), é selecionado um conjunto de triplas de conhecimento de senso comum, caso existam palavras em comum entre a *eu* e a *esc*. Neste trabalho foi utilizado um número máximo de cinco triplas de senso comum para cada entrada de usuário, evitando assim duplicidade. Caso o número de triplas de senso comum seja menor que cinco, algumas triplas existentes são repetidas aleatoriamente até que a lista tenha tamanho cinco. Para o conjunto de treinamento, 174 entradas de usuário não continham conceitos na *ConceptNet*, e para o conjunto de teste, 49 entradas de usuário

não foram cobertas pela *ConceptNet*. Nestes casos, foi adicionado um texto padrão 'não tem conceito'.

4.3. Análise dos Resultados

Para avaliação de desempenho, dois cenários foram definidos e cada um representa um modelo de aprendizagem diferente.

No primeiro cenário, foi criado um modelo de *Chatbot* que não utiliza conhecimento de senso comum para a classificação dos diálogos. Este modelo representa as sentenças com *Word Embeddings*, utiliza arquitetura LSTM e em seguida realiza uma multiplicação entre as entradas. No final, é aplicado uma função *Sigmoid* para gerar o resultado. As sentenças são inicializadas com *Word Embeddings* pré-treinadas do *GloVe* [Pennington et al. 2014] com vetores de dimensão 100. O modelo utilizou um *Batch Size* de tamanho 32 e o *Optimizer* escolhido foi o ADAM. Na LSTM foram utilizados 150 *Units*, com *activation* Tanh, *recurrent dropout* de 0.2 e *dropout* 0.2. O treinamento foi realizado com 20 *epochs*.

No segundo cenário, o sistema de *Chatbot* utiliza conhecimento de senso comum, onde, para cada entrada de usuário, são selecionadas cinco sentenças de senso comum, passadas como uma terceira entrada nas etapas de treinamento e teste do modelo. Os parâmetros utilizados neste cenário foram: *Word Embeddings* do *GloVe* com vetores de 100 dimensões, o *Batch Size* foi de tamanho 32 e a quantidade de *epochs* foi 20. O *Layer* de LSTM utilizado para os pares de diálogos tinham 150 *units*, com *activation* Tanh, *recurrent dropout* de 0.2 e *dropout* 0.2. A rede LSTM utilizada para processar o senso comum utilizou apenas o parâmetro de 150 *units*.

O modelo deve retornar a classificação destes pares de diálogos quanto a coerência em uma escala de 0 a 1, onde 0 implica diálogo incoerente e 1 indica diálogo completamente coerente. Utilizamos como *threshold* (ponto de corte) o valor de 0,5 indicando que o par de sentenças do diálogo (usuário-sistema) é coerente quando o modelo retorna valor superior a 0,50.

Os resultados dos testes em ambos os cenários são apresentados na tabela 2, em termos das medidas de precisão, cobertura e *F1-measure*. No Cenário 2, o qual considera a adição de conhecimento de senso comum, houve melhoria de 13 pontos percentuais na medida *Recall* e menos de 1% de aumento na medida Precisão (0,2 pontos percentuais). Os resultados indicam que, adicionando conhecimento de senso comum ao *ChatBot*, o mesmo responderá ao usuário em torno de 25% mais vezes e com uma taxa equivalente de acerto (57%). A Tabela 3 apresenta a matriz de confusão para ambas as classes "Coerente" e "Incoerente", nos dois cenários.

Tabela 2. Resultados em termos de Precisão, Recall e F1-Measure, referentes aos Cenários 1 e 2

	Coerente		Incoerente	
	Cenário 1	Cenário 2	Cenário 1	Cenário 2
Recall	0.5389	0.6734	0.5905	0.4936
Precisão	0.5682	0.5708	0.5615	0.6001
F1 measure	0.5644	0.5802	0.5755	0.5416

Tabela 3. Matriz de confusão referente aos Cenários 1 e 2

Cenário 1		Real		Cenário 2		Real	
		Coerente	Incoerente			Coerente	Incoerente
Preditivo	Coerente	845	642	Preditivo	Coerente	1056	794
	Incoerente	723	926		Incoerente	512	774

5. Conclusão

Neste trabalho, foi proposto o uso de conhecimento de senso comum em um modelo de aprendizagem profunda para *Chatbots* em língua portuguesa, do tipo *retrieval-based*, que utiliza arquitetura de redes neurais LSTM. Os testes foram realizados com o *córpus* de diálogos de usuários das Clínicas SIM, um grupo de clínicas populares de atendimento médico, instaladas na região nordeste de Brasil.

Estudos anteriores [Young et al. 2018] mostraram melhoria de desempenho quando do uso de conhecimento de senso comum em sistemas de diálogo em língua inglesa. No entanto, não se poderia afirmar o mesmo para outras línguas. Os resultados apresentados neste trabalho mostraram ser promissora a utilização de conhecimento de senso comum na classificação de pares de diálogos, pois houve melhoria significativa na cobertura das respostas, mantendo o mesmo nível de precisão. Entretanto, deve-se ressaltar a importância de realizar testes de significância estatística nos dados dos experimentos, a fim de fortalecer a confiabilidade nos resultados.

Como trabalhos futuros, tem-se a expansão do conjunto de dados de senso comum buscando conhecimento em outras bases, como por exemplo, a *InferenceNet*, especializada em conhecimento de senso comum para a língua portuguesa. Também podem ser avaliadas melhorias no próprio modelo de rede neural, como por exemplo o desenvolvimento de um *Chatbot* do tipo *generation-based* considerado mais natural. Por fim, é importante aplicar métodos mais confiáveis na criação de pares de diálogos incompatíveis, diminuindo o risco de envio de má informação para o modelo de aprendizagem profunda.

6. Agradecimentos

Especial agradecimento à Funcap pelo apoio através do Programa de Inovação Tecnológica (Inovafit) e à empresa Tallos pela disponibilização dos dados conversacionais utilizados neste trabalho.

Referências

- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Bordes, A., Boureau, Y.-L., and Weston, J. (2016). Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.
- Cambria, E., Poria, S., Hazarika, D., and Kwok, K. (2018). Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., and Galley, M. (2018). A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Horn, R. A. (1990). The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, pages 87–169.
- Lowe, R., Pow, N., Serban, I., Charlin, L., and Pineau, J. (2015). Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*.
- Maeda, A. and Moraes, S. (2017). Chatbot baseado em deep learning: um estudo para língua portuguesa.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Moreno, F., Manfio, E., Barbosa, C. R., and Brancher, J. D. (2015). Tical: Chatbot sobre o atlas linguístico do brasil no whatsapp. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 26, page 279.
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Pinheiro, V., Pequeno, T., Furtado, V., and Franco, W. (2010). Inferencenet. br: expression of inferentialist semantic content of the portuguese language. In *International Conference on Computational Processing of the Portuguese Language*, pages 90–99. Springer.
- Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., and Huang, M. (2018). Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Aplicação de Reconhecimento de Entidades Nomeadas em investigação de Crimes Financeiros

Fábio Moreira¹, Renata Vieira¹

¹Pontifícia Universidade Católica do Rio Grande do Sul (Brasil)
fabio.moreira@acad.pucrs.br, renata.vieira@pucrs.br

Abstract. *With the increasing technological advance, criminal organizations are increasingly using technology to commit crimes such as Money Laundering. Individuals and corporations have been replacing the use of physical documents with spreadsheets and digital applications that can store and process large amounts of information. Public security agencies, especially the Federal Police, collect a large volume of digital data from open sources and seized materials (computers, notebooks, smartphones, email servers, etc.), which are stored and analyzed in order to obtain the materiality and authorship of the crime investigated, the primary purposes of the judicial police. Investigative teams, made up of financial police specialists, need technology tools capable of automatically analyzing and extracting information, enabling the work of the experts. The scope of this paper is to apply Natural Language Processing (NLP) in the specific domain of Money Laundering Police Investigation, analyzing the performance of two Named Entity Recognition (NER) tools, LSTM+CNN and Spacy, as well as identify possibilities for future work.*

Resumo. *Com o crescente avanço tecnológico, as organizações criminosas estão cada vez mais se utilizando da tecnologia para o cometimento de crimes como o de Lavagem de Dinheiro. As pessoas físicas e jurídicas foram substituindo o uso de documentos físicos por planilhas e aplicativos digitais, capazes de armazenar e processar um grande número de informações. Os órgãos de segurança pública, em especial a Polícia Federal, arrecada um grande volume de dados digitais, provenientes de fontes abertas e materiais apreendidos (computadores, notebook, smartphone, servidores de email, etc.), que são armazenados e analisados a fim de se obter a materialidade e autoria do delito investigado, fins precípuos da polícia judiciária. As equipes de investigações, compostas por policiais especialistas nas áreas financeiras, necessitam de ferramentas tecnológicas capazes de analisar e extrair informações automaticamente, viabilizando o trabalho dos experts. O escopo deste trabalho consiste na aplicação de técnicas de Processamento de Linguagem Natural (PLN) no domínio específico da investigação policial de crimes de Lavagem de Dinheiro, analisando o desempenho de duas ferramentas de Reconhecimento de Entidades Nomeadas (REN), LSTM+CNN e Spacy, bem como identificar as possibilidades para trabalhos futuros.*

1. Introdução

As investigações policiais de combate ao crime de Lavagem de Dinheiro, previsto na Lei nº 9.613/98 e alterado pela Lei 12.683/12, ocupam um importante papel social com destaque na mídia, a exemplo da Operação Lava Jato, na medida em que enfrentam a corrupção

e expõem o descaso com os recursos que deveriam ser revertidos para a sociedade, mas que estão sendo utilizados para o enriquecimento de poucos.

Nas palavras de [de LEMOS JÚNIOR 2007], no que tange às dificuldades encontradas na investigação do crime de Lavagem de Dinheiro, se destaca a dificuldade de visualizá-lo, pois não há uma vítima pontual e, tampouco, um único agente do delito. O concurso de pessoas é indispensável à consecução penal do delito, o que funciona como fator altamente complicador. Não se trata de uma simples co-autoria, tampouco de uma quadrilha ou bando, mas sim de uma complexa estrutura de pessoas organizadas em torno de um objetivo comum. Para tanto, tais agentes não precisam estar próximos ou agirem juntos.

Considerando a complexidade do delito em estudo e da identificação dos possíveis autores envolvidos na prática delituosa, combinado ao grande volume de dados provenientes dos materiais produzidos e apreendidos em uma investigação policial deste porte, verifica-se a importância da aplicação de técnicas de Processamento de Linguagem Natural (PLN) no auxílio às equipes de análise investigativa, em especial na identificação das Entidades Nomeadas (EN).

Este trabalho apresenta o desempenho de duas ferramentas de Reconhecimento de Entidades Nomeadas (REN), LSTM+CNN e Spacy, aplicadas a um conjunto de documentos produzidos pela Polícia Federal brasileira em investigações de crime de Lavagem de Dinheiro. Foram realizadas pesquisas e entrevistas com os policiais a fim de identificar os desafios enfrentados neste tipo de investigação.

2. Trabalhos Relacionados

Em [van Banerveld et al. 2016] os autores apresentam uma ferramenta, denominada LES, desenvolvida com técnicas de PLN para auxiliar investigadores criminais na análise de grandes quantidades de informação textual de maneira mais eficiente e rápida. Fora avaliado o desempenho da ferramenta com diferentes métricas e apresentados os resultados experimentais com conjuntos de dados grandes e complexos. Os autores identificaram algumas dificuldades enfrentadas pelos analistas no uso de sistemas forenses para a investigação, como: tempo gasto para processar todos os dados apreendidos; os softwares forenses são incapazes de lidar com a enorme quantidade de dados de uma investigação; falha de software ao consultar os bancos de dados; tempo de espera inaceitável ao realizar uma consulta; muitos *hits* de busca para fazer a análise da evidência humanamente possível em muitos casos; e muita abordagem técnica na interface. Os principais ganhos no uso do sistema são apontados como sendo: melhoria no tempo de processamento de dados para manipular grandes quantidades de dados, melhoria no tempo de análise de conjuntos de dados complexos, e permitir que os usuários finais executem tarefas complexas com uma interface muito simples.

Técnicas de aprendizado de máquinas têm sido muito utilizadas em PLN, como exposto no artigo [do Amaral and Vieira 2014] em que as autoras utilizam o algoritmo *Conditional Random Fields* (CRF) para a tarefa de REN em corpora da língua portuguesa e avaliam comparativamente o desempenho desse método com outros sistemas, tendo como base o corpus do HAREM (avaliação conjunta na área do Reconhecimento de Entidades Mencionadas).

No trabalho desenvolvido por [Pires et al. 2015], são abordadas as formas aberta

e convencional de Extração de Informação em textos livres, onde os autores fazem uma comparação dos resultados das ferramentas, indicando uma maior eficiência nas ferramentas abertas. Por outro lado, os resultados apresentaram melhor precisão na técnica convencional, que depende de anotação. Dessa forma, os autores indicam uma análise em relação aos objetivos pretendidos para a utilização da técnica mais apropriada.

A extração e estruturação de relações abertas entre entidades nomeadas é abordada no artigo [Collovini et al. 2016]. Foi aplicado o modelo CRF para a extração de qualquer descritor de relações expressando qualquer tipo de relação entre um par de entidades nomeadas (categorias Pessoa, Lugar e Organização).

[Ku et al. 2008] propõe o desenvolvimento de uma ferramenta para coletar os depoimentos de vítimas e testemunhas de crimes, utilizando PLN para a extração de informações relevantes, visando auxiliar a análise investigativa. Os autores acreditam que as vítimas e testemunhas se sentem inibidas ao relatar os fatos à uma autoridade policial, e que se os dados fossem relatados em texto corrido, de forma anônima, poderia conter maior detalhe e canalizar para uma maior elucidação dos crimes por parte das equipes investigativas.

Seis anos após o trabalho anterior, os autores publicaram um novo trabalho. Em [Ku and Leroy 2014] foi identificada a dificuldade em lidar com os depoimentos anônimos de vítimas e testemunhas de crimes. Para solucionar o problema na identificação de correlação entre os relatos anônimos e determinados crimes, os autores desenvolveram um sistema de suporte à decisões, combinando técnicas de processamento de PLN, medidas de similaridade e aprendizado de máquina, com o uso de um classificador Naive Bayes, para apoiar análises criminais e classificar quais relatórios criminais possuem relação.

3. Dataset

O dataset utilizado é composto por 15 peças policiais produzidas pela Superintendência da Polícia Federal no Rio Grande do Sul, entre os anos de 2010 e 2011. A opção por textos mais antigos se deu em razão do sigilo legal imposto às informações mais recentes.

As peças policiais selecionadas foram submetidas à tarefa de anotação manual, seguindo os padrões utilizados em [Santos and Cardoso 2007], e com o auxílio da ferramenta GATE [Junior and de Carvalho], criando um corpus de referência para fins de avaliação. Obteve-se uma média de 928 palavras em cada peça policial, e um total de 504 Entidades Nomeadas identificadas (187 Pessoas, 167 Organizações e 150 Localizações).

O dataset está dividido em 05 Autos de Qualificação e Interrogatório, 05 Termos de Depoimento e 05 Termos de Declaração:

- **Auto de Qualificação e Interrogatório** (Figura 1): peça policial lavrada para a oitiva do indivíduo que é indiciado, quando há indícios suficientes que apontam ser este o autor do crime investigado.
- **Termo de Declaração** (Figura 2): peça policial utilizada para a oitiva de alguém que se presume ser o autor do crime investigado, mas ainda há dúvidas quanto à autoria.
- **Termo de Depoimento** (Figura 2): peça policial lavrada para a oitiva de testemunhas. A estrutura é a mesma do Termo de Declaração. Das três peças citadas, esta

é a única em que a pessoa se compromete a dizer a verdade, sob pena de incorrer em crime de falso testemunho.

**AUTO DE QUALIFICAÇÃO E INTERROGATÓRIO
DE: [NOME DO INTERROGADO]**

Ao(s) 27 dia(s) do mês de setembro de 2011, nesta Superintendência Regional do Departamento de Polícia Federal, em Porto Alegre/RS, onde se encontrava [NOME DO DELEGADO], Delegado de Polícia Federal, pelo(a) mesmo(a) foi determinado que se formalizasse a qualificação do(a) indiciado(a), o(a) qual RESPONDEU:

NOME: [NOME DO INTERROGADO]

ALCUNHA: não possui

NACIONALIDADE: brasileira

ESTADO CIVIL: casado(a)

PAI: [NOME DO PAI]

MÃE: [NOME DA MÃE]

DATA DE NASCIMENTO: [DATA]

NATURALIDADE: [CIDADE]

PROFISSÃO: [PROFISSÃO]

INSTRUÇÃO: Terceiro Grau Completo

DOCUMENTO DE IDENTIDADE: [RG]

CPF: [CPF]

RESIDÊNCIA: [ENDEREÇO]

ENDEREÇO COMERCIAL: [ENDEREÇO].

Cientificado(a) das imputações que lhe são feitas e de seus direitos constitucionais, inclusive o de permanecer calado(a), PERGUNTADO Qual a profissão e/ou atividade profissional desempenhada pelo interrogado? Qual a remuneração mensal média que recebe nessas atividades? RESPONDEU QUE irá exercer o seu direito de permanecer calado; PERGUNTADO Onde reside? Desde quando? Reside em imóvel próprio, alugado, cedido? RESPONDEU QUE irá exercer o seu direito de permanecer calado; PERGUNTADO se conhece os imóveis localizados nas ruas [ENDEREÇO], registrados

Figura 1. Exemplo de Auto de Qualificação e Interrogatório (os dados foram ocultados)

Verifica-se nos documentos produzidos em sede policial uma característica específica: as sentenças são delimitadas pela expressão “QUE”, que indica seu início, e finalizadas com um ponto-e-vírgula. Em relação às entidades, nota-se que, por convenção, as entidades Pessoas e Organização são escritas em caixa alta, como exemplificado nas Figuras 2 e 1.

**TERMO DE DECLARAÇÕES DE
[NOME DO DECLARANTE]:**

Ao(s) 13 dia(s) do mês de junho de 2011, nesta Superintendência Regional de Polícia Federal, em Porto Alegre/RS, onde se encontrava [NOME DO DELEGADO], Delegado de Polícia Federal, compareceu [NOME DO DECLARANTE], sexo masculino, nacionalidade [NACIONALIDADE], casado(a), filho(a) de [NOME DO PAI] e [NOME DA MÃE], nascido(a) aos [DATA DE NASCIMENTO], natural de [CIDADE], instrução ensino médio ou técnico profissional, profissão Desempregado(a), documento de identidade nº [Nº DA IDENTIDADE], CPF [Nº DO CPF], residente na(o) [ENDERECO], fone [TELEFONE], celular [CELULAR]. Inquirido a respeito dos fatos, RESPONDEU: **QUE**, primeiramente, gostaria de registrar que, inobstante todo o interesse que tenha de colaborar com a investigação, os fatos ora em questão já se passaram há muitos anos e, assim, sua memória já não guarda mais tantos detalhes; **QUE** tem a dizer quanto aos fatos é que muitos dos negócios que, na época dos fatos, envolveram seu nome, foram feitos na realidade em auxílio a outros operadores

Figura 2. Exemplo de Termo de Declarações (idêntico ao Termo de Depoimento, os dados foram ocultados)

Observa-se ainda, que o Auto de Qualificação e Interrogatório (Figura 1) possui um bloco de qualificação, contendo dados do interrogado, que o difere dos Termos de Declaração e Depoimento (Figura 2), onde a qualificação está inserida no texto corrido.

4. Pesquisa

Foi realizada uma pesquisa, por meio de formulário digital, com especialistas da Polícia Federal que atuam diretamente em investigações de crimes financeiros, a fim de identificar quais são as principais informações presentes nas peças policiais que poderiam auxiliar a análise e a identificação de autoria e materialidade dos crimes investigados.

A pesquisa foi respondida por 24 policiais, sendo 12 Delegados de Polícia Federal, 03 Peritos Criminais Federais, 06 Agentes de Polícia Federal e 03 Escrivães de Polícia Federal, todos atuantes nos setores de Combate a Crimes Financeiros.

As principais classes de entidades identificadas estão abaixo ordenadas por relevância:

- | | |
|---|---------------------------------------|
| 1. Pessoas físicas. | 8. Organizações públicas. |
| 2. Contas bancárias, títulos e valores mobiliários. | 9. Bens imóveis. |
| 3. Lugares e endereços. | 10. Período (tempo). |
| 4. Documentos e contratos públicos. | 11. Veículos. |
| 5. Documentos e contratos privados. | 12. Moeda nacional. |
| 6. Organizações privadas. | 13. Moeda estrangeira. |
| 7. Cargo ou função. | 14. Jóias, pedras e metais preciosos. |
| | 15. Obras de arte. |

As principais relações entre as entidades:

- | | |
|--|--------------------------------------|
| 1. É sócio somente no papel (laranja, p. ex.). | cial, p. ex.). |
| 2. É proprietário. | 4. É procurador. |
| 3. É sócio legal (em um contrato so- | 5. Outorgou procuração. |
| | 6. É servidor, funcionário ou empre- |

- gado.
7. É ou foi indicado (para um cargo ou função).
 8. É cônjuge ou companheiro(a).
 9. É parceiro ou auxiliar (conscientemente).
 10. É usuário.
 11. Conhece.
 12. É auxiliar (inconscientemente) ou
 - foi usado.
 13. É amigo.
 14. É credor.
 15. É devedor.
 16. É inimigo.
 17. É vendedor.
 18. Desconhece.
 19. Já ouviu falar.
- E, abaixo, os verbos que, segundo a pesquisa, possuem maior relevância para as investigações de crimes financeiros:
1. Ocultar.
 2. Subornar.
 3. Mandar, determinar ou influenciar.
 4. Auferir ganho, lucrar ou aproveitar.
 5. Adquirir.
 6. Transferir ou remeter.
 7. Pedir, solicitar, exigir ou receber.
 8. Comprar.
 9. Corromper ou enganar.
 10. Assinar.
 11. Usar (sem ser proprietário).
 12. Adulterar (uma coisa).
 13. Vender.
 14. Prejudicar, eliminar ou destruir.
 15. Subtrair.
 16. Doar.
 17. Intermediar.
 18. Emprestar ou ceder gratuitamente.
 19. Alugar.
 20. Outras informações que gostaria de acrescentar.

5. Reconhecimento de Entidades Nomeadas

Reconhecimento de Entidades Nomeadas (REN), segundo [Mohit 2014], é o problema de localizar e categorizar nomes importantes e nomes próprios em um texto livre. Por exemplo, em notícias, nomes de pessoas, organizações e locais são entidades relevantes. [Nadeau and Sekine 2007] definem Entidades Nomeadas (EN) como sendo termos que visam restringir designadores rígidos, que incluem nomes próprios e certos termos naturais, como espécies e substâncias biológicas, além de expressões temporais e algumas expressões numéricas, como quantias em dinheiro e outros tipos de unidades.

As classes de entidades selecionadas para a avaliação foram Pessoa, Organização e Localização, por serem categorias semânticas altamente utilizadas nos eventos de avaliação conjunta na língua portuguesa, como é o caso do HAREM [Santos and Cardoso 2007], e por apresentarem significativa relevância na investigação policial, como se depreende do resultado da pesquisa apresentada na Seção 4.

A LSTM+CNN é uma ferramenta de Reconhecimento de Entidades Nomeadas aplicada e aprimorada pelo grupo de PLN da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), que utiliza uma arquitetura híbrida, combinando uma rede recorrente *Long Short-Term Memory* (LSTM) [Hochreiter and Schmidhuber 1997] bidirecional com uma *Convolutional Neural Network* (CNN) [Kim 2014], treinada com o corpus LeNER-BR [de Araujo et al. 2018], um corpus anotado composto por textos jurídicos, extraídos de jurisprudências, e textos de legislações.

Spacy é uma ferramenta para PLN, de código aberto, escrita na linguagem Python, suporta mais de 33 línguas, possuindo 18 categorias de Entidades Nomeadas, sendo possível sua expansão. Utiliza-se de aprendizado de máquina e modelos de redes neurais convolucionais [Môro et al. 2018]. O corpus utilizado para treino foi o WikiNER em português, as bases provenientes do WikiNER [Nothman et al. 2013] são constituídas por diversos artigos da Wikipedia.

6. Avaliação

As 15 peças policiais selecionadas (seção 3) foram processadas pelas ferramentas LSTM+CNN e Spacy, e submetidas à avaliação de desempenho para as classes de Entidades Nomeadas: Pessoa, Organização e Localização. Os resultados obtidos foram os seguintes:

	Precisão	Abrangência	Medida-F
LSTM+CNN	0,70	0,69	0,68
Spacy	0,50	0,65	0,54

Figura 3. Resultado das ferramentas

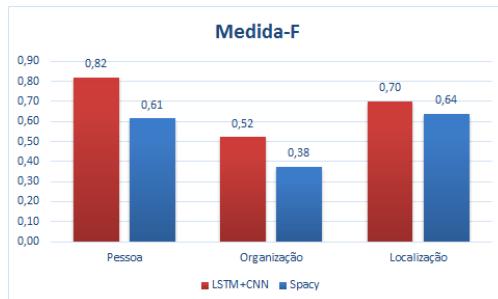


Figura 4. Resultado por classes de entidades

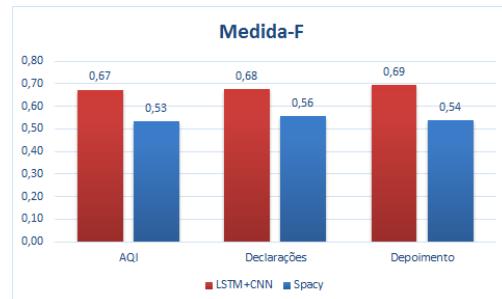


Figura 5. Resultado por tipo de peça policial

	LSTM+CNN						SPACY						Pessoa					
	Pessoa			Organização			Localização			Pessoa			Organização			Localização		
	Prec.	Abr.	M-F	Prec.	Abr.	M-F	Prec.	Abr.	M-F	Prec.	Abr.	M-F	Prec.	Abr.	M-F	Prec.	Abr.	M-F
AQI 01	0,91	0,83	0,87	0,50	0,25	0,33	1,00	0,42	0,59	0,64	0,58	0,61	0,30	0,38	0,33	0,50	0,75	0,60
AQI 02	0,89	0,89	0,89	0,73	0,57	0,64	0,89	0,73	0,80	0,67	0,44	0,53	0,45	0,64	0,53	0,39	1,00	0,56
AQI 03	0,83	0,83	0,83	0,53	0,39	0,45	0,50	0,67	0,57	0,77	0,74	0,76	0,52	0,61	0,56	0,41	0,87	0,55
AQI 04	0,98	0,82	0,86	0,50	0,67	0,57	0,89	0,67	0,76	0,63	0,45	0,53	0,27	0,33	0,30	0,58	0,92	0,71
AQI 05	0,88	0,94	0,91	0,50	0,36	0,42	0,47	0,80	0,59	0,78	0,44	0,56	0,36	0,64	0,46	0,27	0,80	0,40
Declarações 01	0,75	0,67	0,71	0,67	0,44	0,53	0,67	0,57	0,62	0,50	0,33	0,40	0,44	0,44	0,44	0,55	0,86	0,67
Declarações 02	0,95	0,78	0,86	0,60	0,67	0,63	0,43	0,67	0,52	0,77	0,43	0,56	0,21	0,44	0,29	0,58	1,00	0,72
Declarações 03	0,75	0,82	0,78	0,56	0,38	0,45	0,60	0,60	0,60	0,75	0,82	0,78	0,42	0,38	0,40	0,60	0,90	0,72
Declarações 04	0,84	0,75	0,80	0,64	0,60	0,62	0,75	0,67	0,71	0,75	0,38	0,51	0,44	0,53	0,48	0,36	0,89	0,52
Declarações 05	0,94	1,00	0,97	0,83	0,45	0,59	0,77	0,71	0,74	0,73	0,69	0,71	0,33	0,45	0,38	0,64	1,00	0,78
Depoimento 01	0,71	0,63	0,67	0,14	1,00	0,25	0,86	1,00	0,92	0,67	0,75	0,71	0,00	0,00	0,00	0,50	1,00	0,67
Depoimento 02	0,90	0,82	0,86	0,57	0,75	0,65	0,70	0,88	0,78	0,53	0,82	0,64	0,25	0,25	0,25	0,28	0,88	0,42
Depoimento 03	0,78	0,88	0,82	0,40	0,44	0,42	0,73	0,89	0,80	0,55	0,75	0,63	0,27	0,44	0,33	0,53	1,00	0,69
Depoimento 04	1,00	0,80	0,89	0,56	0,56	0,56	0,82	0,90	0,86	0,50	1,00	0,67	0,56	0,56	0,56	0,67	1,00	0,80
Depoimento 05	0,55	0,60	0,57	0,60	0,86	0,71	0,71	0,63	0,67	0,58	0,70	0,64	0,33	0,29	0,31	0,64	0,88	0,74
MÉDIA	0,84	0,80	0,82	0,55	0,56	0,52	0,72	0,72	0,70	0,65	0,62	0,61	0,34	0,43	0,38	0,50	0,92	0,64

Figura 6. Resultado detalhado de todo o experimento

Nota-se na Figura 3 que o desempenho foi melhor com a utilização da ferramenta LSTM+CNN, apresentando uma Medida-F de 0,68, contra 0,54 da Spacy.

Na Figura 5 são apresentados os resultados da Medida-F separados por tipo de peça policial. Pode-se verificar que o tipo da peça não interfere significativamente no resultado.

A Figura 4 apresenta os resultados da Medida-F em relação às classes de entidades avaliadas. Verifica-se um desempenho melhor na classe Pessoa, por outro lado, o pior desempenho foi verificado na classe Organização.

Na Figura 6 são apresentados os resultados de todo o experimento, avaliando a Precisão, Abrangência e Medida-F por peça policial e ferramenta. Verifica-se o bom resultado das ferramentas para a classe Pessoa, sendo que a LSTM+CNN obteve desempenho superior nesta tarefa. Por outro lado, os experimentos para a classe Organização obtiveram os piores resultados. Em relação à Localização, a ferramenta LSTM+CNN obteve melhor Precisão, enquanto a Abrangência foi superior na Spacy, sendo que o resultado da Medida-F da LSTM+CNN foi levemente superior para esta classe.

7. Conclusão e Trabalhos Futuros

Analisando os resultados expostos neste trabalho, verifica-se que a ferramenta LSTM+CNN obteve um desempenho geral melhor que a Spacy, superando os resultados médios em todas as avaliações, exceto quanto à medida de Abrangência da classe Localização. Observa-se, ainda, que o tipo de peça policial submetido à avaliação não interferiu significativamente nos resultados.

Verifica-se que a classe Pessoa foi a que obteve o melhor resultado para as duas ferramentas (Figura 4), por outro lado, a classe Organização apresentou o pior, impactando significativamente para a Medida-F geral (Figura 3).

A pesquisa realizada com Policiais Federais identificou a classe Pessoa como sendo a mais importante para auxiliar as investigações de crimes de Lavagem de Dinheiro, por ser um crime que envolve uma complexa estrutura de pessoas organizadas, sem que para isso necessitem estar próximas ou agirem juntas.

Dessa forma, conclui-se que as técnicas e ferramentas de Reconhecimento de Entidades Nomeadas, propostas neste trabalho, estão aptas a contribuir com o trabalho investigativo no combate ao crime de Lavagem de Dinheiro, na medida em que oferecem a possibilidade dos policiais identificarem de maneira mais eficiente a rede de pessoas envolvidas no cometimento deste delito, dado o grande volume de dados a serem analisados.

Como trabalhos futuros, visando aprimorar a tarefa de Reconhecimento das Entidades Nomeadas aplicada à investigação do crime de Lavagem de Dinheiro, sugere-se a anotação de um corpus específico para treinar as duas ferramentas e submetê-las à nova avaliação. Posteriormente, os resultados poderão ser submetidos à tarefa de extração das Relações entre as Entidades Nomeadas, explorando o elo entre as entidades aqui identificadas.

8. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

- Collovini, S., Machado, G., and Vieira, R. (2016). Extracting and structuring open relations from portuguese text. In *International Conference on Computational Processing of the Portuguese Language*, pages 153–164. Springer.
- de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer.
- de LEMOS JÚNIOR, A. P. (2007). Uma reflexão sobre as dificuldades da investigação criminal do crime de lavagem de dinheiro.
- do Amaral, D. O. F. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Junior, E. A. and de Carvalho, C. L. Processamento de linguagens naturais e a ferramenta gate.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Ku, C. H., Iribarri, A., and Leroy, G. (2008). Natural language processing and e-government: crime information extraction from heterogeneous data sources. In *Proceedings of the 2008 international conference on Digital government research*, pages 162–170. Digital Government Society of North America.
- Ku, C.-H. and Leroy, G. (2014). A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly*, 31(4):534–544.
- Mohit, B. (2014). Named entity recognition. In *Natural language processing of semitic languages*, pages 221–245. Springer.
- Môro, D. K. et al. (2018). Reconhecimento de entidades nomeadas em documentos de língua portuguesa.
- Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.
- Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.
- Pires, J. C. B. et al. (2015). Extração e mineração de informação independente de domínios da web na língua portuguesa.

- Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.
- van Banerveld, M., Kechadi, M.-T., and Le-Khac, N.-A. (2016). A natural language processing tool for white collar crime investigation. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIII*, pages 1–22. Springer.

Detecção Automática dos Heterônimos de Fernando Pessoa por Aprendizado de Máquina

Hugo Queiroz Abonizio¹, Cinthyan R. Sachs C. de Barbosa¹, Arthur A. Artoni¹

¹ Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina (UEL) – Londrina, PR – Brasil

{hugo.abonizio, cinthyan, arthurartoni}@uel.br

Abstract. *Text mining is an subarea of Natural Language Processing, which aims to find patterns on textual documents. An application of this pattern recognition is the authorship attribution. Therefore, this work proposes to apply the techniques of authorship identification to the texts of Fernando Pessoa's heteronyms. A machine learning method was used in order to assign a heteronym to each text based on the extracted features. With 95% accuracy, the model proposed in this work allowed the analysis of objective factors that distinguish heteronyms.*

Resumo. *A mineração de texto é uma área do Processamento de Linguagem Natural que busca encontrar padrões em documentos textuais. Uma aplicação desse reconhecimento de padrões é na atribuição de autoria dos textos. Dessa forma, este trabalho propõe aplicar as técnicas de identificação de autoria aos textos dos heterônimos de Fernando Pessoa. Foi utilizado um método de aprendizado de máquina a fim de atribuir um heterônimo para cada texto com base nas características extraídas. Com 95% de acerto, o modelo proposto neste trabalho possibilita a análise dos fatores objetivos que distinguem os heterônimos.*

1. Introdução

Com início na década de 50, o Processamento de Linguagem Natural (PLN) é uma área de estudo que forma uma interseção entre a Ciência da Computação com a Linguística [Hutchins 2004]. A ideia de usar computadores com habilidade de processar e analisar a linguagem humana já é antiga e estão presentes nas obras de ficção desde o século passado [Jurafsky and Martin 2000]. As aplicações de PLN variam desde interfaces para bancos de conhecimentos [Barbosa 1998], tradução automática gerada por máquina [Papineni et al. 2002] e classificação de documentos textuais [Sebastiani 2002].

Dentre as diversas aplicações dentro da área de PLN, a identificação de autoria tem como objetivo diferenciar, por meio de características textuais mensuráveis, textos escritos por diferentes autores [Stamatatos 2009]. A identificação de autoria, também chamada de verificação de autoria [Zheng et al. 2006], é largamente utilizada em análises forenses de documentos [Juola et al. 2008]. Em [Iqbal et al. 2010], por exemplo, são apresentados métodos que utilizam e-mails e verificação de autoria como evidência em investigações de crimes. Independente da língua em que o texto se apresenta, métricas objetivas podem ser extraídas dos documentos para caracterizar um autor [Kešelj et al. 2003].

Em [Zheng et al. 2006] é proposto um *framework* para identificação de autoria em múltiplas línguas com um conjunto de atributos composto por: características léxicas (tamanho médio das palavras, número de espaços em branco, etc.), sintáticas (frequência de

pontuações), estruturais (número total de linhas, quantidade de frases e parágrafos, etc.) e presença de palavras específicas. Esse conjunto de atributos é similar ao utilizado por [Abbasi and Chen 2006], onde também são extraídas características léxicas, sintáticas, estruturais e palavras associadas ao conteúdo para o inglês e o árabe.

Além disso, a identificação de autoria também é aplicada na detecção de contas comprometidas em redes sociais [Barbon et al. 2017] para identificar gênero, idade e nível de educação dos autores [El and Kassou 2014] e na análise de mensagens de grupos extremistas [Abbasi and Chen 2005]. Assim, fica evidente a importância do estudo da aplicação de técnicas de mineração de textos e PLN no estudo de autoria.

Em geral são utilizados diferentes autores para avaliar as metodologias de identificação de autoria. Assim, um desafio para a área aparece quando um mesmo autor utiliza de artifícios literários para escrever como outros autores, em um processo similar ao usado na pseudonímia. Esses autores incorporados por um mesmo ser humano são conhecidos como heterônimos e o autor mais proeminente no uso da técnica é Fernando Pessoa.

Fernando Pessoa, nascido no século XIX, foi um dos mais importantes escritores modernistas [Pessoa 2001]. Uma característica marcante do autor é o uso de heterônimos, que são autores fictícios que possuem uma personalidade própria. Dessa forma, os heterônimos trazem um desafio para a tarefa de atribuição de autoria, pois um mesmo autor pode produzir textos de diversas autorias fictícias.

Em [Teixeira and Couto 2015] os autores também utilizaram os heterônimos de Pessoa para identificação automática de autores. Entretanto, este artigo propõe avançar essa discussão utilizando os três principais heterônimos que são Ricardo Reis, Álvaro de Campos e Alberto Caeiro, ao invés de apenas dois deles. Além disso, este trabalho propõe uma análise dos dados obtidos a fim de determinar as diferenças objetivas nos textos.

Sendo assim, o objetivo deste trabalho é extrair métricas textuais que caracterizem os documentos, a fim de demonstrar características objetivas que diferenciem os heterônimos e subsidiem trabalhos futuros na área de sistemas de atribuição de autoria.

A seção 2 deste artigo apresenta a metodologia aplicada para a criação e validação do modelo proposto, detalhando os algoritmos e métricas utilizados. Na seção 3 são analisados os resultados e as variáveis preditoras mais importantes são interpretadas em relação a cada autor. Por fim, a seção 4 resume a pesquisa e apresenta direções futuras para este trabalho.

2. Materiais e Métodos

Considerando a finalidade de encontrar medidas objetivas para a classificação dos poemas entre seus autores, este trabalho propõe o uso de um modelo de aprendizado de máquina supervisionado, tendo como entrada um vetor de atributos e como saída o autor do dado documento.

O desenvolvimento do modelo foi dividido em quatro etapas: a aquisição dos dados, a extração das características, o treinamento do modelo de identificação de autoria e a avaliação de sua performance pela validação cruzada [Kohavi et al. 1995].

Primeiramente, foi coletada uma base de dados a partir do site Arquivo Pessoa¹.

¹<http://arquivopessoa.net/>

Foram coletados no total 598 trabalhos dos três heterônimos do escritor: 120 de Alberto Caeiro, 242 de Ricardo Reis e 236 de Álvaro de Campos. Ao total, o conjunto conta com 88.693 palavras com uma média de 148 palavras por poema e desvio padrão de 165, portanto é um conjunto com textos com uma grande diversidade de tamanhos.

Em seguida, na fase de pré-processamento dos dados, foram removidos caracteres especiais e ruídos advindos do processo de aquisição, além da remoção de *stopwords*. Posteriormente, foram extraídas características linguísticas de cada texto, as quais foram:

- frequência de uso de classes gramaticais (verbos, adjetivos, substantivos, dentre outras);
- quantidade de palavras por sentença;
- peso das principais palavras contidas no texto (TF-IDF);
- polaridade média do documento.

Para extrair os aspectos textuais dos textos, inicialmente foi feito um processo de análise léxico-morfológica, onde as palavras são avaliadas e separadas em classes gramaticais. Esse processo de atribuir rótulos às palavras é chamado de POS-tagging [Ratnaparkhi 1996] e foi realizado utilizando a biblioteca spaCy² da linguagem de programação Python. Cada palavra recebe uma classe (VERB para verbos, ADJ adjetivos, NOUN substantivos, etc.) que então contadas e extraída a frequência de uso no documento dividindo a contagem total pela quantidade de tokens. Posteriormente, foram extraídas as sentenças em cada documento para calcular um tamanho médio. Assim, podemos verificar se existe um autor que tem como estilo frases mais curtas ou mais longas. A Figura 1 mostra a árvore sintática de uma frase presente no corpus onde são apresentadas as classes gramaticais que são obtidas no processo de POS-tagging.

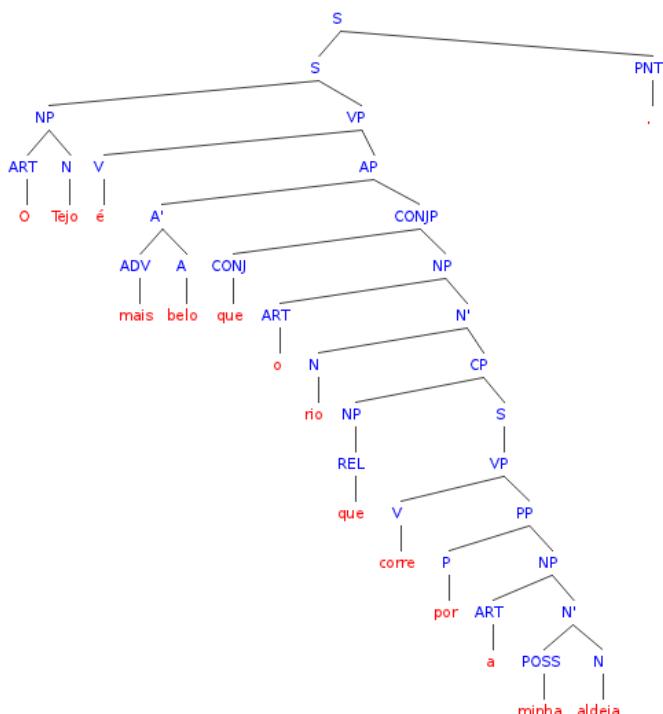


Figura 1. Árvore sintática do trecho "O Tejo é mais belo que o rio que corre pela minha aldeia".

²<https://spacy.io/>

Também foram extraídos os N termos mais frequentes ao longo do corpus pelo do método TF-IDF [Ramos et al. 2003], que atribui pesos para cada termo conforme sua frequência de uso no texto em relação ao uso em todo o corpus. Para cada termo é calculada a frequência que ele ocorre em um documento (TF) e a frequência total de todos os documentos do corpus (IDF). Assim, termos que aparecem muito em todos os documentos, como as *stopwords*, têm pesos baixos, enquanto termos que ocorrem apenas em alguns contextos terão pesos finais maiores.

Para o cálculo dos termos foram utilizados unigramas e bigramas, ou seja, tanto palavras únicas quanto combinações de duas palavras consecutivas como um só termo. No modelo final foram utilizados os 50 unigramas e bigramas mais frequentes.

Além disso, também foi extraída a polaridade dos documentos utilizando análise de sentimentos [Chen and Skiena 2014]. Dessa forma, cada termo no documento é associado um valor de polaridade sentimental que varia de -1 a 1, e para obter a polaridade do documento é feita uma média de todos os termos. Essa variável tem como hipótese encontrar diferenças de tom entre os autores.

Ao final, os atributos foram divididos em três grupos: as características linguísticas, os principais termos e o conjunto com todas as características, portanto são 16 características linguísticas, 50 unigramas e bigramas e 66 atributos no conjunto total.

Para a terceira etapa o algoritmo escolhido foi o Random Forest (RF), pois o algoritmo inherentemente traz a informação sobre a importância das variáveis na classificação, auxiliando no objetivo do trabalho em encontrar as características descritivas dos heterônimos. Além disso, é um método robusto aos ruídos e *outliers* no dataset [Breiman 2001]. O modelo foi implementado utilizando a biblioteca Scikit-learn [Pedregosa et al. 2011].

Para medir a performance em problemas de classificação, as previsões são marcadas como Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN). Assim, várias métricas podem ser calculadas com base nessa contagem. Para este trabalho, além da acurácia, descrita pela Equação 1, foram selecionadas outras três das principais métricas utilizadas em problemas de classificação [Alpaydin 2009]: precisão, revocação e *F1-score*. A precisão, descrita na Equação 2, demonstra o quão corretas estão as previsões, enquanto o revocação, descrito na Equação 3, retorna à fração dos documentos que foram de fato relevantes. A métrica *F1-score* é uma média ponderada da precisão e da revocação, descrita na Equação 4. Essas métricas foram selecionadas devido ao desbalanceamento das classes do problema, pois se fosse utilizada apenas a acurácia, o resultado poderia ser superestimado por conter uma classe com mais representantes que as outras.

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

$$Precisao = \frac{VP}{VP + FP} \quad (2)$$

$$Revocacao = \frac{VP}{VP + FN} \quad (3)$$

$$F1 = \frac{2 * Precisao * Revocacao}{Precisao + Revocacao} \quad (4)$$

Ao final, a fim de avaliar a performance do modelo, foi utilizada uma técnica de validação cruzada para indicar a taxa de acerto do modelo. Foi usado o método de validação cruzada estratificada k -fold [Arlot and Celisse 2010], com k igual a 5.

3. Resultados

Após cada iteração do modelo na validação cruzada foram coletadas as métricas de performance que são mostradas na Tabela 1 que contém a média das métricas de precisão, revocação e $F1$ -score para cada autor. As métricas apresentadas indicam que o modelo obteve uma performance alta nos três heterônimos, conseguindo uma média de $F1$ -score de 0.94 e acurácia de 95%.

Tabela 1. Resultados das métricas precisão, revocação e $F1$ -score em relação a cada autor e a média do resultado para as 3 classes avaliadas.

Autor	Precisão	Revocação	F1
Alberto Caeiro	0.90	0.93	0.92
Ricardo Reis	0.97	0.93	0.95
Álvaro de Campos	0.94	0.97	0.96
Média	0.94	0.94	0.94

A fim de comparação, também foi utilizado como referência um classificador que atribui, aleatoriamente, uma classe a cada amostra. Assim, os resultados desse classificador servem como base de comparação da abordagem que está sendo proposta, pois define uma performance mínima a ser alcançada. A Figura 2 mostra as performances medidas pelo $F1$ -score com validação cruzada.

Os resultados foram separados por grupo de atributos, ou seja, são apresentados os resultados para os modelos treinados apenas com as características linguísticas, apenas com os pesos dos principais termos (TF-IDF), todos os atributos juntos e o resultado base, que serve de referência. Assim, é possível visualizar que as características linguísticas e os pesos dos termos, quando usados separadamente, não atingem uma performance tão alta com as que são usadas em conjunto. Portanto, é possível concluir que os dois conjuntos de atributos se complementam na tarefa de identificar a autoria dos textos.

O gráfico na Figura 3 mostra a importância dos 10 atributos que tiveram maior relevância para a identificação dos autores. Esse valor é obtido por um método interno do algoritmo *Random Forest*, chamado de *out-of-bag error*, que quantifica o quanto cada atributo foi responsável na discriminação das classes [Breiman 2001].

Dos atributos mais importantes é interessante notar que não há uma predominância de um atributo ou grupo de atributos específico. Os quatro atributos considerados mais importantes (frequência de substantivos próprios (PROPN), o termo *deuses*, frequência de pontuações (PUNCT) e o termo *coisa*), têm seus valores apresentados na Figura 4.

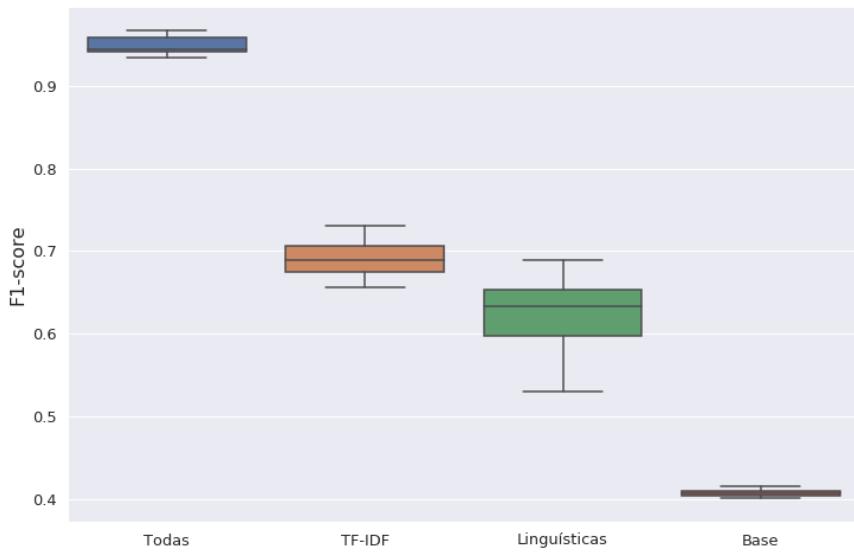


Figura 2. Boxplot dos *F1-score* dos modelos treinados no processo de validação cruzada com cada conjunto de atributos (características linguísticas, peso dos principais termos e uma combinação dos dois conjuntos), e a performance base resultado de uma classificação aleatória servindo de referência.

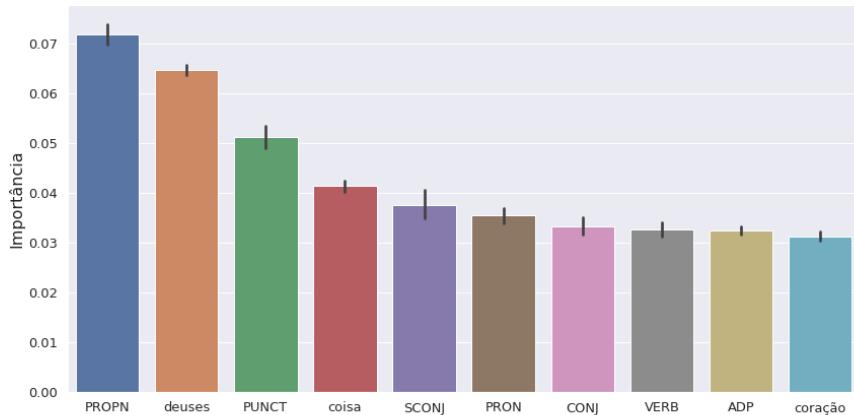


Figura 3. Importância dos 10 atributos mais importantes, extraídos dos modelos treinados com todos os atributos ao longo do processo de validação cruzada.

Dessa forma, é possível ver que as distribuições dos valores dos atributos são diferentes em cada autor, fazendo com que esse padrão seja utilizado para a identificação. Em relação à frequência de substantivos próprios, por exemplo, o heterônimo Ricardo Reis utiliza, em média, uma frequência maior, de 0.07, enquanto os outros dois possuem uma média 0.05.

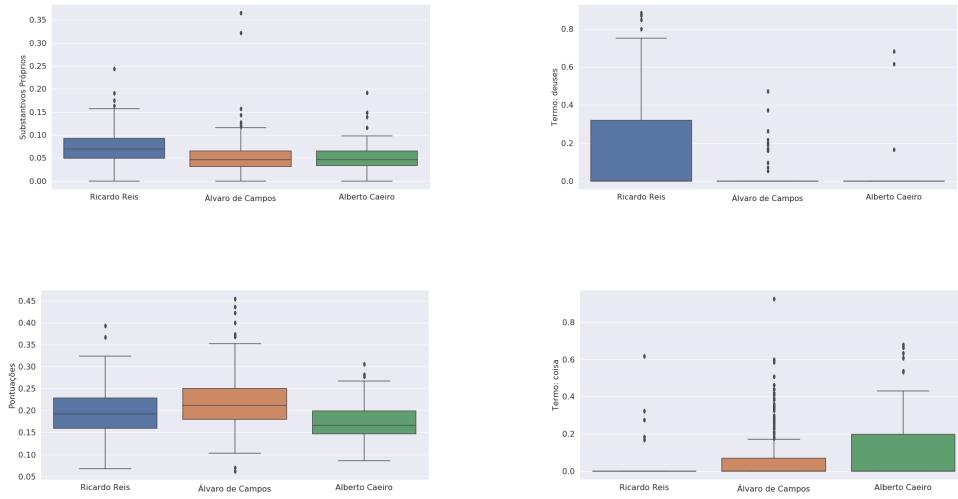


Figura 4. Distribuição dos principais atributos para cada heterônimo

Em relação ao uso dos termos *deuses* e *coisa*, é possível notar que o termo *deuses* tem um peso muito maior nos textos do heterônimo Ricardo Reis, pois este aparece 72 vezes em seus textos, enquanto aparece em 3 e 11 ocasiões em textos de Alberto Caeiro e Álvaro de Campos, respectivamente. O termo *coisa*, por sua vez, tem um peso médio maior para Alberto Caeiro, apesar de também estar presente em textos atribuídos a Álvaro de Campos.

Por meio da média da importância dos termos, também é possível verificar, por exemplo, que Alberto Caeiro tem deuses como tema importante de seus poemas. Apesar de afirmar no poema *Há metafísica bastante em não pensar em nada*, atribuído ao heterônimo, que não acredita em Deus, este é um termo recorrente e importante para caracterizar sua obra, como pode ser visto no excerto do poema:

“(...) *Não acredito em Deus porque nunca o vi. Se ele quisesse que eu acreditasse nele, (...) Mas se Deus é as flores e as árvores E os montes e sol e o luar, Então acredito nele, (...)Mas se Deus é as árvores e as flores E os montes e o luar e o sol, Para que lhe chamo eu Deus? (...)*”

Outro termo que apresenta uma grande diferença é coisa, que tem um peso muito maior para Alberto Caeiro do que para os outros. Um exemplo onde esse termo se mostra importante é no poema *O quê? Valho Mais que uma Flor* que possui 22 versos e a palavra coisa aparece em 6 deles. Os dois outros termos dos mais importantes para as predições foram vida e ter, onde o primeiro claramente é mais importante nos textos de Alberto Caeiro, enquanto o segundo tem um peso maior nos documentos de Álvaro de Campos.

Em relação aos demais atributos, não foi notada uma diferença suficiente entre as distribuições dos valores para serem destacados. Entretanto, os modelos treinados com todos os atributos disponíveis obtiveram um resultado muito superior aos treinados com apenas um dos grupos. Sendo assim, é possível concluir que o uso de características lingüísticas aliadas aos pesos dos principais termos alavanca o poder preditivo dos modelos.

A Tabela 2 apresenta a média e o desvio padrão para cada atributo em cada heterônimo para os 10 atributos considerados mais importantes pelos modelos.

Tabela 2. Média dos valores dos principais atributos e o desvio padrão (esse é exibido entre parênteses) para cada heterônimo.

Variável	Alberto Caeiro	Ricardo Reis	Álvaro de Campos
Frequência PROPN	0.05 (0.03)	0.07 (0.04)	0.05 (0.04)
Termo <i>deuses</i>	0.01 (0.08)	0.15 (0.25)	0.01 (0.05)
Frequência PUNCT	0.17 (0.04)	0.19 (0.05)	0.22 (0.06)
Termo <i>coisa</i>	0.11 (0.18)	0.01 (0.05)	0.07 (0.14)
Frequência SCONJ	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
Frequência PRON	0.09 (0.03)	0.08 (0.04)	0.07 (0.03)
Frequência CONJ	0.03 (0.02)	0.03 (0.02)	0.02 (0.01)
Frequência VERB	0.08 (0.04)	0.07 (0.03)	0.06 (0.03)
Frequência ADP	0.11 (0.04)	0.10 (0.04)	0.12 (0.04)
Termo <i>coração</i>	0.00 (0.03)	0.01 (0.07)	0.08 (0.19)

4. Conclusões e trabalhos futuros

Este trabalho apresentou um sistema de identificação de autoria composto por uma etapa de extração de características textuais e posteriormente a identificação do autor em textos em formato de poema. Foram apresentadas características que diferenciam os autores e foi feita uma validação do modelo de reconhecimento que apresentou um bom poder preditivo, obtendo uma acurácia média de 95%, mesmo com o desafio de realizar a atribuição de autoria em heterônimos de um mesmo autor. Também foram analisadas as diferenças apresentadas na forma de escrita de cada autor, interpretando as métricas objetivas nos textos de cada heterônimo.

Este trabalho propôs uma análise da estilometria das obras do autor por meio um método de aprendizado de máquina, pois assim foi possível quantificar as diferenças de escrita entre as obras dos diferentes heterônimos.

A alta taxa de acerto do modelo demonstra que as diferenças dos textos de Ricardo Reis, Alberto Caeiro e Álvaro de Campos não são apenas referentes aos temas tratados e à personalidade dos heterônimos, mas também às métricas objetivas do texto.

Como continuação deste trabalho, é possível expandir as características extraídas, explorando ainda mais a análise dos sentimentos e utilizar modelagem de tópicos dos documentos. Além disso, pode-se tentar generalizar o procedimento para outros autores que utilizaram a técnica da heteronímia, podendo ser explorada uma abordagem de classificação hierárquica dos autores e seus heterônimos.

Referências

- Abbasi, A. and Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.
- Abbasi, A. and Chen, H. (2006). Visualizing authorship for identification. In *International Conference on Intelligence and Security Informatics 4th Edition*, pages 60–71. Springer.
- Alpaydin, E. (2009). *Introduction to machine learning*. MIT Press.

- Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79.
- Barbon, S., Igawa, R. A., and Zarpelão, B. B. (2017). Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications*, 76(3):3213–3233.
- Barbosa, C. R. S. C. (1998). Gramática para consultas radiológicas em língua portuguesa. Master’s thesis, Instituto de Informática da Universidade Federal do Rio Grande do Sul.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.
- El, S. E. M. and Kassou, I. (2014). Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12).
- Hutchins, W. J. (2004). The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Iqbal, F., Binsalleeh, H., Fung, B. C., and Debbabi, M. (2010). Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation*, 7(1-2):56–64.
- Juola, P. et al. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.
- Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.
- Kešelj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics*, volume 3, pages 255–264.
- Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Pessoa, F. (2001). *The Selected Prose of Fernando Pessoa*. Grove Press.
- Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

- Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.
- Teixeira, J. F. and Couto, M. (2015). Automatic distinction of fernando pessoa's heteronyms. In *Portuguese Conference on Artificial Intelligence*, pages 783–788. Springer.
- Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*, 57(3):378–393.

Atribuição Autoral em Blogs e Redes Sociais

João Trevisan Martins¹, José Eleandro Custódio¹, Ivandré Paraboni¹

¹School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)

{joao.trevisan.martins, eleandro, ivandre}@usp.br

Abstract. This work addresses the computational task of authorship attribution (i.e., the task of identifying the author of a text from a set of possible candidates) using the document classification library fastText based on texts available from Brazilian blogs and social networks. More specifically, the study aims to evaluate authorship attribution models in scenarios with different numbers of candidate authors, and to assess the possible degradation of results as the number of candidates (or classes) to consider increases.

Resumo. Este trabalho implementa a tarefa computacional de atribuição autoral (i.e., a tarefa de identificar o autor de um texto a partir de um conjunto de candidatos possíveis) com uso da biblioteca de classificação de documentos fastText aplicada a textos provenientes de blogs e redes sociais brasileiras. De forma mais específica, o estudo objetiva avaliar modelos de atribuição autoral em cenários com diferentes números de autores candidatos, e a possível degradação dos resultados à medida em que o número de candidatos (ou classes) a considerar é expandido.

1. Introdução

A tarefa computacional de identificar o autor de um texto a partir de um conjunto de autores possíveis é conhecida por atribuição autoral (AA). Esta tarefa pode ser especificada da diversas formas [Gollub et al. 2013], mas neste trabalho será abordado o problema de identificar o autor de um texto dentre um conjunto fechado de candidatos para os quais são conhecidos exemplos de produção textual. Em outras palavras, o problema é aqui tratado como uma tarefa de aprendizado de máquina supervisionado.

Modelos computacionais de AA possuem inúmeras aplicações, incluindo questões de defesa de direitos autorais, identificação de crimes ou fraudes on-line e afins. A área sofreu grande expansão em anos recentes [Pokou et al. 2016, Savoy 2016, Ishihara 2017, Khan et al. 2017, Stamatatos 2017], e é tema recorrente da série de competições (ou *shared tasks*) PAN-CLEF [Rosso et al. 2016, Potthast et al. 2017].

Na construção de modelos de AA, há grande interesse em abordagens independentes de língua e domínio, e no uso de técnicas recentes do processamento de língua natural como os modelos de representação distribuída de palavras (ou *word embeddings*) [Mikolov et al. 2013], frequentemente associados a métodos de aprendizado neural [Ge and Sun 2016, Hitschler et al. 2017, Rocha et al. 2017, Brocardo et al. 2017]. Dentro destes, os recursos de classificação de documentos disponibilizados pela biblioteca fastText [Joulin et al. 2016] têm se demonstrado potencialmente úteis para vários tipos de problemas que envolvem grandes números de classes. Com base nestas observações,

este trabalho discute assim o problema computacional de AA em textos provenientes de blogs e redes sociais brasileiras com uso de modelos fastText. De forma mais específica, o estudo objetiva avaliar modelos de AA em cenários com diferentes números de autores candidatos, e a possível degradação dos resultados à medida em que o número de candidatos a considerar é expandido.

2. Trabalhos relacionados

Nesta seção são revisados alguns estudos recentes da área de AA que tomam por base representação distribuída (ou *embeddings*) de palavras ou caracteres. Cabe entretanto notar de imediato que nenhum destes estudos considera o uso de textos em português em seus experimentos.

2.1. Agrupamento autoral usando k-means

O trabalho em [Sari and Stevenson 2016] aplicou o método *k-means* para o problema de clusterização autoral (i.e., não supervisionado) no contexto da edição de 2016 da competição PAN-CLEF. A extração de características utilizou a base de dados de *embeddings* pré-treinados usando a base *Skip-gram-Google* disponível para as línguas inglesa e alemã, ou modelos de n-gramas de caracteres com peso TF-IDF. O parâmetro *k* do algoritmo K-means foi otimizado utilizando-se as medidas de silhueta e F1. Os resultados baseados em modelos de *embeddings* foram iguais ou inferiores aos dos modelos TF-IDF.

2.2. Modelos Doc2vec para conjuntos pequenos

O trabalho em [Posadas-Durán et al. 2017] aplicou o método Doc2vec [Le and Mikolov 2014] ao cenário de AA com poucos textos, os quais foram extraídos os córpus PAN2012, Reuters RCV1 e The Guardian. Neste modelo, cada documento foi dividido em sequências de n-gramas de palavras e transformado nas formas distribuídas Doc2vec-DM e Doc2vec-DBOW, e ambas as representações foram concatenadas.

O estudo analisou cenários com n-gramas de tamanho 1 a 5, e a classificação foi feita com regressão logística e SVM. Foram reportados 100% de acurácia para o problema PAN2012, enquanto que os conjuntos RCV1_CCAT_10 e RCV1_CCAT_50 apresentaram 84,6% e 75,2% de acurácia utilizando Doc2Vec com n-gramas de tamanho 1 a 3. O córpus The Guardian apresentou acurácia de 77%, contra 59,3% do *baseline* de trigramas de caracteres.

2.3. Modelos Doc2vec com múltiplos tipos de n-gramas

O trabalho em [Adorno et al. 2018] também apresentou um estudo utilizando Doc2vec [Le and Mikolov 2014] para AA de domínio fechado usando o córpus The Guardian. Foram utilizadas 10 amostras por autor, com média de 1500 caracteres e 1000 palavras por documento. O modelo foi treinado com n-gramas de palavras, caracteres e *part-of-speech* (POS) de tamanhos 1 a 5. Os vetores de *embeddings* de tamanho 100 resultantes do treinamento foram utilizados como entrada para regressão logística *softmax*, e demais parâmetros foram avaliados por meio de *grid search*. O número de dimensões do vetor de *embeddings* variou entre 50 e 350, o tamanho da janela dos n-gramas variou entre 3 e 19 e a frequência mínima entre 3 e 4.

A aplicação dos *embeddings* na classificação deu-se por meio de inferência e re-treinamento. Na inferência, o modelo treinamento foi utilizado para predição. No retreinamento, os documentos sob avaliação foram adicionados ao modelo Doc2vec e novas iterações foram executadas. Os melhores resultados foram reportados no caso da inferência e extração de característica usando unigramas e bigramas de palavras. O uso de n-gramas de POS trouxe resultados comparáveis ao uso de n-gramas de palavras e de n-gramas de caracteres.

2.4. Uso de embeddings fastText

O trabalho em [Sari et al. 2017] apresentou um dos poucos estudos do gênero a fazer uso de modelos de *embeddings* fastText [Joulin et al. 2016] baseados em caracteres e palavras. Nesta abordagem, modelos de *embeddings* para o inglês foram utilizados na tarefa de classificação, e os conjuntos Reuters RCV1, IMDb62 e um córpus com textos jurídicos foram utilizados para teste. Os resultados foram comparados com outros estudos e a proposta atingiu patamares similares aos modelos convencionais.

2.5. Embeddings de caracteres com redes convolucionais

Finalmente, o trabalho em [Shrestha et al. 2017] fez uso de redes convolucionais (CNNs) na tarefa de AA em dados de mídia social utilizando n-gramas de caracteres. Nesta abordagem foi utilizado um conjunto de dados do domínio Twitter, o qual foi pré-processado para remoção de nomes de usuários, URLs e dígitos.

A arquitetura proposta consistiu da camada de entrada usando *embeddings* de caracteres, da camada de convolução e da camada de saída *softmax*. Foram utilizados *embeddings* com 300 dimensões, mas não detalhou-se o seu processo de construção. A camada de convolução consistiu-se de filtros de n-gramas de tamanhos 3, 4 e 5, e *pooling* com função *max-over-time* com janela de tamanho 500. Foi utilizado o algoritmo de otimização Adam [Kingma and Ba 2015] e regularização com 0,25 de *dropout*.

O trabalho avaliou os modelos CNN com *embeddings* de unigramas de caracteres (CNN-1), CNN com *embeddings* de bigramas de caracteres (CNN-2), CNN com *embeddings* de palavras *Skip-gram-Google* (CNN-W), o modelo tradicional de n-gramas de caracteres de tamanhos 2 a 4 com regressão logística (CHAR), e por fim LSTM com bigramas de caracteres. De modo geral, as redes CNN-2 e CNN-1 obtiveram os melhores resultados, as redes LSTM-2 obtiveram resultados inferiores aos do *baseline* considerado.

3. Trabalho realizado

O presente trabalho teve como objetivo a criação e comparação de modelos computacionais de AA para textos provenientes de blogs e redes sociais brasileiras, e utilizando modelos do tipo fastText [Joulin et al. 2016] para avaliar cenários com diferentes números de autores candidatos. A seguir descrevemos os conjuntos de dados utilizados e os modelos propriamente ditos.

3.1. Conjuntos de dados

Os documentos considerados no experimento descrito a seguir são provenientes de três domínios: Blogs pessoais (aqui denominado domínio *blog*), postagens Facebook

(domínio *post*) e Twitter (*twit*). Nos três casos, apenas textos em português brasileiro foram considerados.

Para o domínio *blog*, foram selecionados blogs pessoais provenientes do córpus BlogSet-BR [dos Santos et al. 2018], uma coleção de Blogs brasileiros com cerca de 7,4 milhões de postagens escritas por cerca de oitocentos mil autores. Para o domínio *post*, foi utilizado o córpus *b5-post* [Ramos et al. 2018], uma coleção de publicações produzidas por 1020 usuários do Facebook brasileiro. Finalmente, para o domínio *Twitter* foi utilizado um conjunto de mensagens (*tweets*) da rede Twitter do Brasil coletados em um projeto prévio.

Os textos dos três domínios foram pré-processados para remoção de ruído (e.g., código HTML etc.) e adequação ao formato exigido pela ferramenta *fastText*, e também para inclusão de um rótulo identificador de cada autor (correspondendo à classe a ser aprendida) tal qual disponibilizado pelo córpus de origem. Para cada domínio, foram criados conjuntos de dados contendo 5, 10 e 20 autores, selecionados aleatoriamente dentre os autores com maior volume de texto em cada córpus. Estatísticas descritivas dos conjuntos de dados *blog*, *post* e *twit* de diferentes tamanhos são apresentadas na Tabela 1.

Tabela 1. Conjuntos de dados utilizados nos experimentos

Conjunto	Domínio	Autores	Palavras
blog-5	Blog	5	169.925
blog-10	Blog	10	227.892
blog-20	Blog	20	343.338
post-5	Facebook	5	9.596
post-10	Facebook	10	14.312
post-20	Facebook	20	20.502
twit-5	Twitter	5	28.491
twit-10	Twitter	10	53.347
twit-20	Twitter	20	94.121

4. Modelos Desenvolvidos

A tarefa de AA para cada um dos conjuntos de dados definidos na seção anterior foi conduzida com uso da ferramenta *FastText* [Joulin et al. 2016] para classificação de documentos, e utilizando-se *embedding* pré-treinados para o português brasileiro disponibilizados em [Hartmann et al. 2017]. Em todos experimentos, foram utilizados *embeddings* de tamanho 300 por ser esta uma dimensão intermediária e de uso frequente na área.

Dentre os diversos parâmetros configuráveis do modelo *fastText*, foram investigados três em especial: a taxa de aprendizado, o número de épocas e o tamanho dos *word n-grams*. A lista de parâmetros e valores considerados é resumida na Tabela 2.

Tabela 2. Valores utilizados nos parâmetros variados

Parâmetro	Valores considerados
Taxa de Aprendizado	0,2, 0,4, 0,6, 0,8 e 1,0
Número de Épocas	10, 20, 30, 40 e 50
Word n-grams	1, 2 e 3

Os valores ótimos destes parâmetros foram selecionados por meio de *grid search* em uma porção de dados de desenvolvimento. Estes valores correspondem a 30 épocas, taxa de aprendizagem de 0,4 e utilizando-se unigramas. O modelo resultante desta configuração será denominado *fastText*, e seus resultados serão comparados com um *baseline* utilizando regressão logística sobre um modelo do tipo *bag-of-words* com remoção de *stopwords* do português.

5. Avaliação

Nas Tabelas 3, 4 e 5 são apresentados os valores médios de medida F1 obtidos para os três domínios de teste pelos modelos *baseline* e *fastText*. Em todos os casos, os melhores resultados de cada conjunto de dados são destacados.

Tabela 3. Medida F1 média para textos de blogs.

Conjunto	Baseline	fastText
blog-5	0,86	0,92
blog-10	0,76	0,89
blog-20	0,69	0,81
média	0,80	0,87

Tabela 4. Medida F1 média para textos Facebook.

Conjunto	Baseline	fastText
post-5	0,44	0,47
post-10	0,37	0,35
post-20	0,31	0,29
média	0,37	0,37

Tabela 5. Medida F1 média para textos do Twitter.

Conjunto	Baseline	fastText
twit-5	0,79	0,98
twit-10	0,98	0,97
twit-20	0,96	0,95
média	0,91	0,97

A partir destes resultados, observa-se que, de modo geral, os modelos *fastText* tendem a apresentar melhor resultado médio global, embora a diferença seja por vezes

pequena. A exceção neste caso é o domínio Facebook, cujo desempenho médio foi similar ao do modelo de *baseline*.

Com relação às diferenças entre domínios, observa-se que a tarefa de AA pode apresentar graus de dificuldade variados a depender do tipo de texto considerado. Em um extremo estão os textos do domínio Facebook, que apresentaram desempenho muito inferior aos demais; os textos de Blog mantêm uma posição intermediária, e no outro extremo está o domínio Twitter, com resultados em alguns casos próximos a 100% de acurácia. Uma possível explicação para este resultado seria uma combinação de dois fatores: diferenças em volume de dados de treinamento (já que o córpus Facebook é consideravelmente menor do que os demais) e a relativa homogeneidade do domínio Twitter (que contrasta com a maior variedade de tópicos e fontes textuais dos textos de blogs).

Finalmente, e conforme já esperado, observa-se que o aumento do número de autores candidatos (ou classes) degrada o desempenho de todos os classificadores. A figura 1 ilustra este efeito no caso específico dos modelos *fastText* construídos.

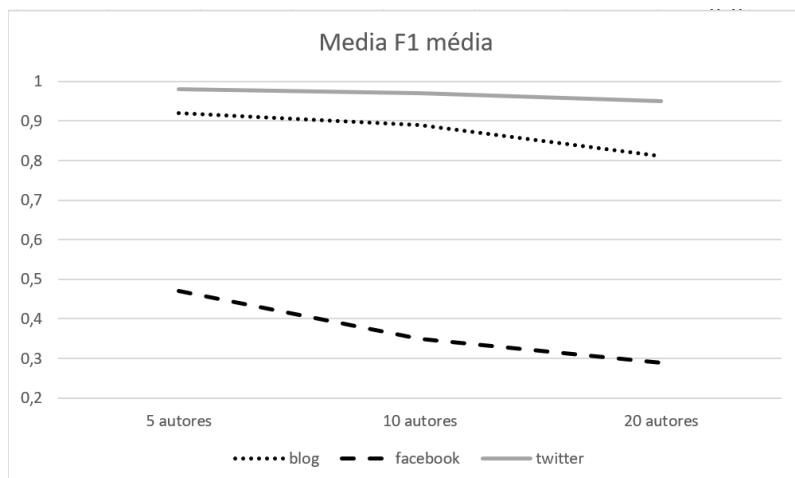


Figura 1. Variação da medida F1 média e número de autores candidatos para modelos fastText de AA.

6. Considerações finais

Este trabalho apresentou uma avaliação de modelos de AA de *embeddings* treinados com uso da ferramenta *fastText* [Joulin et al. 2016] de classificação de documentos. A avaliação fez uso de textos provenientes de blogs e redes sociais brasileiras, e considerou cenários com diferentes quantidades de autores candidatos.

Os resultados obtidos sugerem que modelos *fastText* são em princípio adequados para a modelagem do problema de AA, embora o desempenho sofra redução à medida em que grupos maiores de candidatos são considerados. Como trabalho futuro, pretende-se investigar técnicas para minimizar esta degradação e desenvolver modelos de AA capazes de lidar com grandes massas de autores potenciais de forma conjunta.

Agradecimentos Esta pesquisa contou com suporte FAPESP nro. 2017/20968-0 e 2016/14223-0, e da Universidade de São Paulo.

Referências

- Adorno, H. G., Posadas-Durán, J. P., Sidorov, G., and Pinto, D. (2018). Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):1–16.
- Brocardo, M. L., Traore, I., Woungang, I., and Obaidat, M. S. (2017). Authorship verification using deep belief network systems. *International Journal of Communication Systems*, 30(12).
- dos Santos, H. D. P., Woloszyn, V., and Vieira, R. (2018). BlogSet-BR: A Brazilian Portuguese Blog Corpus. In *11th International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. ELRA.
- Ge, Z. and Sun, Y. (2016). Domain specific author attribution based on feedforward Neural Network Language Models. In *5th International Conference on Pattern Recognition Applications and Methods (ICPRAM-2016)*, pages 597–604. SciTePress.
- Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., and Stein, B. (2013). Recent trends in digital text forensics and its evaluation: Plagiarism detection, author identification, and author profiling. In *LNCS 8138*, pages 282–302.
- Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *11th Brazilian Symposium in Information and Human Language Technology - STIL*, pages 122–131, Uberlândia, Brazil.
- Hitschler, J., van den Berg, E., and Rehbein, I. (2017). Authorship Attribution with Convolutional Neural Networks and POS-eliding. *Association for Computational Linguistics*, pages 53–58.
- Ishihara, S. (2017). Strength of forensic text comparison evidence from stylometric features: A multivariate likelihood ratio-based analysis. *International Journal of Speech, Language and the Law*, 24(1):67–98.
- Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Khan, F. A., Tahir, M. A., Khelifi, F., Bouridane, A., and Almotaeryi, R. (2017). Robust off-line text independent writer identification using bagged discrete cosine transform features. *Expert Systems with Applications*, 71:404–415.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations - ICLR 2015*, page 13. Ithaca, NY: arXiv.org.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proc. of Machine Learning Research 32(2)*, pages 1188–1196, Beijing, China. PMLR.
- Mikolov, T., Wen-tau, S., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proc. of NAACL-HLT-2013*, pages 746–751, Atlanta, USA. Association for Computational Linguistics.
- Pokou, Y. J. M., Fournier-Viger, P., and Moghrabi, C. (2016). Authorship attribution using small sets of frequent part-of-speech skip-grams. In *29th International Florida Artificial Intelligence Research Symposium (FAIR)*, Orlando, USA. FAIR.

- ficial Intelligence Research Society Conference (FLAIRS-2016), pages 86–91. AAAI Press.
- Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., and Chanona-Hernández, L. (2017). Applications of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3):627–639.
- Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., and Stein, B. (2017). Overview of PAN 17: Author identification, author profiling, and author obfuscation. In *LNCS 10456*, pages 275–290.
- Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., and Dias, R. F. S. (2018). Building a corpus for personality-dependent natural language understanding and generation. In *11th International Conference on Language Resources and Evaluation (LREC-2018)*, pages 1138–1145, Miyazaki, Japan. ELRA.
- Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A. R. B., and Stamatatos, E. (2017). Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33.
- Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., and Stein, B. (2016). Overview of PAN 16: New challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In *LNCS 9822*, pages 332–350.
- Sari, Y. and Stevenson, M. (2016). Exploring Word Embeddings and Character N -Grams for Author Clustering Notebook for PAN at CLEF 2016. *CEUR Workshop Proceedings*.
- Sari, Y., Vlachos, A., and Stevenson, M. (2017). Continuous N-gram Representations for Authorship Attribution. In *EACL-2017*, volume 2, pages 267–273. Association for Computational Linguistics.
- Savoy, J. (2016). Estimating the probability of an authorship attribution. *Journal of the Association for Information Science and Technology*, 67(6):1462–1472.
- Shrestha, P., Sierra, S., González, F., Rosso, P., Montes-Y-Gómez, M., and Solorio, T. (2017). Convolutional Neural Networks for Authorship Attribution of Short Texts. In *EACL-2017*, pages 669–674. Association for Computational Linguistics.
- Stamatatos, E. (2017). Authorship attribution using text distortion. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017)*, Valencia, Spain. Association for Computational Linguistics.

Reavaliando o dicionário LIWC português: o caso do reconhecimento de traços de personalidade e gênero autoral

Ricelli Moreira Silva Ramos¹, Ivandré Paraboni¹

¹School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)

{ricelliramos, ivandre}@usp.br

Abstract. *We present an experiment that compares the use of the LIWC dictionary for Portuguese (containing 64 lexical categories) and English (with 93 categories) applied to the personality recognition and author gender profiling tasks. The goal of the experiment is to verify whether the lower coverage of the Portuguese dictionary may represent a drawback if compared to similar studies developed for the English language, or whether it may be still applied to lexical tasks of this kind without significant losses.*

Resumo. *Propõe-se um experimento que compara o uso dos dicionários LIWC para o português (contendo 64 categorias lexicais) e inglês (com 93 categorias) aplicados às tarefas de reconhecimento de traços de personalidade e classificação de gênero autoral. O objetivo do experimento é verificar se a menor cobertura do dicionário português representa uma desvantagem em relação a estudos similares desenvolvidos para o inglês, ou se este recurso pode ser empregado sem prejuízo em tarefas de natureza lexical deste tipo.*

1. Introdução

Em anos recentes, o reconhecimento computacional de traços da personalidade humana - aqui entendida como um conjunto estável de características individuais representando padrões do comportamento em grande parte previsíveis [Allport and Allport 1921] - tem despertado grande interesse na pesquisa em Processamento de Línguas Naturais (PLN). Este interesse se deve, dentre outros motivos, à observação prática de que usuários de sistemas computacionais não apenas atribuem características humanas às máquinas com as quais interagem, como também preferem aquelas que demonstram ter uma personalidade semelhante à sua própria [Mairesse et al. 2007].

Traços de personalidade fundamentais são consistentemente refletidos nas escolhas lingüísticas que um indivíduo faz ao se comunicar, e a relação entre personalidade e língua natural é assim explorada por diversos modelos de personalidade propostos pela Psicologia. Em especial, consideramos neste trabalho o modelo dos Cinco Grandes Fatores (CGF) ou *Big Five* [Goldberg 1990], que contempla cinco dimensões fundamentais amplamente aceitas como a base adequada para a representação da personalidade humana: Abertura à experiência, Conscienciosidade, Extroversão, Agradabilidade e Neuroticismo [de Andrade 2008].

O estudo da personalidade humana foi a principal motivação para o desenvolvimento de recursos como o dicionário LIWC (*Linguistic Inquiry and Word Count*) [Pennebaker et al. 2001]. Este recurso contempla categorias lexicais como ‘dinheiro’,

‘família’, ‘emoção negativa’ etc. que são correlacionadas a traços de personalidade e, não por acaso, a base de um grande número de abordagens computacionais de reconhecimento de personalidade a partir de texto [Mairesse et al. 2007, Iacobelli et al. 2011].

O estudo em [Balage Filho et al. 2013] demonstrou que o dicionário LIWC português possui desempenho satisfatório na tarefa de análise de sentimentos. Entretanto, uma vez que uma das principais motivações para a criação do dicionário LIWC é na verdade a tarefa de reconhecimento de traços de personalidade a partir de textos [Pennebaker et al. 2001], observa-se que o dicionário LIWC português ainda carece de uma avaliação sob esta perspectiva. Além disso, observa-se que, em sua versão mais recente, o dicionário LIWC inclui 93 categorias lexicais para o inglês, e que no caso do português [Balage Filho et al. 2013] este recurso é limitado a 64 categorias baseadas em uma versão anterior do dicionário. Assim, o dicionário LIWC português não contempla categorias mais recentes como ‘afiliação’, ‘poder’, ‘risco’, ‘recompensa’, ‘informal’ e várias outras potencialmente úteis à tarefa de reconhecimento de traços de personalidade.

Embora não tenha se demonstrado significativa para a tarefa de análise de sentimentos em [Balage Filho et al. 2013], a maior cobertura do dicionário LIWC inglês é apontada em [Isbister et al. 2017] como a possível explicação para o melhor desempenho da tarefa de classificação de gênero autoral para o inglês em relação a outros idiomas. Sob a perspectiva do presente trabalho, coloca-se assim a questão de como outras tarefas para o português se comparam às desenvolvidas para o inglês no dicionário LIWC.

Com base nestas observações, este trabalho estende a análise apresentada em [Balage Filho et al. 2013] com o acréscimo duas novas tarefas desenvolvidas utilizando-se o dicionário LIWC português - o reconhecimento de traços de personalidade propriamente dito, e a tarefa complementar de classificação de gênero autoral (masculino ou feminino) a partir de textos - e comparar seus resultados com os de tarefas similares fazendo uso do dicionário LIWC inglês.

De forma mais específica, o presente estudo descreve um experimento que compara o desempenho de modelos construídos a partir de textos em português (e utilizando o dicionário LIWC próprio, de 64 categorias) com modelos construídos a partir de textos traduzidos para o inglês (utilizando o dicionário LIWC inglês de 93 categorias). O objetivo desta comparação é o de verificar se a menor cobertura do dicionário português representa uma desvantagem em relação a estudos similares desenvolvidos para o inglês, tal qual sugerido em [Isbister et al. 2017], ou se o presente método pode ser considerado uma alternativa válida para tarefas de PLN de natureza lexical deste tipo.

O restante deste artigo está organizado da seguinte forma. A seção 2 faz um breve levantamento de estudos na área de reconhecimento de traços de personalidade e classificação de gênero autoral a partir de textos. A seção 3 descreve os experimentos propostos, cujos resultados são apresentados na seção 4. Finalmente, a seção 5 apresenta as conclusões do estudo e sugestões de trabalhos futuros.

2. Trabalhos relacionados

2.1. Reconhecimento de traços de personalidade a partir de textos

O reconhecimento de traços de personalidade a partir de textos é tipicamente modelado como um problema de aprendizado de máquina supervisionado, ou seja, baseado

em córpus de textos rotulado com informações (classes) de personalidade. Estes rótulos podem ser estimados por diversos métodos consagrados em Psicologia, sendo o mais comum o uso de inventários (ou questionários) de traços de personalidade como o BFI-44 [John et al. 1991], composto de 44 itens na forma de frases breves contendo adjetivos que capturam os aspectos mais essenciais de cada fator do modelo *CGF*, como ‘é prestativo e ajuda os outros’. Cada item do inventário é respondido em uma escala de 1 (discordo totalmente) a 5 (concordo totalmente), e as respostas de determinados itens são posteriormente combinadas por meio de uma média ponderada para compor um valor escalar representando cada uma das dimensões *CGF*.

O inventário BFI-44 tem sido replicado em dezenas de outros idiomas, incluindo alguns estudos dedicados ao português. Em especial, o estudo em [de Andrade 2008] validou este inventário para o português brasileiro por meio de uma análise fatorial envolvendo uma amostra de 5.089 respondentes das cinco regiões brasileiras. O inventário considerado em [de Andrade 2008], denominado IGFP-5, foi empregado também na construção do córpus utilizado no presente trabalho.

Um dos estudos mais influentes na área de reconhecimento de personalidade para o inglês é o trabalho em [Mairesse et al. 2007], que utiliza textos escritos e transcrições de fala. Dentre as várias contribuições apresentadas, destaca-se o uso de conhecimento psicolinguístico proveniente dos dicionários LIWC [Pennebaker et al. 2001] e MRC [Coltheart 1981], e a comparação entre métodos de classificação, regressão e *ranking* para a tarefa. A acurácia média dos modelos propostos varia de 50% a 62% com o uso de máquinas de vetores de suporte (SVM).

Como alternativa ao uso de conhecimento psicolinguístico, o estudo em [Nowson and Oberlander 2007] faz uso de modelos baseados unicamente em contagens de n-gramas. O estudo se concentrou na tarefa de reconhecimento de indivíduos nos extremos de personalidade (que são mais fáceis de identificar do que os casos intermediários) e fez uso de métodos do tipo Naive-Bayes e SVM. A acurácia média obtida em um córpus de 1600 blogs pessoais foi de cerca de 65%.

Diversos estudos de reconhecimento de personalidade a partir de textos do Twitter rotulados com informações obtidas por meio de um inventário de personalidade simplificado (de apenas dez itens) [Rammstedt and John 2007] foram apresentados no contexto da série de competições PAN-CLEF [Rangel et al. 2015]. Estudos deste tipo incluem o vencedor da edição 2015 da competição em [Álvarez-Carmona et al. 2015], que combina atributos de segunda ordem representando relações entre palavras, documentos e perfis autorais com modelos de análise de semântica latente (LSA); o uso de n-gramas de caracteres e *part-of-speech* (POS) em [González-Gallardo and et. al. 2015], e o uso de contagens TF-IDF e atributos estilísticos (e.g., pontuação, uso de emoticons etc.) em [Şulea and Dichiù 2015], dentre outros.

Mais recentemente, métodos de aprendizado profundo combinados com córpus de grandes proporções, que se tornaram comuns na área de PLN, começam a ser aplicados também ao problema de reconhecimento de traços de personalidade. Estudos deste tipo incluem, por exemplo, a abordagem composicional baseada em *Gated Recurrent Units* (GRUs) bidirecionais em [Liu et al. 2017], que faz uso de modelos baseados em caracteres para construir representações de palavras e subsequentemente sentenças

com o objetivo de reconhecer traços de personalidade a partir do córpus PAN-2015 [Rangel et al. 2015].

Finalmente, observamos que, embora o modelo CGF seja predominante na literatura científica, há estudos que abordam modelos alternativos. Dentre estes, o mais popular é provavelmente o modelo Myers-Briggs Type Indicator (MBTI) [Myers and Myers 2010], amplamente utilizado para recrutamento de recursos humanos. O modelo MBTI define 16 tipos de personalidade ao longo de quatro dimensões (Extroversão e Introversão; Sensorial e Intuição; Razão e Sentimento e Julgamento e Percepção), que podem ser convenientemente modeladas na forma de problemas de classificação binária, como em [Verhoeven et al. 2016].

2.2. Classificação de gênero autoral

A classificação de gênero autoral (masculino e feminino) é uma das tarefas mais populares na área de caracterização autoral (do inglês, *author profiling*). A tarefa é normalmente tratada como um problema de aprendizado de máquina supervisionado, tendo sido também tema de uma competição do tipo *shared task*, a PAN-CLEF 2017 [Rangel et al. 2017]. Alguns estudos deste tipo são brevemente relatados a seguir.

Os estudos em [Basile et al. 2017, Martinc et al. 2017, Sierra et al. 2017] foram desenvolvidos para participação da competição PAN-CLEF 2017 e tratam, além da tarefa de detecção de variedade linguística, da classificação de gênero autoral no domínio Twitter em inglês, espanhol, árabe e português. O trabalho em [Basile et al. 2017], que apresenta o modelo mais simples dos três, fez uso de n-gramas de caracteres e palavras e classificação do tipo SVM. Este trabalho foi o vencedor da competição, apresentando resultados ligeiramente superiores ao uso de informação do tipo *part-of-speech* (POS) em [Martinc et al. 2017]. A abordagem mais sofisticada das três, em [Sierra et al. 2017], faz uso de *word embeddings* e redes convolutivas, mas teve desempenho abaixo de 10 outros participantes (de um total de 22).

Informações de POS são também empregadas em [Reddy et al. 2017]. Neste trabalho, um modelo de n-gramas de POS ponderado por contagens de TF-IDF supera diversas alternativas simples, como modelos do tipo ‘*bag of words*’ e outros, na classificação de gênero autoral em recomendações de hotel do site *Trip Advisor*.

Finalmente, o estudo em [Isbister et al. 2017] é um dos poucos a considerar o uso de conhecimento psicolinguístico disponibilizado pelo dicionário LIWC [Tausczik and Pennebaker 2010] na tarefa de classificação de gênero autoral. O estudo avalia o papel de diferentes categorias lexicais na predição de gênero, e as diferenças entre dicionários deste tipo disponibilizados em cinco idiomas, concluindo que a maior cobertura da versão em inglês representa uma vantagem em relação as outras versões.

3. Trabalho realizado

3.1. Visão geral

O presente trabalho propõe um método em que textos originalmente em português são submetidos à tradução automática para o inglês usando o Google Tradutor¹ sem pós-edição, e então processados com uso do dicionário LIWC2015 inglês (de 93 categorias)

¹<https://translate.google.com.br/>

no contexto de duas tarefas de PLN: o reconhecimento de traços de personalidade e a classificação de gênero autoral (masculino / feminino). Em ambos os casos, o modelo baseado em tradução, aqui denominado LIWC-En, é comparado ao modelo baseado nos textos originais em português e fazendo uso do dicionário próprio (de 64 categorias), aqui denominado LIWC-Pt. O objetivo do experimento é assim o de comparar o desempenho dos modelos LIWC-En e LIWC-Pt nas duas tarefas em questão.

A tarefa de reconhecimento de personalidade foi modelada na forma de cinco problemas de classificação binária independentes, sendo uma para cada fator de personalidade CGF. Para cada um destes fatores, definiu-se uma instância positiva como sendo aquela para a qual o indivíduo obteve pontuação acima da média do córpus de acordo com os resultados do seu inventário de personalidade, ou uma instância negativa em caso contrário. A tarefa de classificação de gênero consiste de um problema de classificação binária (feminino / masculino) único.

3.2. Conjunto de dados

O experimento utiliza um córpus de publicações do Facebook brasileiro de 1020 usuários, rotulado com informações de personalidade do modelo CGF e idade. As informações de personalidade foram obtidas por meio de inventários do tipo IGFP-5 [de Andrade 2008] respondidos pelos próprios usuários, que também cederam suas publicações Facebook para a composição do córpus. As informações de idade (quando disponibilizadas pelos usuários) foram extraídas automaticamente por meio da API Facebook.

A distribuição de classes para as duas tarefas é sumarizada na Tabela 1.

Tabela 1. Distribuição de classes para as tarefas de reconhecimento de personalidade (parte superior) e gênero autoral (inferior).

	Pos.	Neg.
Extroversão	514	506
Agradabilidade	482	538
Conscienciosidade	513	507
Neuroticismo	526	494
Abertura	487	533
<hr/>		
Gênero	Fem 578	Masc. 441

4. Avaliação

O experimento foi realizado utilizando-se regressão logística estratificada com validação cruzada de 10 partições. Além dos dois modelos avaliados - LIWC-En e LIWC-Pt - um *baseline* de classe majoritária foi acrescentado para fins ilustrativos.

4.1. Reconhecimento de traços de personalidade

A tabela 2 apresenta os resultado da tarefa de classificação binária de traços de personalidade utilizando os três modelos propostos.

Os resultados da Tabela 2 podem ser interpretados como uma boa notícia para a comunidade de PLN do português. Os resultados de ambos os modelos LIWC - baseados

Tabela 2. Média F1 ponderada para classificação binária de personalidade.

Tarefa	Baseline	LIWC.En	LIWC.Pt
Extroversão	0.36	0.62	0.61
Agradabilidade	0.38	0.59	0.58
Conscienciosidade	0.36	0.57	0.59
Neuroticismo	0.35	0.57	0.53
Abertura	0.36	0.54	0.56
Média	0.36	0.58	0.57

em textos na versão original em português ou na versão traduzida para o inglês, respectivamente - são praticamente idênticos. Não houve perda significativa de acurácia no uso de textos traduzidos (muito embora a tradução tenha sido feita de forma automática e sem pós-edição), e também não houve ganho significativo em usar o conjunto de características mais completo oferecido pela versão em inglês do dicionário LIWC.

Este resultado pode ser explicado, ao menos em parte, pela observação de que o reconhecimento de traços de personalidade do modelo CGF é um problema de natureza essencialmente lexical, e que este tipo de conhecimento tende a ser mais preservado na tradução para o inglês do que, por exemplo, a estrutura sintática das sentenças. Assim, o método de tradução automática adotado não chega a ser prejudicial ao desempenho da tarefa.

4.2. Classificação de gênero autoral

A tabela 3 apresenta os resultado da tarefa de classificação de gênero para os modelos discutidos.

Tabela 3. Média F1 ponderada para classificação binária de gênero.

Tarefa	Baseline	LIWC.En	LIWC.Pt
Feminino	0.72	0.45	0.49
Masculino	0.00	0.41	0.51
Média	0.36	0.43	0.50

Os resultados da Tabela 3 sugerem um cenário diferente do anterior. Neste caso, houve uma vantagem considerável no uso dos textos e dicionário em português sobre a versão traduzida para o inglês. No entanto, isso não necessariamente deve ser interpretado como uma vantagem do dicionário LIWC português sobre a versão inglesa. Uma possível explicação para este resultado seria a de que a tradução dos textos para o inglês suprime boa parte dos marcadores de gênero presentes no texto em português, o que pode ter favorecido o uso do modelo no idioma original.

Finalmente, observa-se ainda que o presente resultado não é consistente com os resultados relatados em [Isbister et al. 2017], que apontam uma vantagem da classificação de gênero autoral em inglês sobre outros idiomas. Em especial, o presente experimento coloca em questão se o efeito observado em [Isbister et al. 2017] seria realmente devido à maior cobertura do dicionário LIWC inglês, ou talvez decorrente de características específicas dos idiomas analisados.

5. Discussão

Este artigo apresentou uma avaliação do uso do dicionário LIWC português aplicado ao seu propósito possivelmente mais fundamental - a tarefa de reconhecimento de traços de personalidade a partir de texto - e também de uma tarefa adicional de identificação de gênero autoral. No experimento realizado, modelos baseados em características do português foram comparados com modelos baseados em um conjunto estendido de características para o inglês, e utilizando para isso versões traduzidas dos textos de validação.

Mesmo tendo menor cobertura do que a versão mais recente em inglês, o uso do dicionário português permitiu desempenhar a tarefa de reconhecimento de personalidade com resultados semelhantes aos do método baseado em tradução, o que reforça sua aplicabilidade em tarefas de PLN para o português. Estes resultados podem assim ser vistos como um complemento às evidências já apresentadas em [Balage Filho et al. 2013] no contexto da tarefa de análise de sentimentos para o português e, adicionalmente, sugerem que o método de tradução pode facilitar o desenvolvimento de aplicações de natureza lexical para o português quando os recursos necessários só se encontram disponíveis para outro idioma (tipicamente, o inglês).

Como trabalhos futuros, pretende-se investigar o uso de conhecimento psico-linguístico em outros problemas de inferência de informações demográficas, e também de atribuição autoral a partir de textos originais e traduzidos.

Agradecimentos

Esta pesquisa contou com suporte FAPESP nro. 2017/20968-0 e 2016/14223-0, e da Universidade de São Paulo.

Referências

- Allport, F. H. and Allport, G. W. (1921). Personality traits: Their classification and measurement. *Journal of Abnormal And Social Psychology*, 16:6–40.
- Álvarez-Carmona, M., López-Monroy, A., y Gómez, M. M., Villaseñor-Pineda, L., and Escalante, H. (2015). INAOE's participation at PAN'15: Author Profiling task. In *CLEF 2015*.
- Balage Filho, P. P., Aluísio, S. M., and Pardo, T. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology - STIL*, pages 215–219, Fortaleza, Brazil.
- Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., and Nissim, M. (2017). N-GrAM: New groningen author-profiling model. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33(4):497–505.
- de Andrade, J. M. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. PhD thesis, Universidade de Brasília.
- Goldberg, L. R. (1990). An alternative description of personality: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.
- González-Gallardo, C. and et. al. (2015). Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams. In *CLEF 2015*.
- Iacobelli, F., Gill, A. J., Nowson, S., and Oberlander, J. (2011). Large scale personality classification of bloggers. In D'Mello, S. K., Graesser, A. C., Schuller, B., and Martin, J.-C., editors, *ACII (2)*, volume 6975 of *Lecture Notes in Computer Science*, pages 568–577, Memphis, TN, USA. Springer.
- Isbister, T., Kaati, L., and Cohen, K. (2017). Gender classification with data independent features in multiple languages. In *European Intelligence and Security Informatics Conference (EISIC-2017)*, pages 54–60, Athens, Greece. IEEE Computer Society.
- John, O. P., Donahue, E., and Kentle, R. (1991). The Big Five inventory - versions 4a and 54. Technical report, Inst. Personality Social Research, University of California, Berkeley, CA, USA.
- Liu, F., Perez, J., and Nowson, S. (2017). A language-independent and compositional model for personality trait recognition from short texts. In *15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–764, Valencia, Spain. Association for Computational Linguistics.
- Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.
- Martinc, M., Skrjanec, I., Zupan, K., and Pollak, S. (2017). PAN 2017: Author profiling - gender and language variety prediction. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.

- Myers, I. B. and Myers, P. (2010). *Gifts differing: Understanding personality type*. Nicholas Brealey Publishing.
- Nowson, S. and Oberlander, J. (2007). Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, Colorado, USA.
- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.
- Rammstedt, B. and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.
- Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF 2015 Evaluation Labs and Workshop*, Toulouse, France. CEUR-WS.org.
- Rangel, F. M., Rosso, P., Potthast, M., and Stein, B. (2017). Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.
- Reddy, T. R., Vardhan, B. V., and Reddy, P. V. (2017). N-Gram approach for gender prediction. In *Advance Computing Conference (IACC)*, pages 860–865.
- Sierra, S., y Gómez, M. M., Solorio, T., and González, F. A. (2017). Convolutional neural networks for author profiling. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.
- Sulea, O.-M. and Dichiu, D. (2015). Automatic Profiling of Twitter Users Based on Their Tweets. In *CLEF 2015*.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Verhoeven, B., Daelemans, W., and Plank, B. (2016). TwiSty: a multilingual Twitter Stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC-2016)*, pages 1632–1637, Portoroz, Slovenia. ELRA.

Avaliação Extrínseca de Analisadores de Dependência através da Extração de Informação Aberta

Maurício do V. D. Wanderley¹, Daniela Barreiro Claro¹,
Marlo Souza¹, Leandro S. de Oliveira¹

¹FORMAS - Formalismos e Aplicações Semânticas

Departamento de Ciência da Computação – LaSiD – IME – Universidade Federal da Bahia
Av. Adhemar de Barros, s/n, Campus de Ondina, Salvador - Bahia - Brasil

mauricio.wanderley@ufba.br, dclaro@ufba.br, msouza1@ufba.br, leo.053993@gmail.com

Resumo. *Analisadores de Dependência (AD) são sistemas computacionais que analisam as relações gramaticais existentes entre as unidades lexicais em uma sentença. Enquanto nos últimos anos a avaliação de ADs recebeu grande atenção da literatura, tais trabalhos restringiram-se a métodos com avaliação intrínseca. Entretanto, uma avaliação extrínseca de ADs permite um panorama mais flexível para comparar sistemas baseados em diferentes escolhas e arcabouços linguísticos. Por esse motivo, este trabalho propõe uma avaliação extrínseca de ADs, aplicando-os em um sistema de Extração de Informação Aberta (EIA) do estado da arte para a língua portuguesa. Os experimentos demonstraram relevantes diferenças entre as avaliações intrínseca e extrínseca e, portanto, a importância de se avaliar diferentes ADs em um contexto de aplicação.*

1. Introdução

A Análise de Dependência é uma tarefa de análise sintática na qual a estrutura da sentença é descrita não pela forma de suas componentes (ou constituintes) mas pelas relações gramaticais/funcionais entre as palavras da mesma.

Analisadores automáticos da estrutura de dependência de uma sentença, comumente chamados de Analisadores de Dependência (AD) ou *dependency parsers* em inglês, ganharam bastante proeminência na literatura nos últimos anos [Dozat et al. 2017, Oepen et al. 2017, Zeman et al. 2018, Fares et al. 2018, Gamallo and Garcia 2018], dado o surgimento de recursos e métodos robustos para desenvolvimento desses sistemas e suas potenciais aplicações a outras tarefas de Processamento de Linguagem Natural (PLN). Enquanto a tecnologia para a construção de ADs se desenvolveu enormemente nessas últimas décadas, dando origem a diversos sistemas de análise sintática automatizada como [McDonald et al. 2005, Nivre et al. 2006], surge a necessidade do desenvolvimento de metodologias para a avaliação e comparação do desempenho de tais sistemas.

Sistemas de PLN podem ser avaliados em relação a um certo conjunto de dados previamente anotado e a um conjunto de métricas de qualidade aplicáveis (avaliação intrínseca do sistema) ou através de sua aplicação em um contexto de uso, como a aplicação de um método de *parsing* em um sistema maior (avaliação extrínseca do sistema). Como apontado por Jones [Jones 1994], ambas as formas de avaliação são necessárias para compreender o desempenho de um tal sistema e aplicáveis em circunstâncias distintas.

Atualmente, a forma mais comum para avaliar e comparar ADs é através de avaliações intrínsecas, usando um *treebank* manualmente anotado, por meio de métricas como a *Labeled Attachment Accuracy* (LAS) e *Unlabeled Attachment Accuracy* (UAS). Tais métricas se tornaram populares para avaliação de ADs pelo seu uso em tarefas de avaliação conjunta de *parsing* como a CoNLL [Buchholz and Marsi 2006, Zeman et al. 2017, Zeman et al. 2018]. O LAS apresenta a taxa de acerto dos tokens que apontam corretamente para o nó pai e a relação de dependência rotulados corretamente. O UAS apresenta, somente, a taxa de acerto dos tokens que apontam para o nó pai corretamente [Kübler et al. 2009].

Entretanto, a comparação de sistemas reais construídos para essa tarefa através de avaliação intrínseca envolve algumas dificuldades. Isso se dá pelo fato que este tipo de avaliação se baseia em um conjunto de dados previamente anotado e esses dados são dependentes das teorias e escolhas linguísticas específicas, como métodos de *tokenização*, o conjunto de relações gramaticais considerados (*tagset*), etc. De fato, as dificuldades em se comparar analisadores sintáticos baseados em diferentes arcabouços linguísticos já foi reconhecido pela comunidade científica [Carroll et al. 1998, Tsarfaty et al. 2012].

Este trabalho propõe uma avaliação extrínseca de ADs através de uma tarefa de aplicação, nomeadamente Extração de Informação Aberta (EIA), como primeiramente proposto por Gamallo e Garcia [Gamallo and Garcia 2018]. O desempenho de um sistema do estado-da-arte de EIA para a língua portuguesa - nomeadamente o dptOIE [Oliveira and Claro 2019]¹ - é comparado ao receber como entrada as anotações realizadas por diferentes ADs disponíveis na literatura para a língua portuguesa. Com esse tipo de avaliação, é possível comparar de forma contextualizada o desempenho de analisadores sintáticos baseados em diferentes arcabouços linguísticos e com variadas escolhas de implementação, como estratégias de *tokenização*, classes morfossintáticas consideradas (*POS tagset* em inglês) etc.

Note que a EIA é uma tarefa de PLN que consiste na extração de fatos (ou proposições básicas) a partir de bases de dados textuais grandes e heterogêneas [Banko et al. 2007]. Esses sistemas normalmente extraem fatos na forma de triplas $t = (arg1, rel, arg2)$, em que *rel* descreve uma relação semântica entre as entidades *arg1* e *arg2*. Esta tarefa está intrinsecamente associada à análise sintática da sentença, uma vez que consistem na identificação de certos aspectos estruturais que conotam relações semânticas.

Este trabalho está estruturado da seguinte forma: a seção 2 descreve os trabalhos relacionados. A seção 3 apresenta os principais conceitos de Extração da Informação Aberta como mecanismo da avaliação extrínseca da AD. A seção 4 descreve os experimentos realizados e apresenta os resultados obtidos. Finalmente a seção 5 discute as conclusões e descreve os trabalhos futuros.

2. Trabalhos Relacionados

A avaliação de *parsers* é uma preocupação antiga na área de PLN. Para a comparação de analisadores de constituição na língua inglesa, ou seja, analisadores sintáticos que identificam os sintagmas que constituem uma sentença, diversas métricas foram propostas

¹Disponível em <http://formas.ufba.br/dclaro/tools.html#dptoe>

[Carroll et al. 1998] e tornaram-se padrão para avaliar *parsers* baseados no Penn Treebank [Marcus et al. 1993].

A avaliação intrínseca de Analisadores de Dependência, por outro lado, tem como marco a realização da tarefa de avaliação conjunta de analisadores de dependência multilíngue na *SIGNLL Conference on Computational Natural Language Learning - CoNLL* em 2006 [Buchholz and Marsi 2006]. Em sucessivas edições desde 2006 a 2018, a tarefa de avaliação de ADs na CoNLL levou ao estabelecimento de métricas-padrão para avaliação de ADs baseados no padrão *Universal Dependencies* (UD) [McDonald et al. 2013]. A métrica principal adotada é o LAS, sendo esta a principal métrica da área, e tendo suas variações como UAS e *labeled accuracy* (LA), *Content-word Labeled Attachement* (CLA), entre outras[Zeman et al. 2017, Zeman et al. 2018].

A submissão dos diversos analisadores à tarefa permitiu, já em 2006, uma avaliação aprofundada das árvores sintáticas geradas por sistemas, extrapolando limitações de métricas convencionais. O trabalho de [McDonald and Nivre 2007] concentra-se numa análise refinada dos erros gerados por dois ADs com desempenho geral semelhante, representando o estado da arte à época: o *MaltParser*[Nivre et al. 2006], baseado em transições, e o *MSTParser* [McDonald et al. 2005], baseado em grafos. Nesse trabalho, os autores caracterizaram os erros da análise de dependência em relação ao comprimento da sentença, distância entre arcos, presença de não-projetividade e rotulação de POS e tipos de relações de dependência.

A estimativa da utilidade relativa provida por analisadores de dependência é insuficientemente medida por suas métricas padrão (intrínsecas) para sua aplicação em alguns sistemas de PLN. Mais ainda, a avaliação intrínseca de ADs é inadequada para comparação de sistemas baseados em diferentes arcabouços linguísticos. De fato, a avaliação extrínseca de *parsers* já havia sido proposta anteriormente, como quando da realização da tarefa de avaliação conjunta *Extrinsic Parser Evaluation* (EPE), realizada em edições dos anos 2017 [Oepen et al. 2017] e 2018 [Fares et al. 2018], sendo a última uma trilha da CoNLL Shared Task. Nelas são avaliadas as saídas em diferentes representações de diversos *parsers* para aplicações de: extração de eventos biológicos, análise de opiniões e resolução de negação. Em 2018, a saída resultante foi restrita ao padrão UD.

Ao compararem os resultados das avaliações extrínsecas entre as edições 2017 e 2018 da EPE, os organizadores do evento creditam o desempenho mais baixo das aplicações alcançado em 2018 devido à adoção pelos *parsers* do padrão UD e treino limitado a dados em língua inglesa. Essa limitação não é encontrada nesta abordagem, uma vez que a ferramenta de EIA escolhida se vale do formato UD para aplicar suas regras de extração de informação.

O trabalho de Gamallo e Garcia [Gamallo and Garcia 2018], por outro lado, propõe a utilização de sistemas de EIA para realizar avaliação extrínseca de ADs e utiliza o módulo de EIA presente na suíte multilíngue LinguaKit [Gamallo and Garcia 2017] para avaliar ADs distintos, de modo semelhante ao proposto neste trabalho. Os autores utilizam para a sua avaliação um conjunto de 103 sentenças provenientes de um *corpus* de domínio específico.

O presente trabalho se distingue em dois aspectos: (i) na escolha dos ADs avalia-

dos e (ii) na escolha dos dados para a avaliação. Quanto à escolha dos ADs, ao escolher os sistemas com melhor desempenho na avaliação da CoNLL, é possível contrastar os resultados dos diferentes métodos de avaliação e investigar a complementaridade dos métodos de avaliação de ADs. Quanto à escolha dos dados utilizados, o presente trabalho utilizou sentenças aleatórias provenientes de textos de estilos linguísticos heterogêneos e de diferentes domínios do conhecimento. Tal variedade é uma importante característica na utilização de EIA, uma vez que a tarefa foi projetada para lidar com essa diversidade de domínios e estilos.

3. Extração de Informações Aberta e Análise de Dependência

Extração de Informação (EI) é a área que estuda métodos para obter informação estruturada a partir de dados não estruturados, como textos. Os métodos tradicionais dessa área se baseiam na identificação de padrões textuais/linguísticos para identificação de instâncias de um conjunto fechado e conhecido *a priori* de relações semânticas relevantes para um domínio de aplicação. Por exemplo, sistemas de EI podem extrair a informação nasceuEm(Aristóteles, a Estagira) na sentença “*Aristóteles nasceu na Estagira*”.

Recentemente, o paradigma de Extração de Informação Aberta (EIA) foi proposto visando generalizar o problema tratado pela Extração de Informação de forma a não restringir a identificação a um conjunto pré-definido de relações. A EIA extrai toda e qualquer relação semântica expressa em um texto, resolvendo assim os problemas de escalabilidade e adaptabilidade dos sistemas a diferentes domínios [Banko et al. 2007, Xavier et al. 2015, Glauber and Claro 2018].

Analisadores de dependência são ferramentas muito utilizadas em sistemas de EIA, pois permitem explorar e manipular a estrutura dessas sentenças de forma explícita para identificar as relações semânticas, diferentemente dos métodos baseados em análise superficial como padrões morfossintáticos. De fato, os sistemas de EIA que utilizam AD normalmente tem apresentado melhores resultados [Del Corro and Gemulla 2013, Oliveira et al. 2017, Oliveira and Claro 2019].

Sistemas de EIA podem ser divididos em duas categorias amplas: os sistemas baseados em regras pré-definidas e baseados em aprendizagem automática [Glauber and Claro 2018]. Além disso, essas categorias podem ser separadas em dois subtipos: os sistemas que realizam análise rasa da sentença, geralmente através de POS *taggers* e/ou *chunkers*, e os que utilizam uma análise mais profunda, através da análise de dependência. Os sistemas que fazem análise rasa normalmente obtém precisão alta, mas possuem baixa cobertura. As abordagens que utilizam análise de dependência geralmente tem um custo maior de processamento, mas apresentam uma melhora tanto na precisão quanto na cobertura. Assim, é possível verificar que o desempenho de um AD exerce grande influência nos sistemas de EIA que os utilizam [Oliveira et al. 2017, Sena and Claro 2019].

4. Experimentos e Resultados

Os experimentos para avaliação extrínseca de Analisadores de Dependência para a língua Portuguesa deste trabalho foram realizados através da tarefa de Extração de Informação Aberta. Os materiais e o conjunto de dados utilizado são apresentados, assim como o

projeto experimental proposto. Posteriormente, discute-se os resultados obtidos a partir dos experimentos realizados.

4.1. Materiais

Neste trabalho foram selecionados 4 Analisadores de Dependência de equipes participantes da CoNLL Shared Task 2017 que alcançaram os melhores resultados de LAS para o Português Brasileiro. Para tal, dois critérios de inclusão do sistema foram adotados: (i) o código-fonte deveria estar disponibilizado e (ii) possibilidade de replicar o treino do modelo, utilizando os mesmos dados da tarefa. Com base nesses critério, foram selecionados os sistemas: *Stanford*² [Dozat et al. 2017], *Orange – Deskiñ*³ [Heinecke and Asadullah 2017], *MQuni*⁴ [Nguyen et al. 2017] e *LyS-FASTPARSE*⁵ [Vilares and Gómez-Rodríguez 2017], e o *Stanford CoreNLP* [Chen and Manning 2014], presente atualmente no *dptOIE*. A Tabela 1 apresenta o desempenho alcançado pelos ADs em sua avaliação intrínseca.

Tabela 1. Características dos Analisadores de Dependência participantes da CoNLL Shared Task 2017

Equipe	LAS	Abordagem	Posição
Stanford	91.36	Grafo	1
MQuni	87.91	Grafo	5
Orange – Deskiñ	87.07	Transição	8
LyS-FASTPARSE	86.74	Transição	9

Os ADs selecionados foram posteriormente incorporados ao sistema de EIA *dptOIE* [Oliveira and Claro 2019] para a realização da avaliação. O *dptOIE* é um sistema do estado da arte para EIA na língua portuguesa e recebe como entrada textos sintaticamente anotados seguindo o padrão *CoNLL-U* para árvores de dependência, o que permite ser facilmente adaptado para a saída de diferentes ADs seguindo esse padrão. Os ADs participantes da CoNLL Shared Task 2018 não foram considerados pois os *treebanks* avaliados nesta edição são incompatíveis com o *dptOIE*. Além disso, ADs que obtiveram alto LAS para a língua portuguesa basearam-se em [Dozat et al. 2017], sistema com maior LAS em 2017, o que traria menor diversidade de abordagens.

4.2. Conjunto de dados

O conjunto de dados de treino para os analisadores de dependência foi o *treebank* do português brasileiro *UD_Portuguese-BR*⁶. Este *treebank* fornece 11998 sentenças com suas árvores de dependências anotadas, provindas de textos extraídos de artigos de jornais, *blogs* e *reviews* de consumidores. O conjunto de dados de entrada para o *dptOIE* é formado por um *corpus* de 100 sentenças aleatoriamente selecionadas do Corpus de Extratos de Textos Eletrônicos NILC/Folha de S. Paulo (*CETENFolha*⁷) e da *Wikipedia*.

²<https://github.com/CoNLL-UD-2017/Stanford>

³<https://github.com/CoNLL-UD-2017/Orange-Deskin>

⁴<https://github.com/datquocnguyen/jPTDP>

⁵<https://github.com/CoNLL-UD-2017/LyS-FASTPARSE>

⁶https://github.com/UniversalDependencies/UD_Portuguese-GSD

⁷<http://www.linguateca.pt/cetenfolha/>

4.3. Métricas

Os fatos extraídos pelo *dptOIE* foram avaliados por 3 especialistas com domínio na área de EIA, sendo consideradas corretas as extrações julgadas coerentes por pelo menos 2 dos votos. A coerência aqui é entendida como fatos cujos argumentos conectados preservam a informação da sentença que a originou, como é mostrado na Tabela 2. O grau de concordância entre os especialistas foi determinada pela medida *kappa* [Carletta 1996].

Tabela 2. Avaliação de coerência de fatos extraídos de uma sentença pelo *dptOIE*

<i>Horntown é uma cidade localizada no estado norte-americano de Oklahoma , no Condado de Hughes.</i>			
Argumento 1	Relação	Argumento 2	Coerente
Horntown	é uma cidade localizada em	o Condado de Hughes	sim
Horntown	é uma cidade localizada em	o estado norte - americano de Oklahoma	sim
Horntown	é	uma cidade localizada	não

As métricas adotadas para avaliação do desempenho obtido pelo sistema de EIA são aplicadas às extrações obtidas ao se integrar cada um dos ADs ao *dptOIE*. São elas:

- quantidade de fatos extraídos;
- *yield* - a quantidade de fatos coerentes extraídos;
- quantidade de fatos distintos;
- quantidade de sentenças sem fato algum extraído;
- a precisão - medida pela fração de triplas coerentes entre todas triplas recuperadas;
- a cobertura - a fração de triplas coerentes dentre o conjunto de todas triplas coerentes;
- a precisão x *yield*
- a área sob a curva da precisão x *yield*

4.4. Descrição dos Experimentos

Os Analisadores de Dependência foram avaliados a partir das extrações obtidas pelo *dptOIE*, ao fornecerem ao sistema as árvores de dependência para o mesmo conjunto de 100 sentenças. A aplicação de regras gramaticais sob as árvores sintáticas geradas pelos ADs permite a segmentação das sentenças em argumentos mediados por uma relação, e o *dptOIE* o faz, de modo geral, através da identificação do sujeito, relações e seus dependentes, como objetos e complementos. Além disso, o *dptOIE* tem tratamento específico para conjunções coordenadas, orações subordinadas e tratamento de aposto.

4.5. Resultados

Foram obtidos 1211 fatos extraídos de 100 sentenças pelo *dptOIE*, havendo triplas em comum entre os ADs adotados. Destes, 516 foram julgados coerentes, obtendo a medida de concordância *kappa* de 0.68. Segmentando as extrações por AD, a Figura 1 demonstra que o *Orange - Deskiñ* propiciou um maior *yield*, o *MQuni* a maior quantidade de extrações (incoerentes \cup coerentes), e o *Lys-FASTPARSE* a menor quantidade de fatos coerentes.

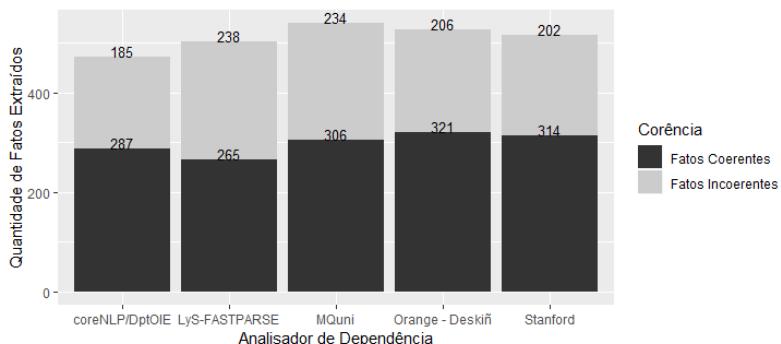


Figura 1. Quantitativo de fatos extraídos pelos analisadores de dependência avaliados

Os valores absolutos de *yield* são importantes para um sistema de EIA, entretanto as métricas de precisão e cobertura permitem avaliar o desempenho em relação aos erros obtidos e ao universo de fatos coerentes. Como mostra a Tabela 3, o AD *Orange - Deskiñ* propicia a maior precisão (0.609), com resultado quase idêntico ao de *Stanford* e de *coreNLP/dptOIE* (0.608 ambos). Entretanto, ao observar a cobertura alcançada com estes 3 ADs, o resultado do *Orange - Deskiñ* é superior, enquanto o *coreNLP/dptOIE* apresenta a segunda menor cobertura. A relação entre precisão e cobertura é dada pela medida-F, a qual confirma o melhor desempenho do *dptOIE* é resultado do AD *Orange - Deskiñ*, seguido do *Stanford*. Utilizando estes dois ADs, o sistema de EIA obtém mais fatos coerentes com um menor custo trazido por fatos incoerentes.

Tabela 3. Desempenho alcançado pelo dptOIE de acordo com o AD adotado

AD	Extrações	Yield	Precisao	Cobertura	Medida-F	AUC-PY
coreNLP/dptOIE	472	287	0.608	0.556	0.580	175.55
Stanford	516	314	0.608	0.608	0.608	192.41
Orange - Deskiñ	527	321	0.609	0.622	0.615	207.38
MQuni	540	306	0.566	0.593	0.579	186.40
LyS-FASTPARSE	503	265	0.526	0.513	0.520	131.12

A área sob a curva da relação entre *yield* e precisão foi calculada a partir do gráfico da Figura 2, e corrobora o desempenho superior ao se adotar os ADs *Stanford* e *Orange - Deskiñ*, mostrando que realizam mais extrações coerentes mantendo melhor precisão.

O conjunto de fatos obtidos é formado em grande parte por extrações comuns a todos os ADs. Entretanto, eles podem identificar árvores de dependência que tanto não levem a extração alguma ou a extrações coerentes distintas dos demais, como é mostrado na Tabela 4.

Os fatos distintos são importantes por demonstrarem relações não identificadas por nenhum outro AD, e neste sentido, o *Orange - Deskiñ* apresenta um previsível melhor resultado, seguido do *coreNLP/dptOIE*, o que não seria demonstrado por sua medida-F. Deve-se notar que mesmo com a superioridade em todas as métricas, o *Orange - Deskiñ* não permitiu extração para a sentença “*Rebeldes islâmicos fugiram em 1992 de o Tadjiquistão para o Afeganistão , após derrota para forças controladas por ex-comunistas.*”, enquanto todos os outros ADs proveram fatos coerentes.

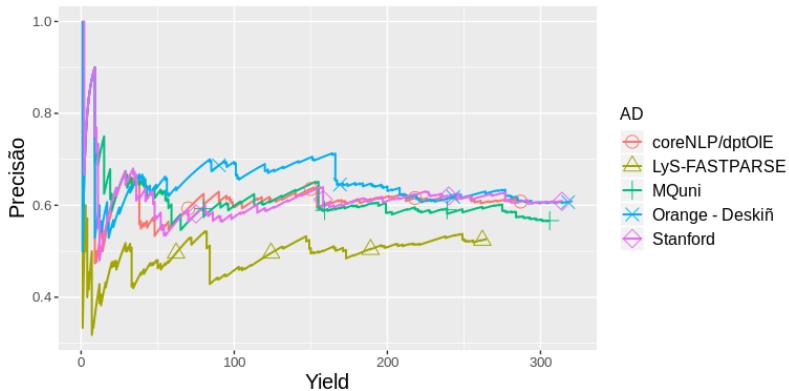


Figura 2. Precisão x Yield

Tabela 4. Sentenças sem extração e fatos distintos por AD

AD	Sem extração	Distintos
coreNLP/dptOIE	11	40
Stanford	11	35
Orange - Deskiñ	11	41
MQuni	9	34
Lys-FASTPARSE	9	31

5. Conclusões e Trabalhos Futuros

A avaliação intrínseca de analisadores de dependência serve como uma boa medida da evolução destes sistemas: quanto mais as medidas de LAS e UAS se aproximam de seu valor máximo, melhores foram os aperfeiçoamentos adotados ao seu método. Mesmo assim, estas métricas pouco dizem sobre o impacto do erro de uma rotulação ou relação de dependência em métodos de PLN baseados em regras gramaticais, como o *dptOIE*. A avaliação dos fatos extraídos por este sistema requer uma interpretação semântica, de modo que a correta identificação de algumas estruturas sintáticas ganham maior relevância em detrimento de outras em determinados contextos.

O melhor desempenho observado pelo *Orange - Deskiñ* corrobora nossa hipótese de que as métricas de LAS e UAS são importantes mas insuficientes para estimar a contribuição fornecida por um AD específico. Mesmo que o constante aumento do LAS nos últimos anos traga melhorias aos métodos de PLN, a adoção indiscriminada de um AD considerando unicamente esta métrica pode não trazer a contribuição esperada. Por outro lado, vimos que mesmo ADs que tragam menor precisão e cobertura poderão gerar fatos não identificados por outros.

Como trabalho futuro, pretende-se explorar outras métricas de avaliação de EIA para avaliar o desempenho de ADs. Além disso, uma análise refinada da estrutura da sentença relacionada aos erros deve ser realizada para melhor compreender os erros de cada AD.

Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.
- Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning, CoNLL-X '06*, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22:249–254.
- Carroll, J., Briscoe, T., and Sanfilippo, A. (1998). Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 447–454. Granada.
- Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.
- Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.
- Dozat, T., Qi, P., and Manning, C. D. (2017). Stanford’s graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.
- Fares, M., Oepen, S., Øvrelid, L., Björne, J., and Johansson, R. (2018). The 2018 shared task on extrinsic parser evaluation: On the downstream utility of english universal dependency parsers. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 22–33.
- Gamallo, P. and Garcia, M. (2017). Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática*, 9(1):19–28.
- Gamallo, P. and Garcia, M. (2018). Task-oriented evaluation of dependency parsing with open information extraction. In *International Conference on Computational Processing of the Portuguese Language*, pages 77–82. Springer.
- Glauber, R. and Claro, D. B. (2018). A systematic mapping study on open information extraction. *Expert Systems with Applications*, 112:372–387.
- Heinecke, J. and Asadullah, M. (2017). Multi-model and crosslingual dependency analysis. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 111–118.
- Jones, K. S. (1994). Towards better nlp system evaluation. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Kübler, S., McDonald, R., and Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127.

- Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- McDonald, R. and Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131.
- McDonald, R., Nivre, J., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.
- McDonald, R. T., Pereira, F., Ribarov, K., and Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *HLT/EMNLP*.
- Nguyen, D. Q., Dras, M., and Johnson, M. (2017). A novel neural network model for joint pos tagging and graph-based dependency parsing. *arXiv preprint arXiv:1705.05952*.
- Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.
- Oepen, S., Ovrelid, L., et al. (2017). The 2017 shared task on extrinsic parser evaluation towards a reusable community infrastructure. *Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation*, pages 1–16.
- Oliveira, L. and Claro, D. B. (2019). dptoie: A portuguese open information extraction system based on dependency analysis. *Linguamatica, (Under review). Obtained directly from the authors*.
- Oliveira, L., Glauber, R., and Claro, D. B. (2017). Dependentie: An open information extraction system on portuguese by a dependence analysis. In *ENIAC - 2017 XIV Encontro Nacional de Inteligência Artificial e Computacional*.
- Sena, C. F. L. and Claro, D. B. (2019). Inferportoie: A portuguese open information extraction system with inferences. *Natural Language Engineering*, 25(2):287–306.
- Tsarfaty, R., Nivre, J., and Andersson, E. (2012). Cross-framework evaluation for statistical parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 44–54. Association for Computational Linguistics.
- Vilares, D. and Gómez-Rodríguez, C. (2017). A non-projective greedy dependency parser with bidirectional lstms. *arXiv preprint arXiv:1707.03228*.
- Xavier, C. C., Lima, V. L. S., and Souza, M. (2015). Open information extraction based on lexical semantics. *Journal of the Brazilian Computer Society*, 21(4):1–14.
- Zeman, D. et al. (2017). Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.
- Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

Using linguistic cues to detect fake news on the Brazilian Portuguese parallel corpus Fake.BR

Emerson Yoshiaki Okano¹, Evandro Eduardo Seron Ruiz¹

¹Departamento de Computação e Matemática, FFCLRP, Universidade de São Paulo
Av. Bandeirantes, 3900, Monte Alegre. CEP: 14040-901, Ribeirão Preto - SP
Brazil

{okano700, evandro}@usp.br

Abstract. *Fake news has become a significant concern because of their ease propagation through social networks. The use of natural language processing (NLP) tasks to detect fake news has soared lately as human beings usually fail on this task. The purpose of the current work is to detect fake news written in Portuguese using psycholinguistic cues. We tested 76 linguistic characteristics on the Fake.Br parallel corpus. Results show an accuracy of 75% using three distinctive machine learning methods.*

Resumo. *As Fake news se tornaram uma grande preocupação devido a fácil propagação através de redes sociais. O uso de processamento de linguagem natural (PLN) para detectar fake news tem aumentado ultimamente, já que os seres humanos normalmente falham nessa tarefa. O objetivo deste trabalho é detectar fake news escritas em Português usando características psicolinguísticas. Nós testamos 76 características psicolinguísticas no corpus paralelo Fake.Br. Os resultados apresentados foram uma acurácia de 75% utilizando três métodos de aprendizado de máquina distintos.*

1. Introduction

The web constitutes a vast source of textual information, but not every piece of information presented therein is reliable. Some users might use the web to spread false information to manipulate and deceive other users [Ott et al. 2011]. One kind of false information is what is called fake news, which is everywhere nowadays. The term fake news is usually referred to false news published with the aim of achieving some advantage by spreading false information. Fake news has increased as the ease of access to the web and also due to the growing use of web-based social networks.

The term ‘fake news’ became popular with key political events including the Brexit vote in the UK and the US presidential election in 2016. In Brazil, it became well known just after the 2018 presidential elections.

Recently, there have been many national and international efforts to reduce the effects of false or misleading news or even reveal the hidden truth in fake news creating web pages known as fact pages checking. Two well known Brazilian initiatives are: Me engana que eu posto¹, and Agência Lupa². Following this same line, companies such as

¹<https://veja.abril.com.br/blog/me-engana-que-eu-posto/>

²<https://piaui.folha.uol.com.br/lupa/>

Google, Facebook, and Bing have joined forces to create The Trust Project³, which consists of an international consortium of news organizations that are collaborating intending to create journalism standards to produce a more reliable press.

Some authors used psycholinguistic features trying to classify deceptive texts [Hauch et al. 2015, Ott et al. 2011]. Myle Ott [Ott et al. 2011] achieved an accuracy of 89.8% using Linguistic Inquiry and Word Count (LIWC)+Bigrams and SVM as classifier. Valerie Hauch [Hauch et al. 2015] presented a meta-analysis of linguistic cues to deception, showing which cues are better to detect deception in texts.

In our paper, we followed the cues showed through a meta-analysis conducted by Hauch [Hauch et al. 2015] to detect fake news. As a novelty, we used these cues as features for some classic machine learning classification algorithms. We also used the effect size measure to show which cue has more influence in fake and true news characterization.

We organized this paper as follows: Section 2 presents state-of-the-art detection of opinion spam. Section 3 shows the methodology used by Hauch [Hauch et al. 2015] and in this work. Section 4 presents the dataset used in this work. Section 5 we show the results obtained with our methodology. Section 6 presents the conclusion of this paper.

2. Related work

Recent work in linguists has explored textual indicators to be implemented by computer systems to discriminate between truths and lies.

Myle Ott and colleagues [Ott et al. 2011] pioneered in the study of psycholinguistic cues using NLP. They used three different automated approaches to detect deceptive opinion spam, which are: a) Frequency distribution of the part-of-speech (POS) tags (word categories) in a text which are often dependent on the genre of the text; b) Psycholinguistic deception detection, whereas the Linguistic Inquiry and Word Count (LIWC) [Tausczik and Pennebaker 2010] is a powerful tool for textual analysis and it has been used to analyze deception [Hancock et al. 2007], and, finally; c) Text categorization approaches, like bag of words and n-gram. They consider n-grams as effective features for deception detection, allowing to model both content and context of the message. Ott and co-workers also achieved their best results using bigrams + LIWC features with SVM as classifier [Ott et al. 2011].

According to [Rubin et al. 2015a] there are three types of Fake news: a) Serious Fabrications: that are the fraudulent news, fake contents with the purpose of manipulating and deceiving the public; b) Large-Scale Hoaxes: rumors that are not confirmed that attempts to deceive audiences masquerade as news; c) Humorous Fakes: that are the one humorous content like News satire sites and news parody. According to Rubin, hybrid methods combining linguist cues and machine learning are very promising approaches. In an earlier article [Rubin et al. 2015b], Rubin and co-workers applied a vector space model to cluster the news by discourse feature similarity, achieving 63% accuracy.

Mihalcea and colleagues [Pérez-Rosas et al. 2018] also worked on fake news detection. They introduced two novel datasets for the task of fake news detection and con-

³<https://thetrustproject.org/>

ducted a set of learning experiments to build fake news detectors. They achieved accuracies of up to 76%.

3. Methods

We based our work on a meta-analysis proposed by Hauch and co-workers [Hauch et al. 2015], where the authors investigated 79 linguistic cues compiled from 44 other academic studies. Hauch *et al.* allocated these cues to six research questions:

1. Do liars experience greater cognitive load?
2. Are liars less certain than truth-tellers?
3. (a) Do liars use more negations and negative emotion words?
(b) Do liars use fewer positive emotion words?
(c) Do liars express more or less unspecified emotion words?
4. Do liars distance themselves more from events?
5. Do liars use fewer (sensory and contextual) details?
6. Do liars refer less often to cognitive processes?

In our work, we investigated 76 linguistic cues from the 79 cues studied by [Hauch et al. 2015]. Three of them (Writing errors, Pleasantness and unpleasantness, and redundancy) have not been implemented due to a lack of resources. Some linguistic cues were originally allocated to a miscellaneous (M) category, as Hauch showed they did not fit into any research question. To extract some of the linguistic cues, we mainly used the Brazilian Portuguese LIWC [Balage Filho et al. 2013], which is a translation of the original one [Tausczik and Pennebaker 2010]. LIWC is an English lexicon that classifies words in psycholinguistic and linguistic categories.

For some linguistic cues (e.g. pausality and verb quantity) we used Spacy⁴ [Honnibal and Montani 2017], a free, open-source library for Natural Language Processing in Python. From Spacy we used the *pt_core_news_sm* module which performs part-of-speech tagging in Portuguese texts.

We conducted our experiments in two steps: 1) first, we calculated the effect size measure g of each of the 76 linguistic cues. The top ten g measures were selected and shown for the fake and the real news; then; 2) We applied a machine learning approach to classify fake news from the real ones.

The g proposed by Hedges's [Hedges and I. 1985] is an unbiased estimator of the standardized mean difference. Equation 1 presents g where M_1 is the mean value of the attribute in true news and M_2 is the mean value of the attribute in fake news. Otherwise, s_{pooled}^* , the weighted average of standard deviations for the two groups, is given by Equation 2, where n_1, s_1 and n_2, s_2 are the numbers and standard deviation of true and fake news respectively. Positive g indicates truth and negatives g are indicative of fake news.

$$g = \frac{M_1 - M_2}{s_{pooled}^*} \quad (1)$$

⁴<https://spacy.io/>

$$s_{pooled}^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (2)$$

To interpret the hedges g effect size we used the “Rules of thumb” adopted by Sawilowsky [Sawilowsky 2009], here presented in Table 1.

g	Interpretation
0.01	Very small
0.2	Small
0.5	Medium
0.8	Large
1.2	Very large
2	Huge

Table 1. Rules of Thumb.

Next, after calculating the values for all the 76 cues for the 7,200 news (equally divided into two classes, fake and real), we applied three machine learning algorithms to perform an automatic classification of news, they were: Support Vector Machine, SVM, with a linear kernel; Random Forest, and; Logistic regression. We used the 5-fold cross-validation procedure and computed accuracy, precision, recall, and F-score for each class.

4. Fake.Br corpus

In our work, to process the linguistic cues proposed by [Hauch et al. 2015], we used the Fake.Br corpus [Monteiro et al. 2018], which is the first corpus containing fake news in Brazilian Portuguese. To build this corpus, the authors manually collected and checked 3,600 fake news on the internet and semi-automatically looked for corresponding true news for each fake one. The authors collected all the fake news from an interval of two years (from 01/2016 to 01/2018) from 4 websites: *Diário do Brasil*, *A Folha do Brasil*, *The Journal Brasil* and *Top Five TV*. On the other hand, the authors collected the true news using a web crawler in three major news agencies in Brazil: *G1*, *Folha de São Paulo* and *Estadão*. After the web crawler, the authors used a lexical similarity measure to choose the most similar to the fake news collected. They also performed a manual verification to guarantee that the fake news and true news are subject related.

5. Results

It is worth mentioning that often the real news has longer sentences than their corresponding fake news. In this section, we will present the results obtained using the cues, mentioned in Section 3, in the Fake.Br corpus, using two different approaches: 1) using the full text, and; 2) using the truncated text [Monteiro et al. 2018], wherein each fake-true pair, the longer text is truncated to the size (number of tokens) of the shorter text.

5.1. Full text

First, we performed our experiments using the full text of the corpus. In Table 2 we presented the top 10 positive and negative hedges, g . For this experiment the average absolute effect size was $\bar{g} = 0.31$ and its standard deviation was $s_g = 0.53$.

Analyzing the results obtained in Table 2, we observed that we obtained five cues from the research question 5 showing that, in general, fake news uses more sensory and contextual details than the ones used for true news. These results contradict the results obtained from [Hauch et al. 2015] for the same research question (5). We observed some cues have higher values of g than others (e.g., Type token ration, number of verbs). After analyzing these clues, we realized that they are determined by news containing a large number of tokens when compared to their equivalent fake version. Generally, the number of tokens per news is not a well-balanced feature. The mean number of tokens for fake news is 185.92 while for true news is 1110.40.

RQ	Cue	g	\bar{x}_f	s_f	\bar{x}_t	s_t
1	Type-token ratio	-2.72	0.6541	0.0809	0.4518	0.0673
5	Perceptual processes	-0.36	0.0345	0.0212	0.0284	0.0110
5	Motion	-0.29	0.0567	0.0249	0.0508	0.0138
M	Leisure	-0.26	0.0148	0.0156	0.0116	0.0083
5	Feel	-0.25	0.0116	0.0123	0.0092	0.0050
5	Hear	-0.24	0.0139	0.0130	0.0114	0.0068
M	Auxiliary verbs	-0.23	0.0505	0.0219	0.0462	0.0145
5	Quantifiers	-0.23	0.0360	0.0192	0.0324	0.0118
3(a)	Negative emotion	-0.21	0.0383	0.0236	0.0341	0.0144
3(a)	Sadness	-0.21	0.0092	0.0105	0.0075	0.0050
M	Pausality	0.50	2.5329	0.7777	2.8804	0.6166
5	Prepositions	0.52	0.1777	0.0370	0.1938	0.0238
1	Average sentence length	0.73	109.7863	34.6865	136.4863	38.6173
2	Modal verbs	1.01	0.8456	1.4014	4.2136	4.4796
4	Passive voice verbs	1.48	1.8303	1.9291	10.0644	7.6385
1	Sentence quantity	1.52	10.3014	7.2094	52.3358	38.4219
1	Verb	1.71	28.0431	20.0078	156.0322	104.2411
4	Past tense verbs	1.74	14.3539	10.4727	81.2647	53.2152
1	Length	1.86	185.9167	128.1151	1110.3969	689.9285
4	Present tense verbs	1.91	62.4783	41.6863	378.0736	229.7128

Table 2. Top 10 negative and positives linguistic cues extracted from full text
where: RQ = Research question for the cue it belongs to (where M stands for miscellaneous category); Cue = Linguistic cue itself; g = Hedges effect size; \bar{x}_f = mean value for fake news; s_f = standard deviation for fake news; \bar{x}_t = mean value in true news; s_t = standard deviation in true news.

5.2. Truncated Texts

Using the truncated text, we obtained more reliable results, as shown in Table 3. In Table 4 we described the best cues shown in Table 3 for better understanding. In this experiment we obtained a lower average absolute effect size $\bar{g} = 0.24$ and a lower standard deviation $s_g = 0.17$ than using full text.

In Table 3, we observed that we have seven cues from the overall twenty presented in the same table that belongs to the miscellaneous category, which is a category that contains cues that could not be allocated in any of the research questions. From these

cues, we can notice that social-related words and auxiliary verbs are frequently used in fake news having $g = -0.6$, a medium effect size.

Another interesting cue is the Average sentence length that has a hedges $g = 0.74$ showing that true news has more words per sentence than fake news. This feature may be interpreted as true news are being more elaborate texts than fake news.

RQ	Cue	g	\bar{x}_f	s_f	\bar{x}_t	s_t
4	Pronouns	-0.65	0.1420	0.0374	0.1180	0.0367
M	Social processes	-0.60	0.1756	0.0446	0.1497	0.0420
M	Auxiliary verbs	-0.60	0.0504	0.0219	0.0381	0.0193
4	Impersonal pronouns	-0.55	0.1189	0.0330	0.1009	0.0328
5	Personal pronouns	-0.54	0.0986	0.0279	0.0838	0.0270
2	Tentative	-0.50	0.0705	0.0283	0.0567	0.0268
5	Quantifiers	-0.49	0.0360	0.0192	0.0270	0.0180
M	Articles	-0.48	0.0974	0.0248	0.0854	0.0257
M	Function words	-0.46	0.4271	0.0488	0.4046	0.0496
4	3rd pers. singular	-0.44	0.0836	0.0256	0.0724	0.0250
M	Work	0.14	0.0482	0.0250	0.0519	0.0284
M	Pausality	0.15	2.5595	0.7788	2.6823	0.8972
5	Time	0.17	0.0654	0.0253	0.0701	0.0302
4	Present tense verbs	0.20	61.6869	38.6631	69.4503	40.0187
5	Space	0.21	0.1264	0.0346	0.1339	0.0363
1	Six-letter words	0.24	0.3800	0.0523	0.3923	0.0492
1	Average word length	0.27	4.8338	0.3025	4.9149	0.2965
M	Inclusive	0.27	0.1400	0.0336	0.1494	0.0357
5	Prepositions	0.63	0.1777	0.0371	0.2001	0.0338
1	Average sentence length	0.74	109.7942	34.7963	140.1972	46.4695

Table 3. Top 10 negative and positives linguistic cues extracted from truncated text where: RQ = Research question for the cue it belongs to (where M stands for miscellaneous category); Cue = Linguistic cue itself; g = Hedges effect size; \bar{x}_f = mean value for fake news; s_f = standard deviation for fake news; \bar{x}_t = mean value in true news; s_t = standard deviation in true news.

5.3. Classification task

To verify the performance of the psycholinguistic features in automatic classifying of fake news, we used some classic machine learning algorithms with the psycholinguistic features extracted from the truncated texts.

Following the work of [Monteiro et al. 2018], we used the 5-fold cross-validation computing accuracy, precision, recall, and F-score for each class. In table 5, we show the results obtained using the psycholinguistic features. Observing the classification results presented in Table 5, we achieved good results taking into account that we used only psycholinguistic features.

6. Conclusion

In our paper, we presented the research questions, and the clues previously showed by Hauch and colleagues [Hauch et al. 2015] and, as they did in their work, we used the

RQ	Cue	Operational definition
4	Pronouns	% of pronouns (e.g., eu, nós)
M	Social processes	% of words that express social processes (e.g., falar, amigos)
M	Auxiliary verbs	Number of auxiliary verbs (e.g., ter, ser)
4	Impersonal pronouns	% of impersonal pronouns (e.g., aquele, esse)
5	Personal pronouns	% of personal pronouns (e.g., eu, ele)
2	Tentative	% of tentative words (e.g., talvez, possivelmente)
5	Quantifiers	% of quantifiers words (e.g., alguns, muitos)
M	Articles	% of articles
M	Function words	% of articles, pronouns, verbs
4	3rd pers. singular	% of 3rd pers. singular (e.g., ele, ela)
M	Work	% of words that express job issues (e.g., chefe, carreira)
M	Pausality	Total number of punctuation marks divided by total number of sentences
5	Time	% of temporal words (e.g., tarde, hora)
4	Present tense verbs	Number of present tense verbs
5	Space	% of spatial words (e.g., fora, espaçoso)
1	Six-letter words	% of words that are longer than six letters
1	Average word length	Total number of letters divided by the total number of words
M	Inclusive	% of inclusive words (e.g., dentro, junto)
5	Prepositions	% of prepositions
1	Average sentence length	Total number of words divided by total numbers of sentences

Table 4. Top 10 negative and positives linguistic cues extracted from truncated text where: RQ = Research question for the cue it belongs to (where M stands for miscellaneous category); Cue = Linguistic cue itself; Operational definition = describes the cue itself.

Hedges g effect size to measure the influence of each clue in true and fake news.

To verify the performance of these clues under automatic classification, we applied three machine learning algorithms to classify the news as fake or real. The classification results presented in Table 5 are comparable to [Monteiro et al. 2018] except when they used the bag of words approach. Although [Pérez-Rosas et al. 2018] used two different corpora, they also achieved similar results for fake news detection. Fake news generally uses words commonly adopted at the time the news are created. Given this fact, gazetteer approaches able to discern between fake and true news would have to rely on this time-dependent dictionaries. One can say that the method used in this paper relies on a steady, reliable, and less dependent time dictionary, the LIWC.

7. Acknowledgements

This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process number: 2018/03129-8. We would like to include a special note of thanks to Dr. Tiago Agostinho de Almeida for helpful comments on this project.

Method	Precision		Recall		F-score		
	Accuracy	Fake	True	Fake	True	Fake	True
SVM Linear	0.75	0.76	0.74	0.73	0.77	0.75	0.76
Random forest	0.75	0.75	0.75	0.74	0.75	0.75	0.75
Logistic regression	0.75	0.76	0.74	0.73	0.76	0.74	0.75

Table 5. Results obtained using machine learning in truncated texts.

References

- Balage Filho, P. P., Aluísio, S. M., and Pardo, T. A. S. (2013). An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. *9th Brazilian Symposium in Information and Human Language Technology – STIL*, pages 215–219.
- Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.
- Hauch, V., Blandón-Gitlin, I., Masip, J., and Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, 19(4):307–342.
- Hedges, L. V. and I., O. (1985). *Statistical methods for meta-analysis*. Academic Press, San Diego, CA.
- Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.
- Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, pages 324–334. Springer International Publishing.
- Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1*, pages 309–319. Association for Computational Linguistics.
- Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Rubin, V. L., Chen, Y., and Conroy, N. J. (2015a). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.
- Rubin, V. L., Conroy, N. J., and Chen, Y. (2015b). Towards news verification: Deception detection methods for news discourse. In *Hawaii International Conference on System Sciences*.
- Sawilowsky, S. S. (2009). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

LexPorBr Infantil: uma base lexical tripartida e com interface Web de textos ouvidos, produzidos, e lidos por crianças

**Gustavo Estivalet¹, Nathan S. Hartmann², Vanessa Marquiafável³,
Katerina Lukasova⁴, Maria T. Carthery-Goulart⁴, Sandra M. Aluísio²**

¹Federal University of Paraíba, Department of Classical and Vernacular Letters

²University of São Paulo, Institute of Mathematics and Computer Sciences

³SpeechTera Desenvolvimento de Programas para Computadores Ltda

⁴Federal University of ABC, Center for Mathematics, Computation and Cognition

gustavoestivalet@hotmail.com

{nathansh, sandra}@icmc.usp.br

marquiafavel@gmail.com

{teresa.carthery, katerina.lukasova}@ufabc.edu.br

Abstract. *Corpora of words have been widely used in the selection of stimuli in psycholinguistic experiments, research in lexicology, among others uses. Word frequency is a great proxy in psycholinguistics research, since it provides a good time response in word recognition. This paper presents the first of the three corpora of the LexPorBr Infantil project: subtitles of family, comedy, and children movies and series listen by children in Brazilian Portuguese. This work provides a large lexicon of words, publicly available, with almost 130 million tokens and 880 thousand types where each word is annotated with 48 categories for psycholinguistic research, corpus analysis and research in education.*

Resumo. *Córpus de palavras têm sido largamente utilizados na seleção de estímulos em experimentos psicolinguísticos, pesquisas em lexicologia, dentre outros usos. A frequência de palavras é um importante proxy em pesquisas psicolinguísticas, pois prevê com boa precisão os tempos de reação para o reconhecimento de palavras. Este artigo apresenta o primeiro dos três córpus do projeto LexPorBR Infantil: legendas de filmes e séries de comédia, família e animações em português brasileiro ouvidos por crianças. Este trabalho disponibiliza publicamente um léxico de 130 milhões de tokens e 880 mil types, disponibilizando 48 categorias de informações para pesquisas em psicolinguística, análise de córpus e aplicadas à educação.*

1. Introdução

Pesquisas em linguística de córpus têm sido uma poderosa ferramenta para o estudo das línguas naturais nos mais diversos níveis de análise assim como na interface com várias disciplinas. Podemos diferenciar córpus de textos baseados na análise de sentenças e córpus de palavras baseados na análise de entradas lexicais [Brysbaert and New 2009].

A frequência de palavras é uma das variáveis mais importantes nas pesquisas em psicolinguística, estabelecendo que palavras que são lidas e ouvidas mais frequentemente são reconhecidas mais rapidamente do que palavras menos frequentes [Brysbaert et al. 2011a]. Os círpus a partir dos quais as frequências das palavras são calculadas devem conter entradas variadas e representativas da língua (ouvida e/ou escrita). Além disso, é interessante que estes círpus possam ser ajustados para populações específicas, uma vez que a exposição a conteúdos linguísticos varia conforme a idade, nível socioeconômico, grau de escolarização, entre outros.

Em geral, os círpus são compostos por um grande conjunto de livros, jornais e revistas que têm oferecido parâmetros satisfatórios para a população adulta e escolarizada, apesar de alguns contrastes em relação à familiaridade de entradas lexicais cuja ocorrência predomina na linguagem oral ou na escrita. Do ponto de vista lexical, a linguagem escrita tende a ser mais redundante em função da falta de interação comunicativa, o que pressupõe a escolha de itens lexicais menos ambíguos. Assim, círpus construídos unicamente a partir de textos escritos são mais problemáticos para estudos psicolinguísticos envolvendo crianças.

Para lidar com estes vieses, pesquisas em linguística de círpus têm incluído materiais que possam se aproximar da linguagem oral para maior precisão na obtenção de parâmetros de frequência. Nesse sentido, mensagens de texto, e-mails e legendas de filmes e programas de televisão têm sido apontados como materiais mais representativos da língua falada. Alguns estudos mostraram que a frequência obtida a partir de legendas prevê com mais precisão os tempos de reação para o reconhecimento de palavras [Brysbaert and New 2009, Brysbaert et al. 2011a, Brysbaert et al. 2011b, van Heuven et al. 2014].

Uma vantagem adicional do uso de legendas é a possibilidade de se ajustar o círpus para finalidades específicas a partir de filtros, por exemplo, estudos com crianças de diferentes faixas etárias e escolaridades. O objetivo das bases lexicais infantis é oferecer uma ferramenta sensível para estudos psicolinguísticos e de desenvolvimento com controle das variáveis psicolinguísticas para cada faixa etária [Corral et al. 2009]. Para tais propósitos, o uso de círpus baseado em textos escritos que somente a população adulta está exposta tem recebido críticas, pois as variáveis psicolinguísticas, tais como frequência das palavras, tendem a não refletir a realidade linguística do mundo infantil, uma tendência observada nos círpus já existentes. Assim, o crescente interesse em desenvolver estudos para crianças com foco experimental requer a existência de bases de dados lexicais contendo informação psicolinguística ajustada para esse público.

Com o objetivo de se responder a seguinte pergunta: *quais são as palavras que as crianças escutam com maior frequência*, este artigo apresenta o primeiro dos três círpus do projeto LexPorBR Infantil, que inclui: i) legendas de filmes e séries ouvidos por crianças, ii) textos escritos por crianças e iii) textos lidos por crianças. O LexPorBR Infantil - Oral (Legendas) foi compilado a partir de filmes e séries classificados de acordo com o círpus SubIMDb-PT [Paetzold and Specia 2016] como gêneros familiares (filmes/séries para família, comédia, crianças e animação), pois esses gêneros são destinados a crianças, famílias e/ou para todos, apresentando assim uma linguagem

acessível¹. Legendas de filmes e séries têm sido consideradas material linguístico que apresentam o vocabulário mais próximo do dia-a-dia e da linguagem oral, logo, possuem normas de frequência otimizadas para a formação de córpus no estudo sobre a aquisição, o processamento e a utilização da linguagem tanto por crianças como por adultos [Brysbaert and New 2009, Soares et al. 2014a].

Este trabalho teve dois objetivos principais: i) disponibilizar de forma pública um grande léxico de palavras, com 129.053.297 *tokens* e 874.887 *types*, acessado por uma interface Web; ii) calcular 48 categorias importantes para pesquisas em psicolinguística, análise de córpus e aplicadas à educação (por exemplo: lexema, forma fonológica, POS, flexão nominal (gênero e número) e flexão verbal (modo, tempo, pessoa e número), silabação, sílaba tônica, número de letras, fonemas e sílabas, vizinhos ortográficos e fonológicos, OLD20/PLD20 (Orthographic/Phonological Levenshtein Distance).

2. Trabalhos Relacionados

Lexin [Corral et al. 2009] é uma base de vocabulário infantil do Espanhol composta por 134 livros destinados a estimular a leitura e a escrita em crianças na pré-escola (76 livros) e no primeiro ano do ensino fundamental (58 livros). A base contém 13.184 palavras (*types*) e 178.839 *tokens*. Manulex [Lété et al. 2004] é uma base do Francês composta por textos da 1^a a 5^a série da escola fundamental (idade 6 a 11 anos). A seleção do material foi feita considerando-se a frequência cumulativa referente aos livros mais vendidos pelas principais editoras francesas no ano de 1996. A base contém um total de 48.886 palavras e 23.812 lemas. Novlex [Lambert and Chesnet 2001] é mais uma base do Francês que descreve material didático da 3^a série e de leitura correspondente (19 títulos). A base totaliza 20.600 palavras e 9.300 lemas.

Entre as variáveis que são geralmente computadas nos referidos córpus, encontram-se: frequência das palavras, ocorrência de uma palavra em diferentes textos, frequência por mil/milhão, log da frequência por mil/milhão, categoria gramatical, número de letras, estrutura silábica, ano escolar com maior probabilidade da criança se deparar com a palavra, além de outras variáveis relacionadas com as estrutura ortográfica e fonológicas das palavras.

Em português europeu, Escolex [Soares et al. 2014b] é um córpus composto por 3,2 milhões de palavras (3.211.805 *tokens* e 48.381 *types*) coletadas de uma base de 171 livros escolares do 1º ao 6º ano do Ensino Fundamental (crianças de 6 a 11 anos). Além das medidas já relatadas nas outras bases, Escolex estima também diversidade contextual.

Em português brasileiro, o projeto que mais se aproxima do LexPorBR Infantil - Oral (Legendas) é o SUBTLEX-PT-BR [Tang 2012], compilado a partir do site de legendas OpenSubtitles em dezembro de 2012, com 61 milhões de *tokens* e 136.147 *types*, mas que não traz conteúdo direcionado para crianças. Este córpus disponibiliza: i) unigramas com OLD20; ii) lemas e POS *tagging*; e iii) bigramas, que são úteis para obtenção da frequência das colocações e para a identificação de palavras compostas.

¹Entretanto, cabe aqui uma ressalva: o córpus foi compilado automaticamente e assim incluiu filmes com temas adultos do gênero comédia, em alguns casos.

3. O Processamento do córpus LexPorBR Infantil - Oral (Legendas)

O SubIMDb-PT é um córpus composto por legendas de filmes e séries infantis, utilizado no trabalho de [dos Santos et al. 2017], mas não descrito anteriormente. SubIMDb-PT foi compilado a partir do site de legendas OpenSubtitles em janeiro de 2017 e foi utilizado para o desenvolvimento do presente trabalho, o LexPorBR Infantil - Oral (Legendas), seguindo a mesma metodologia apresentada para a compilação do SubIMDb-EN [Paetzold and Specia 2016]. O córpus base não contém marcações que distinguem o início do fim da legenda de cada produção, pois cada filme/série dos gêneros familiares em português brasileiro foi concatenado em um único arquivo, sem anotação.

3.1. Segmentação, limpeza e tokenização das legendas

O gênero legendas é caracterizado por sentenças curtas, pois as mesmas devem ser dispostas na tela da televisão. Muitas sentenças, no entanto, não são suficientemente curtas e parte delas é exibida em diferentes *frames* na televisão. No córpus SubIMDb-PT, esse fenômeno é representado por sentenças quebradas por orações, seguidas por quebras de linhas. Para lidar com esse fenômeno, foi necessária a reconstrução das sentenças do córpus. Ainda, foram aplicados filtros de remoção de: (i) marcações de fala (travessão), tornando o córpus de legendas em um texto corrido; (ii) endereços de sites; (iii) referências ao editor/criador da legenda (conteúdo recorrente em legendas traduzidas autores amadores); (iv) *tokens* contendo caracteres diferentes do alfabeto do português brasileiro. Após o pré-processamento inicial, o córpus foi tokenizado com o *TreebankWordTokenizer* do NLTK². A Tabela 1 apresenta estatísticas em relação aos *tokens*, *types* e *type/token ratio* (TTR) identificados durante o pré-processamento, assim como estatísticas de dois projetos relacionados ao LexPorBR Infantil - Oral (Legendas).

Córpus	Tokens	Types	TTR
LexPorBR Infantil - Oral (Legendas) original	168.888.430	927.023	0,55%
Aplicação de filtros de limpeza realizados	129.053.297	874.887	0,68%
Aplicação de Léxico de língua UNITEM (DELAF)	121.281.557	289.001	0,24%
Escolex	3.211.805	48.381	1,50%
SUBTLEX-PT-BR	61.000.000	136.147	0,22%

Tabela 1. Distribuição de *tokens*, *types* e *type/token ratio* do LexPorBR Infantil.

Com base no TTR, percebemos que o córpus possui maior riqueza lexical do que o SUBTLEX-PT-BR e menor riqueza lexical do que o Escolex. Observa-se ainda que, após a aplicação dos filtros, houve um aumento da riqueza lexical tendo em vista que houve, proporcionalmente, uma grande diminuição de *tokens* e uma pequena diminuição de *types*. Quando comparado com o léxico do dicionário UNITEM-PB DELAF [Muniz 2004], houve uma diminuição da riqueza lexical em função da grande diminuição de *types* de baixa frequência e, consequentemente, baixa diminuição de *tokens*.

Calculamos também a distribuição de frequência de *types* por decil, onde 50% deles possuem uma única ocorrência (cauda longa), 10% possuem 2 ocorrências, 10% possuem 3 ou 4 ocorrências, 10% possuem de 5 a 7 ocorrências, 10% possuem entre 8 e 24 ocorrências e somente 10% possuem mais que 25 ocorrências, podendo chegar a

²<https://www.nltk.org>

frequência máxima de 3.480.284. A baixa frequência para a grande maioria dos *types* do córpus pode ser justificada por erros de digitação, problemas de codificação do arquivo original da legenda (produzindo caracteres estranhos) e, claro, palavras raras para o gênero. Exemplos de *types* com frequência 1 no córpus são: N'attend, Novo.então, Peregrinaã§ãues, Opelette, L-e-v-e-d-a-ç-ã-o, Estiércol, Pórticos, Gabiente, Ruptured.

Os *types* com frequência 1 foram avaliados no DELAF, sendo que apenas 33% deles são palavras da língua. Essa mesma análise para todo o córpus (linha 3 na Tabela 1) implica na redução de 68,8% no número de *types*, indicando o alto número de *tokens* ruidosos ou não presentes no DELAF. No entanto, vale lembrar que este dicionário não contempla uma grande variedade de nomes próprios, neologismos e estrangeirismos. Assim, com o objetivo de preservar palavras raras e nomes próprios, não removemos *types* de baixa frequência.

3.2. Tagging e Lematização

Para realizar o *POS tagging* e lematização no LexPorBR Infantil - Oral (Legendas), utilizamos o nlpnet [Fonseca et al. 2015] alinhado com o dicionário UNITEX-PB DELAF para mapearmos os lexemas com suas respectivas etiquetas morfossintáticas para o lema adequado. O nlpnet é um etiquetador morfossintático amplamente utilizado, treinado no córpus MacMorpho [Aluísio et al. 2003], que foi revisado para melhorar a tarefa de *POS tagging*³, mas que não possui o mesmo conjunto de etiquetas morfossintáticas do DELAF. Para realizar o mapeamento entre as categorias do DELAF e as 25 categorias morfossintáticas da versão 3 do MacMorpho utilizada no nlpnet, fizemos um relaxamento nas etiquetas do DELAF, considerando somente as categoria principal e geral.

3.3. Outros léxicos utilizados no estudo

A fim de estudarmos a aderência do vocabulário utilizado nas legendas do nosso córpus com o esperado por crianças, fizemos uso de dicionários sugeridos pelo Programa Nacional do Livro Didático (PNLD) do Ministério da Educação (MEC). Esses dicionários foram categorizados por níveis de complexidade lexical esperada em cada etapa escolar, previamente compilados no trabalho de [Hartmann et al. 2018]. O dicionário de Tipo 1, composto aqui pelo dicionário Caldas Aulete com a Turma do Cocoricó, contempla o 1º ciclo do Ensino Fundamental 1 (1º ao 3º ano) e possui 1.371 entradas; o dicionário de Tipo 2, composto pelo Dicionário Escolar da Língua Portuguesa, Dicionario Ilustrado de Português e Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Pica-Pau Amarelo, contempla o 2º ciclo do Ensino Fundamental 1 (4º ao 5º ano) e possui 8.171 entradas; e o dicionário de Tipo 3, composto aqui pelo Minidicionário Contemporâneo da Língua Portuguesa, contempla o Ensino Fundamental 2 (6º ao 9º ano) e possui 29.970 entradas.

Também utilizamos léxicos amplamente utilizados em pesquisas do Processamento de Linguagem Natural: o UNITEX-PB, já apresentado nessa seção, contendo 7.580.357 palavras; e o Hunspell, dicionário eletrônico frequentemente utilizado em recursos computacionais, contendo 312.418 palavras.

³<http://nilc.icmc.usp.br/macmorpho>

4. O transcritor fonético Petrus e as categorias de silabação, sílaba tônica e transcrição fonológica

O processo de conversão de textos ortográficos em seus correlatos sonoros é chamado de conversão grafema-fonema (do inglês *grapheme-to-phoneme* - G2P) ou transcrição letras-som. O Petrus 2.0 (*PhonEtic TRanscriber for User Support*) [Serrani 2015] foi o sistema de conversão G2P utilizado para a obtenção da silabação, da indicação do acento primário e da transcrição fonológica das palavras contidas no LexPorBR Infantil - Oral (Legendas).

4.1. Divisão silábica e sílaba tônica

A metodologia adotada para a marcação da sílaba tônica (acento primário) em palavras simples do português brasileiro foi baseada nas regras publicadas por [Silva et al. 2006] devido à completa documentação e disponibilização dos algoritmos desenvolvidos. A taxa de acerto de 93% obtida em um córpus de teste composto por 52.525 palavras também foi decisiva para sua escolha, visto ser a maior entre os sistemas testados.

Vale mencionar que existem diferenças entre o conjunto de regras adotados para uma divisão silábica para efeitos de translineação e uma divisão silábica feita com base fonológica. Os algoritmos de silabificação propostos por [Silva 2011] foram desenvolvidos com a intenção de conciliar as teorias fonológicas da língua com as necessidades de sistema de síntese de fala. A Tabela 2 traz exemplos de palavras divididas siladicamente conforme as regras adotadas pelo Petrus e pelo Dicionário online Caldas Aulete.

	Petrus	Dicionário online Caldas Aulete
obstrução	o.bs.tru.ção	obs.tru.ção
advogado	a.d.vo.ga.do	ad.vo.ga.do
arredondar	a.rre.don.dar	ar.re.don.dar
assado	a.ssa.do	as.sa.do

Tabela 2. Divisão silábica do Petrus e do Dicionário online Caldas Aulete.

4.2. Transcrição fonológica

A obtenção da transcrição fonológica se deu da seguinte forma: o último módulo do referido sistema é responsável por realizar a transcrição fonética das palavras, que é feita a partir de diferentes níveis de informação sobre a palavra em análise, desde a presença ou não de um prefixo até a categorização gramatical, identificação da vogal tônica e da divisão silábica. Dessa forma, criou-se manualmente um conjunto de regras linguísticas dependentes de contexto (silábico, acentual e gramatical), que aliado ao uso de um dicionário fonético (ou seja, uma lista de palavras cujas transcrições fonéticas não seguem as regras de transcrição propostas) indica como transcrever os grafemas em suas respectivas unidades fonéticas. Os resultados obtidos com o Petrus indicaram uma taxa de acerto de 97.5% ao fone [Serrani 2015].

Por fim, a partir do *output* fonético gerado ao fim do processamento, e com base em uma lista de correspondência fone-fonema, fez-se a conversão da transcrição fonética gerada em IPA para uma transcrição fonêmica em alfabeto SAMPA adaptado. O Petrus foi desenvolvido para transcrever palavras simples do português brasileiro. Portanto, palavras compostas, estrangeirismos que não estejam em sua base de dados, palavras com

erros ortográficos ou qualquer outro tipo de sequência gráfica que não seja natural do português brasileiro apresentará transcrição, segmentação e marcação de tônica inadequadas no LexPorBR Infantil - Oral (Legendas).

5. Plataforma Web LexPorBR Infantil: construção e pesquisa

Com o objetivo de disponibilizar o máximo de informação sobre as palavras do LexPorBR Infantil - Oral (Legendas), 48 colunas com dados lexicais e metalingüísticas foram criadas e derivadas. As colunas com categorias de informações do LexPorBR Infantil - Oral (Legendas) são listadas na Tabela 3.

1. Lexema	13. Orto_freq/M	25. Lema_freq_log10	37. CVCV_sílaba
2. Fonologia	14. Orto_freq_log10	26. Lema_escala_zipf	38. Tipo_1
3. POS	15. Orto_escala_zipf	27. Nb_homógrafas	39. Tipo_2
4. POS_MM	16. Orto_zipf_rank	28. Vizinhos_ortográficos	40. Tipo_3
5. POS_DELAF	17. Fono_freq	29. OLD20	41. UNITEX
6. Sílabas/Tônica	18. Fono_freq_laplace	30. PUO	42. Hunspell
7. Nb_letras	19. Fono_freq_lexema/M	31. Nb_homófonas	43. Dicionário
8. Nb_fonemas	20. Fono_freq_log10	32. Vizinhos_fonológicos	44. Invertida_lexema
9. Nb_sílabas	21. Fono_escala_zipf	33. PLD20	45. Invertida_fono
10. Lema	22. Fono_zipf_rank	34. PUF	46. Invertida_lema
11. Orto_freq	23. Lema_freq	35. CVCV_lexema	47. Invertida_sílabas
12. Orto_freq_laplace	24. Lema_freq/M	36. CVCV_fonologia	48. Invertida_CVCV

Tabela 3. Categorias de informação do LexPorBR Infantil - Oral (Legendas).

As células número de letras/fonemas/sílabas foram calculadas através de um contador de caracteres; as células número de homógrafas/homófonas foram calculadas a partir de um contador de lexemas/fonologia repetidos, respectivamente; a célula fono_freq_laplace foi calculada a partir da frequência do córpus + 1, para comparação com outros córpus [Brysbaert and Diependaele 2012]; a célula fono_freq/M, contendo a frequência da palavra entre 1 milhão de palavras, foi calculada a partir da divisão da frequência de La Place pelo total de *tokens* do córpus; a célula fono_freq_log10 foi calculada a partir do log base 10 das frequências por milhão, com o objetivo de se linearizar a distribuição das frequências; as células fono_escala_zipf e fono_zipf_rank foram calculadas com objetivo de comparação das frequências entre diferentes córpus usando uma distribuição linearizada e ranqueada, respectivamente [van Heuven et al. 2014]. As células correspondentes às frequências ortográficas foram derivadas do Léxico do Português Brasileiro [Estivalet and Meunier 2015]. As células vizinhos ortográficos/fonológicos foram calculadas através da comparação de cada entrada lexical com todas as demais palavras do córpus a partir da distância de Hamming = 1 (substituição de 1 letra por vez); as células distância ortográfica/fonológica de Levenshtein 20 (OLD20/PLD20) apresentam uma medida mais flexível de semelhança lexical a partir do cálculo do número de inserções, exclusões ou substituições das 20 palavras mais próximas [Yarkoni et al. 2008]. Para uma maior precisão destas normas, estas quatro categorias foram calculadas primeiramente entre as palavras existentes em pelo menos um dicionário dentre os utilizados neste trabalho e posteriormente para as demais entradas lexicais do córpus. As células ponto de unicidade ortográfico/fonológico (PUO-PUF) apresentam a informação sobre a partir de que letra/fonema a entrada lexical é única no léxico através de uma comparação com a entrada anterior e posterior no córpus organizado em ordem alfabética. As células CVCV apresentam a estrutura de consoantes e

vogais das formas, calculadas substituindo-se as vogais por V e consoantes por C. As células invertidas apresentam as respectivas entradas lexicais na forma invertida. Enfim, a célula POS, contendo 9 etiquetas gerais (ADJ, ADV, FUNC, IN, N, NUM, PCP, PRO, V), foi criada através da simplificação das 25 etiquetas específicas da célula POS_MM; considerou-se PROSUB = N, PROADJ = ADJ, CUR = NUM, ART + CONJ + PDEN + PREP + PRO = FUNC e desconsideraram-se as contrações, como por exemplo ADV-KS = ADV. A célula POS_DELAF, contendo informações sobre flexão nominal e flexão verbal foi derivada a partir das definições do dicionário UNITEX-PB DELAF; as células Tipo_1/Tipo_2/Tipo_3/UNITEX/Hunspell apresentam uma marcação binária marcando se a entrada lexical está presente nestes materiais; ainda, a célula dicionário apresenta esta marcação se a palavra está presente em pelo menos um dentre estes cinco dicionários.

A interface apresenta dois motores de pesquisa: a pesquisa simples permite a inserção de uma lista de palavras a serem procuradas e a pesquisa complexa permite a inserção de uma série de especificações lexicais a serem procuradas ou evitadas.

6. Conclusões e Trabalhos Futuros

Os dois objetivos principais do presente estudo foram atingidos com êxito: i) criamos e disponibilizamos publicamente o LexPorBR Infantil - Oral (Legendas) com 129.053.297 *tokens* e 874.887 *types*; e ii) calculamos e derivamos 48 categorias de informações lexicais e metalingüísticas, contribuindo para o desenvolvimento de recursos lexicais carentes na pesquisa em psicolinguística e análise de córpuses. O córpus criado apresenta alta riqueza lexical de um vocabulário oral de fala cotidiana derivado de legendas de filmes e séries familiares/infantis. No melhor do nosso conhecimento, este é o primeiro córpus baseado em palavras que apresenta as formas fonológicas e silábicas do português brasileiro. O LexPorBR Infantil - Oral (Legendas) pode ser acessado por interface web⁴.

O LexPorBr Infantil pode ser usado de diferentes maneiras; seguem alguns exemplos abaixo. A base lexical é uma fonte valiosa de estímulos para estudos de aprendizagem e cognição de crianças. Portanto, pesquisadores da psicologia cognitiva, linguistas, neurocientistas, professores e outros podem se beneficiar da nossa base lexical para selecionar e combinar palavras de seu interesse e para monitorar suas propriedades psicolinguísticas. Isso tem sido feito em diferentes estudos sobre leitura [Schuster et al. 2015], memória de trabalho [Dominic et al. 2018], desenvolvimento de linguagem [Tomasello 2003] e muitos outros. Outra forma de utilização da base lexical é a exploração do próprio conteúdo para análise de redes semânticas, comparações com outras linguagens e geração de trajetórias interlingüísticas.

Os trabalhos futuros incluem a compilação de mais dois córpuses para o LexPorBr Infantil: produzido por crianças e escrito (textos), para que tenhamos mais abrangência nos materiais, gêneros e frequências das palavras escutadas, lidas e escritas por crianças. Para estes córpuses, pretende-se calcular a diversidade contextual. Também, pretendemos aprimorar o módulo de silabação e transcrição fonológica, enriquecer as entradas com informação morfológica e possibilitar a pesquisa dos contextos linguísticos originais (córpus de texto) onde as palavras ocorrem.

⁴<http://lexicodoportugues.com/infantil>

Referências

- Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., and Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 110–117. Springer.
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., and Böhl, A. (2011a). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in german. *Experimental Psychology*, 58:412–424.
- Brysbaert, M. and Diependaele, K. (2012). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, 45(2):422–430.
- Brysbaert, M., Keuleers, E., and New, B. (2011b). Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2:27.
- Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.
- Corral, S., Ferrero, M., and Goikoetxea, E. (2009). Lexin: a lexical database from spanish kindergarten and first-grade readers. *Behavior Research Methods*, 41(4):1009–17.
- Dominic, G., Jean, S.-A., Gerald, T., and Anne, T. (2018). Does neighborhood size really cause the word length effect? *Memory cognition*, 46(2):244–260.
- dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., and Aluísio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. In *International Conference on Text, Speech, and Dialogue*, pages 281–289. Springer.
- Estivalet, G. L. and Meunier, F. (2015). The brazilian portuguese lexicon: An instrument for psycholinguistic research. *PLOS ONE*, 10(12).
- Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.
- Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. (2018). Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.
- Lambert, E. and Chesnet, D. (2001). Novlex: Une base de données lexicales pour les élèves de primaire [novlex: A lexical database for primary school children]. *L'Année Psychologique*, 2:215–235.
- Lété, B., Sprenger-Charolles, L., and Colé, P. (2004). Manulex: A grade-level lexical database from french elementary school readers. *Behavior Research Methods*, 36:156–66.
- Muniz, M. C. M. (2004). A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, ICMC-USP, São Carlos.

- Paetzold, G. and Specia, L. (2016). Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.
- Schuster, S., Hawelka, S., Richlan, F., Ludersdorfer, P., and Hutzler, F. (2015). Eyes on words: A fixation-related fmri study of the left occipito-temporal cortex during self-paced silent reading of words and pseudowords. *Scientific Reports*, 5(12686).
- Serrani, V. M. (2015). *Ambiente web de suporte à transcrição fonética automática de lemas em verbetes de dicionários do português do Brasil*. PhD thesis.
- Silva, D. C., de Lima, A. A., Maia, R., Braga, D., de Moraes, J. F., de Moraes, J. A., and Resende, F. G. (2006). A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing. In *2006 International Telecommunications Symposium*, pages 550–554. IEEE.
- Silva, D. d. C. (2011). Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em hmm. *Doutorado, Programa de Engenharia Elétrica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE/UFRJ), Rio de Janeiro*.
- Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J. J., Pinheiro, A. P., and Comesaña, M. (2014a). ESCOLEX: A grade-level lexical database from european portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1):240–253.
- Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J. J., Pinheiro, A. P., and Comesaña, M. (2014b). Escolex: A grade-level lexical database from european portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1):240–253.
- Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, 24:208–214.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, Cambridge, MA, US.
- van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.
- Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

B2W-Reviews01

An open product reviews corpus

Livy Real¹, Marcio Oshiro¹, Alexandre Mafra¹

¹B2W Digital – Tech Labs
São Paulo – BR – Brazil

{livy.coelho,marcio.oshiro,alexandre.mafra}@b2wdigital.com

Abstract. This paper describes B2W-Reviews01, an open corpus of product reviews. B2W-Reviews01 contains more than 130k e-commerce customer reviews, collected from the Americanas.com website between January and May, 2018. B2W-Reviews01 offers rich information about the reviewer profile, such as gender, age, and geographical location. The corpus also has two different review rates: the usual 5 point scale rate, represented by stars in most e-commerce websites, and also a ‘recommend to a friend’ label, a ‘yes or no’ question representing the willingness of the customer to recommend the product to someone else. By comparing these two rates, we found that the common approach of conducting sentiment analysis, based on a simplification over the 5 point scale, does not always reflect users’ sentiments about the product. It suggests that, for production applications, the approach of analyzing the 5 point scale rate as a three level scale can lead to wrong conclusions.

1. Introduction

In the era of machine learning and big data, one of the biggest bottlenecks of Natural Language Processing (NLP) and Computational Linguistics (CL) is having open corpora available. While there is a lot of information available on the internet, it is still difficult to have structured, high quality, curated data. This work aims to help fill this gap, presenting a new open corpus of product reviews in Brazilian Portuguese, the *B2W-Reviews01*.

In particular, customer product reviews represent a difficult information source for web crawlers, since there is no standardization on how to represent them on e-commerce websites. For example, on the *Americanas.com* website, reviews are displayed in product pages, but only the latest 5 left by customers automatically appear on that page. Another click is required to display 5 more reviews, as shown in figure 1.

Although customer generated data is often seen as a valuable by-product of on-line companies, few of them actually notice the value of making this information generally available. *B2W Digital* is a major e-commerce platform in Latin America. Its first e-commerce brand was released in 1999: *Americanas.com*. Today, the *B2W Digital* marketplace platform has three major brands in Brazilian e-commerce: *Americanas.com*, *Submarino*, and *Shoptime*. These brands support more than 25,500 sellers trading on its digital platform. For a company such as *B2W Digital*, being able to analyze and extract information from user reviews became a critical task. User reviews not only have a high impact on the reputation of products, sellers and services, but can also be seen as the first and most direct way to obtain feedback from customers. Although digital companies count on several techniques to track customer activity and assess customer satisfaction

ratings, analyzing product reviews often represents the easiest way to get customer feedback. Review analysis becomes especially relevant when the customer journey works as expected and, therefore, the customer does not need to contact Customer Service.

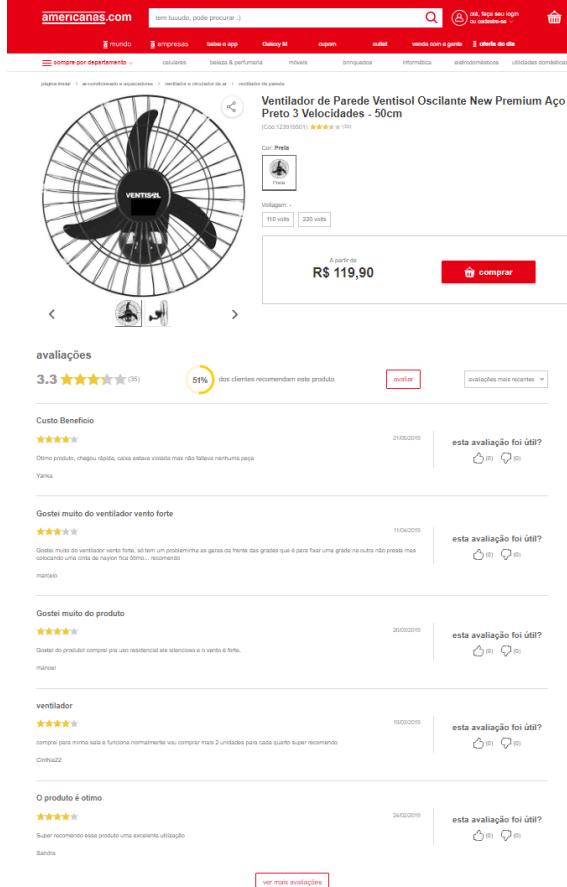


Figure 1. Example of product and product reviews display on *Americanas.com*

This corpus is a set of product reviews submitted to *Americanas.com* from January through May, 2018. *B2W-Reviews01* has both the review text and the meta-data related to each review: dates and times, ratings, geographical locations and ages of reviewers. *B2W-Reviews01* is available at <https://github.com/b2wdigital/b2w-reviews01/> under the license CC BY-NC-SA 4.0¹, which means that licensees may only copy, distribute, display, work on and make derivative works and remixes based on it if they give credit to *B2W Digital* in the manner specified in this work. Also, licensees may only distribute derivative works under a license identical (“not more restrictive”) to the license that governs the original work. Finally, licensees may use it for non-commercial purposes.

1.1. Aims and Usages

B2W-Reviews01 is a corpus which contains such varied information, that it can be useful for several NLP/CL tasks. The first that comes to mind is probably sentiment analysis. Sentiment analysis is the task of assigning a sentiment (or a position) to the content of a

¹<https://creativecommons.org/licenses/by-nc-sa/4.0/>.

given text. Thereunto, *B2W-Reviews01* offers two different evaluation ratings, described in section 3. Product reviews often have complex information, related not only to the product that was purchased, but also to the online shopping experience, payment methods, or even the product delivery process. Therefore, different facts and opinions can be extracted from such a corpus, and classifying sentiment may not be enough to capture the content of reviews [Wachsmuth et al. 2014]. For real world applications, dealing with topic modeling, user intent identification and feature extraction also become necessary. It is relevant to know not only the reviewer’s sentiment, but also the object of this feeling.

Since *B2W-Reviews01* offers the exact text written by users, this corpus also offers rich material for those interested on out-of-vocabulary words, slang identification, or spell-checker tasks. For those interested on socio-linguistics analysis, the present corpus offers a rich possibility of crossing reviewer information considering gender, age and geographical location. One can, for example, find easily how negative or positive reviews are distributed among age groups or which product categories receive more reviews from women or men. It is also possible to conduct a study on bias in reviews by joining and aggregating data. Although *B2W-Reviews01* is mainly a product review dataset, we believe that important insights about the current language in use in the web register can be made, since *Americanas.com* customers are spread throughout Brazil and have different social backgrounds.

2. Previous Works

To the best of our knowledge, the biggest product review data available is the *Amazon Customer Reviews Dataset*², which was made available by the *Amazon.com*³ website, containing more than 130M reviews in four different languages. A particularly interesting subset of the *Amazon Customer Reviews Dataset* corpus is the work of [Filatova 2012], who produced the *Sarcasm Amazon Reviews Corpus*⁴, a crowd-sourced annotated corpus that contains both sarcastic and non-sarcastic reviews of the same product. Another dataset made available by a marketplace is the *Rakuten*⁵ marketplace dataset, the major Japanese digital marketplace. *Rakuten* offers some 64M reviews in Japanese available upon request⁶.

For Portuguese, the situation of freely available data is not great. We have the Brazilian E-Commerce Public Dataset by *Olist*⁷, which has 100k product reviews, collected from 2016 to 2018 by *Olist*, which is also one of the biggest sellers in the *B2W Digital* marketplace. Although the Brazilian E-Commerce Public Dataset by *Olist* is open, it is distributed in several files, so anyone who is interested in capturing very basic information, such as the amount of positive or negative reviews per geographical location information, or what is the average product rating for a given category, has to process, join, and aggregate data. Although *Olist* and *B2W-Reviews01* data can be seen as two datasets of the same nature, the *Olist* corpus has information related to payment methods, for example, which *B2W-Reviews01* does not have, while our corpus offers more infor-

²<https://s3.amazonaws.com/amazon-reviews-pds/readme.html>

³<https://www.amazon.com>

⁴<https://github.com/ef2020/SarcasmAmazonReviewsCorpus/wiki>

⁵<https://www.rakuten.co.jp>

⁶https://rit.rakuten.co.jp/data_release

⁷<https://www.kaggle.com/olistbr/brazilian-e-commerce>

mation about the reviewer, making her/his gender and birth year available, for example. Therefore, one interested in the reviews considering the user profile will probably find more useful information in *B2W-Reviews01*.

Another corpus of reviews, collected from *Mercado Livre*, is the one used in [Hartmann et al. 2014], made available by its creators upon request. Few different works can also be found about the use of product reviews in Portuguese, such as [Avanço 2015], a Master’s thesis on normalization and classification of products that uses the data from [Hartmann et al. 2014]. [Siqueira and Barros 2010] uses reviews collected from the *Ebit* services web-page to extract features and polarity from users’ feedbacks, but the used corpus is not available. [Ribeiro et al. 2012] is another work on feature extraction and opinion classification from reviews obtained at the *Carros na Web*⁸ blog. However, collected data is also not available. [Nobre et al. 2016] also works on sentiment analysis using reviews collected from *Amazon.com* and *Buscape*⁹, but, yet again, the used corpora are not available.

Considering the amount of work done in information extraction, customer review analysis, and the fact that most of the works use different, not always available, corpora, we believe *B2W-Reviews01* can be really useful for the NLP Portuguese community. At least, this data can serve as a public and generic dataset, and further works interested on opinion mining and topic modeling of product reviews, among other subjects, could be compared and reproduced.

3. Corpus Description

The *B2W-Reviews01* corpus has 132,373 reviews, left by 112,993 different users regarding 48,001 unique products. The reviews were collected from January to May, 2018. All reviews submitted to the *Americanas.com* website are present in the corpus. It means that one can find offensive language, repeated reviews and reviews composed by only one word in the present data. These kinds of reviews are often not accepted to be displayed on the *Americanas.com* website. So, this means that this resource is richer than one could get crawling the *Americanas.com* website.

The reviews present in *B2W-Reviews01* have 3,160,781 tokens and 148,541 unique tokens¹⁰. The median number of tokens per review is 16 tokens, while the minimum found is 1 token and the maximum 795. The median number of tokens present in titles is 2, while the minimum found is 1 and the maximum is 30 tokens per title. The user names, identification codes, and nicknames were anonymized, but a unique `user_id` was kept to make identification of all reviews written by the same user possible. User attributes also include their gender, birth date and geographical location, making this corpus particularly interesting for social analysis of e-commerce customer behavior.

B2W-Reviews01 is distributed as a comma-separated values file (`.csv`), each line representing one customer review. Each review has with 14 fields, described in Table 1.

⁸<https://www.carrosnaweb.com.br>

⁹<http://www.buscape.com.br>

¹⁰Here, we consider token as the content between two white spaces. We did not pre-process the corpus to correct typos, for example.

#	Field	Data type	Description
1	submission_date	date/time	review submission date (format YYYY-MM-DD hh:mm:ss)
2	reviewer_id	string	unique reviewer id
3	product_id	integer	unique product id
4	product_name	string	product name
5	product_brand	string	product brand
6	site.category_lvl1	string	product category - first level
7	site.category_lvl2	string	product category - second level
8	overall_rating	integer	overall customer rating, from 1 to 5
9	recommend_to_a_friend	string	answer to "would you recommend this product to a friend?" ("Yes"/"No")
10	review_title	text	review title, introduces or summarizes the review content
11	review_text	text	main text content of the review
12	reviewer_birth_year	integer	reviewer's birth year
13	reviewer_gender	string	reviewer's gender ('F' for female; 'M' for male)
14	reviewer_location	string	reviewer's Brazilian State, according to the delivery address

Table 1. Fields, data types, and descriptions.

3.1. Reviews collection in *Americanas.com*

After a product is successfully delivered to the customer, the company sends an e-mail with a link to the product review form (Figure 2). This form can also be reached from the product page, so that any customer can write a review anytime, without needing to have bought the product on *Americanas.com*. The review form consists of an overall rating of the product ranging from 1 (bad) to 5 (excellent), a question on whether the customer recommends the product, a review title and the review text. All fields are required and the review text must have at least 50 characters.

The figure shows a screenshot of a product review form. At the top, it says "avalie este produto". Below that, there are five radio buttons for rating: "ruim", "regular", "bom", "ótimo", and "excelente". Next, there's a question "Você recomenda esse produto?*" with two radio buttons: "sim" and "não". Then, there's a field labeled "Título da avaliação*" with a placeholder "Exemplo: Gostei muito do produto!". Below that is a larger text area labeled "Escreva sua opinião*" with a placeholder "Escreva aqui sobre o produto". At the bottom left is a red button labeled "avaliar". At the very bottom, a small note reads "* A sua avaliação pode ser publicada em nossos sites e ajudará outras pessoas na escolha de seus produtos."

Figure 2. Product review form.

3.2. Examples and Discussion

One of the main goals of this work is to provide a data collection of opinions left by customers, with which one can get perceptions of evaluations across different user profiles and product features. Since we release the reviews exactly as they were written in the

B2W-Reviews01 corpus, one can find all kinds of noisy user-generated texts: simple typos, abbreviations, internet register and a vast amount of constructions hugely influenced by orality. One can also find offensive language and sarcasm. Here we present a few examples from the corpus.

Example 1.¹¹ In this example of a review scored as 5, one can see the language register present in many reviews. Even a review that can be considered well written lacks diacritics. In this example, we see a typical case of difficulty when processing the lack of diacritics in Portuguese: we miss the distinction of two of the most used words in Portuguese: ‘e’ (and) and ‘é’ (is). Of course, this characteristic imposes several challenges for one interested in processing the corpus. A task as simple as lemmatizing the data becomes a complex task when several kinds of mistakes, registers and typos appear indistinctly in the corpus.

Field	Value
submission_date	“2018-01-01 02:02:13”
reviewer_id	“a0fd1ad35b08d3b764ad6f884ef7183bf29fc7eb(...)”
product_id	122776350
product_name	“Ventilador de Teto Ventisol Fenix Premium Branco 3 velocidades com Controle Remoto”
product_brand	“ventisol”
site_category_lv1	“Casa e Construção”
site_category_lv2	“Climatização”
overall_rating	5
recommend_to_a_friend	“Yes”
review_title	“Gostei do produto”
review_text	“O barulho é minimo e o vento é bem forte na velocidade 2”
reviewer_birth_year	1987
reviewer_gender	“M”
reviewer_location	“SP”

Example 2.¹² In this example, the field `product_brand` has its value set to null as it is not present in the database.

Field	Value
submission_date	“2018-03-18 06:10:10”
reviewer_id	“ecd8648fee87789e041522b6d2e0ee5e22bcacb7(...)”
product_id	132326651
product_name	“Smart TV LED 48” Sony KDL-48W655D com Conversor Digital 2 HDMI 2 USB Wi-Fi Foto Sharing Plus Miracast Preta”
product_brand	null
site_category_lv1	“TV e Home Theater”
site_category_lv2	“TV”
overall_rating	5
recommend_to_a_friend	“Yes”
review_title	“Gostei muito da minha TV Smart.”
review_text	“Vocês estão de parabéns fez uma excelente entrega o produto chegou em perfeito estado gostei muito.obrigado.”
reviewer_birth_year	1986
reviewer_gender	“F”
reviewer_location	“MG”

Example 3.¹³ This example shows a typical case of offensive language and sarcasm being used, while the review is still positive. Considering the two rates we have

¹¹I liked the product: The noise is minimal and the wind is very strong at speed 2.

¹²I liked very much my Smart TV: Congratulations you made an excellent delivery the product arrived in perfect condition I enjoyed it very much.thank you.

¹³good: it sucks why cannot only give stars need to write something to evaluate whatafuck americanas change it.

related to the user's opinion, the user liked the product. However, the user's text is quite negative, but it is about the process of reviewing offered by *Americanas.com*, and not the product itself. The user pointed out that he preferred to only rate the product with stars and not be forced to write something about it.

Field	Value
submission_date	"2018-05-05 22:05:54"
reviewer_id	"a4297137bb957850899982a232218(...)"
product_id	31053501
product_name	"Smartphone Multilaser MS80 4G 32GB 5,7 HD 3GB RAM Android 7.1 Dual Camera 20MP+8MP dourado- P9065"
product_brand	null
site.category_lv1	"Celulares e Smartphones"
site.category_lv2	"Smartphone"
overall_rating	5
recommend_to_a_friend	"Yes"
review_title	"bom"
review_text	"que merdapq n pode so da estrela tem que escrever alguma coisa para avaliar koé americanas muda isso"
reviewer_birth_year	1999
reviewer_gender	"M"
reviewer_location	"SP"

Considering these three examples, one can have a perception about the challenges this dataset imposes. One can naively think that classifying opinions using user generated text content can be enough to tackle the problem of analyzing users' perceptions of products, but many times the text content does not match the rate given by the user. That is mostly because the review is frequently not about the product itself, but about a specific aspect of the customer purchase journey. In Example 1, the review is about the product, in Example 2, the review is about the delivery and finally, in Example 3, the review is about the process of writing a review. In all these cases, the product is rated as a 5 star product.

4. Polarity Classification Challenges

Polarity classification is a subtask of sentiment analysis often done with corpora such as *B2W-Reviews01*, since such data offers not only the user text content but also a score that is supposed to express the user's sentiment in relation to what s/he is writing about. However, as the examples show, the polarity of the text doesn't always express the reviewer's sentiment: Example 3 shows how the text content can be negative while the customer is seemingly satisfied with the purchased product. Therefore, polarity text classification, which is a task on its own, is not a forward way to get the sentiment of the customer and, if available, other information about the customer and the review can be used together to better understand customer sentiment about the product or her/his experience.

As in the *Amazon Customer Reviews Dataset* the main rate associated with a review in *B2W-Reviews01* is a 5 points scale rate, here called an **overall_rating**. [Rain 2013, Avanço and Nunes 2014] and [Nobre et al. 2016], following many other works looking at corpora reviews also rated in a 5 point scale, argue that 'it is better to only consider reviews rated from 0 to 2 as negative and only those rated 5 as positive. Rates 3 and 4 could be positive, negative or neutral within the same universe (ambiguous) and should be discarded for polarity classification' [Nobre et al. 2016]. Table 2 shows the distribution of polarity reviews in *B2W-Reviews01*, considering the literature.

However, when we consider the field **recommend_to_a_friend**, we see a unexpected, but complementary information: 72.8% of the customers actually recommend the

Negative (1-2)	35,758	27.01%
Neutral (3-4)	48,660	36.76%
Positive (5)	47,955	36.23%

Table 2. Reviews polarity classification following literature

product to a friend, most of them included in the ‘neutral’ portion of the reviews: 31,837 users that rated the reviews 4 stars and 14,434 users that rated the reviews 3 stars recommended the product to a friend.

This analysis suggests that the typical approach of distributing reviews scored 4 and 3 as neutral can lead one to wrongfully analyze user sentiment. As roughly pointed by [Liu 2012, Chap.03], the simplification of these scores is often related to the computational feasibility/precision of a given model. This is to say that this simplified analysis is often used because of technical issues and not based on what the data really shows. It is easier for an automated system to be correct when scoring a review among only three very different scores — and not among five slightly different classifications. Of course, future analyses of this case need to be carried out to really confirm that the cited approach of analyzing scored reviews is not the best. But it suggests that, for real case applications, this simplified analysis is not appropriate and that performing analyses based on what a model can do, and not based on data driven insights can lead to wrong perceptions of what your data actually says.

5. Conclusions and Future work

This work introduced the *B2W-Reviews01* corpus, a dataset composed by products reviews and the metadata related to them submitted to the *Americanas.com* marketplace between January and May, 2018. The main goal of this work is to describe an open and freely available dataset of product reviews that can be useful for further works interested on different NLP/CL tasks that can use such data.

We leave several tasks as future work. In particular, we are interested on topic modeling in reviews. Also seen as a feature extraction task, knowing the topic of the review — what the review is actually talking about: the delivery process, the whole user experience when buying in the marketplace, the product itself or a specific feature of it, for example — is a critical task for *B2W Digital*, since mining the opinions of users only becomes relevant when one can figure out what the opinion is about. Another experiment we leave as future work is the analysis of polarity reviews started in section 4. While computational models for classification work better when we have few and very distinct categories, if the final result of the carried out analysis does not represent a trustworthy conclusion that can be effectively applied to a business model, it would be better to invest more on augmenting the precision of said models before applying the output results to the business case we are interested in.

We also plan to have a smaller part of the data annotated for sarcasm, following [Filatova 2012], since detecting sarcasm in Portuguese is still an open task and *B2W-Reviews01* offers a rich content for it. Moreover, *B2W Digital* plans to periodically release open reviews corpora, considering the relevance of such data and the difficulty in having established open data in Portuguese.

References

- Avanço, L. V. (2015). Sobre normalização e classificação de polaridade de textos opinativos na web. Master's thesis, ICMC/USP.
- Avanço, L. V. and Nunes, M. G. V. (2014). Lexicon-based sentiment analysis for reviews of products in brazilian portuguese. *3th Brazilian Conference on Intelligent Systems*.
- Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. *LREC*.
- Hartmann, N. S., Avanço, L. V., Balage, P. P., Duran, M. S., Nunes, M. G., Pardo, T. A., and Aluísio, S. M. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. *LREC*.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Nobre, G., Justino, A., Tadao, F., Nunes, D., Takabayashi, D., and Küllian, R. (2016). Booviews: Aspect-based sentiment analysis on product reviews combining svm and crf in portuguese. *Student Research Workshop - PROPOR*.
- Rain, C. (2013). Sentiment analysis in amazon reviews using probabilistic machine learning. Master's thesis, Swarthmore College.
- Ribeiro, S. S., Junior, Z., Meira, W., and Pappa, G. L. (2012). Positive or negative? using blogs to assess vehicles features. *ENIA*.
- Siqueira, H. and Barros, F. M. M. (2010). A feature extraction process for sentiment analysis of opinions on services. In *WTI 2010*.
- Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., and Palakarska, T. (2014). A review corpus for argumentation analysis. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 115–127, Berlin, Heidelberg. Springer Berlin Heidelberg.

A bunch of helpfulness and sentiment corpora in Brazilian Portuguese

Rogério Figueiredo de Sousa, Henrico Bertini Brum, Maria das Graças Volpe Nunes

¹Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP – Brazil

{rogerfig, henrico.brum}@usp.br, gracan@icmc.usp.br

Abstract. This paper presents the UTLcorpus, a novel corpus in Brazilian Portuguese for helpfulness classification of online reviews. There is a lack of corpora in Brazilian Portuguese annotated with helpfulness information, therefore there are also few works on modeling and predicting helpfulness of online reviews in this language. Moreover, there is no reference corpus to ground those results. This work tries to partially solve this problem by presenting UTLcorpus, a huge amount of annotated online reviews regarding helpfulness. Since the source data also contain star score labels, this paper also explores polarity labels in the data set. Some experiments show that both tasks of predicting helpfulness and polarity are benefited by the use of this corpus.

1. Introduction

Navigating websites for buying clothes, picking travel locations or choosing a movie can be a hard task considering the number of choices we can find. It may be even harder if we are looking for user reviews to support our final decision. The high amount of comments in such websites can be a hold back for user looking for opinions that may help them to choose products/services, or even alert them for flaws already known to previous acquirers. Helpfulness Prediction (HP) is the task that aims to correctly predict whether a review or opinion is helpful for a user to read before acquiring a product or hiring a service.

Prediction models are usually based on supervised learning, therefore, demanding for linguistic resources (corpora) of labeled reviews. One of the challenges for helpfulness research in Brazilian Portuguese is the few number of available data sets. We present UTLcorpus, a data set composed of two automatic annotated corpora for helpfulness in that language. The data were extracted from two different domains: movie reviews from a Brazilian social network for movies¹ and app reviews from Google App Store².

The data was anonymized and preprocessed. Evaluations were carried out to bootstrap the corpora for the HP task and the results were compared to other literature corpora in Brazilian Portuguese. Since UTLcorpus also contains labels for binary polarity classification (star score indicative of positive and negative reviews) we also used literature methods for evaluating the data for this task.

¹www.filmow.com. Accessed in May 19th, 2019.

²play.google.com. Accessed in May 19th, 2019.

The main contribution of this work is the creation of resources mainly for helpfulness prediction, but also for the polarity classification task. The corpus created in this work should be useful to increase the research in these areas and help to find the particularities of the helpfulness modeling task, thus enabling the understanding of this phenomenon in Brazilian Portuguese.

The paper is organized as follows. Section 2 presents an overview of Helpfulness Modeling and Prediction Task. The UTLcorpus is presented in Section 3. Experiments with the corpus in the tasks of helpfulness prediction and polarity classification are presented in Section 4. In Section 5 some important literature works on Polarity Classification and Helpfulness Prediction are discussed. Finally in Section 6 some conclusions and future works are presented.

2. Helpfulness Prediction Task

Modeling and prediction online reviews helpfulness (quality, usefulness or utility [Liu 2012]) are relevant for ranking and displaying comments to users who search comments on products or services. Most e-commerce websites present the most useful ones first and delegate to the users the task of evaluating whether they are helpful or not. Questions like "Was this review helpful to you?" are presented to the users and the feedback allows the system to re-rank eventually the set of reviews.

The drawback of this functionality is that the reviews can take a long time to accumulate a good number of user feedback. This is especially noticeable in new reviews, which can even be useful, but because of their low posting time, they can not get sufficient votes to achieve the top of the ranking. This fact demonstrates one of the advantages of automating the task. Websites that do not have ranking systems can benefit as well as the rankings themselves can be improved by the use of helpfulness prediction. In addition, the prediction of helpfulness can be used to filter off low-quality reviews, which can improve other tasks, such as the reviews summarization [Anchieta et al. 2017].

Helpfulness prediction tasks mainly include score regression, binary review classification and review ranking. These three methods depend on the helpfulness score which is usually calculated for each review by the Equation 1. The score regression aims to predict the helpfulness score $h \in [0, 1]$. The binary review classification seeks to decide whether comments are helpful or not based on a specific threshold (e.g. $h > 0.5$). And the review ranking needs to order the reviews by their helpfulness according to a reference ranking.

$$h = \frac{\text{helpful votes}}{\text{helpful votes} + \text{unhelpful votes}} \quad (1)$$

Several features have been used to characterize helpfulness in the literature. Usually they are split in two categories: Content and Context features [Diaz and Ng 2018]. The content features are related to the information that can be extracted directly from the review, such as the text and the stars given by the author. And the context features are those extracted from outside the review, such as reviewer information. In the survey of [Diaz and Ng 2018] one can find the most important features of the literature.

Contrary to what occurs for Portuguese, some large English corpora with utility annotations are available:

- Multi-Domain Sentiment Dataset (MDSD) [He and McAuley 2016]³: Collected from Amazon.com. Contains 25 product categories and 1.422.530 reviews.
- Amazon Review Dataset (ARD) [Blitzer et al. 2007, McAuley et al. 2015]⁴: Also collected from Amazon.com, contains 24 product categories and 142.8 million reviews and includes more metadata information than MDSD.
- Ciao Dataset [Tang et al. 2013]⁵: Was collected from an extinct e-commerce website and contains 302.232 reviews. The main difference from the previous ones is that it contains a social network between their users.

For the best of our knowledge, there is only one available corpus in Brazilian Portuguese, the *Buscapé* [Hartmann et al. 2014], containing 28.774 product reviews annotated with information that can be used for calculating helpfulness. Therefore, this work presents a new *corpus* containing information of helpfulness to promote researches in this task.

3. The UTLcorpus

The data set is a collection of reviews extracted from two domains: movies and apps. 2.881.589 reviews (1.839.851 of movies and 1.041.738 of apps) were collected using two web crawlers. The domains were chosen for the popularity, the high amount of data and the presence of a public “like” counter in each review, which makes possible to infer a helpfulness label. Besides the “like” counter, the data also contains scores given by users to the movie/app they are evaluating. We used the later for inferring positive and negative labels.

The methodology for labeling the polarity was proposed in [Avanço 2015]. Each review has a 5-star score according to the author’s evaluation of the related movie/app. Reviews with 0 and 5 stars are ignored to avoid those cases in which the users stars are not coherent with the review text. Also the 3 star reviews are discarded because they usually contain positive and negative sentiment about the entity.

In order to label the data we looked for the utility labels in the data set. Both domains provide the number of “likes” a review received (indicating it was helpful for other users) and the main issue we faced was the lack of a counterpart indicating the number of “dislikes” were attributed to the review. The majority of works in the literature [Kim et al. 2006, Malik and Hussain 2017] divide the number of positive likes by the sum of likes and dislikes to obtain a value and determine a threshold of helpfulness for a data set.

Since we can not count on dislikes, we define helpfulness in UTLcorpus as following. First, we group the data by category (movie titles and app names). This is performed because more popular apps/movies aggregate more likes by review than the less popular ones. Then we sort the reviews by the number of likes each of them has received (ignoring the ones with zero likes, since we can not identify if they have anything helpful in

³<https://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

⁴<http://jmcauley.ucsd.edu/data/amazon/>

⁵<https://www.cse.msu.edu/~tangjili/trust.html>

	Movie review subset	App review subset
# documents	1.834.702	921.257
# types	1.828.647	419.713
# tokens	60.177.264	11.919.636
Avg. token per doc.	32.7994	12.9384
Helpfulness Labeled	1.833.691 <i>helpful: 381.083 (20%)</i>	898.847 <i>helpful: 50.166 (5%)</i>
Sentiment Labeled	862.768 <i>positive: 702.720 (81%)</i>	320.255 <i>positive: 113.351 (35%)</i>

Table 1. UTLcorpus information

Subset	Movie Review		App Review	
	Positive	Negative	Positive	Negative
Helpful	155.672	37.699	6.533	18.473
Unhelpful	546.357	122.029	98.374	174.465

Table 2. Intersection between classes

it). Next, we remove replications and then consider helpful any comment with more likes than the first percentile of the distribution. To determine the negative samples we consider any review with fewer likes than the threshold previously defined and crawled at least five days after the review was published; the observation of the timespan is important since recent reviews take longer to achieve higher like counts thus becoming a false negative noise in the data set.

Online reviews are classified as User Generated Content (UGC) [Krishnamoorthy 2015], a type of text which carries many noisy linguistic phenomenons such as typos and Internet slangs. In order to reduce the noise in UTLcorpus the data was normalized using Enelvo [Bertaglia 2017], a tool for normalizing UGC in Brazilian Portuguese⁶.

The data set is composed of two subsets representing two different domains:

Movie review corpus: The movie reviews corpus was obtained by crawling Filimow, a popular film social network containing reviews, scores, evaluations and general movie information. We crawled reviews from the 4.283 most popular movies in the platform (we stopped for storage reasons) and the reviews generally represent opinions about films, actors/actresses and all kind of experience the users can have in watching a movie. Reviews can be directed to recent blockbusters as well as classics and we made available the movie titles which the reviews are about. The class distribution is skewed and the majority class for helpfulness is of unhelpful (80% of the documents) and positive (around 80% of the documents). Even though the helpfulness and polarity tags come from the

⁶Available in <https://github.com/tfcbertaglia/enelvo>. Visited in March 19th, 2019.

same data set, some reviews do not have enough information to be part of both of the subsets (972.945 documents do not the overlap).

The most frequent terms in the corpus (ignoring stop-words such as demonstratives pronouns, conjunctions and punctuation) are: *movie*, *well*, *good*, *story* and *best*.

App review corpus: The App Review corpus was obtained by crawling Google Play Store⁷. The corpus contains app reviews and one important feature of this data set is the absence of reviews with zero stars, since it is mandatory to evaluate with a star score any review. We gathered reviews from 243 apps (the most popular ones) and the whole corpus contains 921.257 reviews. The data is also skewed in both labels, being unhelpful the majority class for helpfulness (95%), and negative the majority class for polarity (65%). 371.651 reviews have only one label and do not overlap.

The most frequent terms in the corpus (also ignoring stop-words) are: *app*, *best*, *great*, *can* and *cool*. It is interesting no notice that several words are more frequent in both corpora even though they are skewed for different polarity classes.

Table 1 contains detailed data set information. The skewing of the data is a challenge for machine learning classification methods and we address this issue in section 4. Table 2 presents the intersection between the classes (Helpfulness and Polarity) for both datasets. It is possible to see that, in the movie reviews subcorpus, 80% of the helpful comments have positive polarity, and so are 81% of the unhelpful comments. Moreover, in the app reviews subcorpus, most of the helpful reviews (73%) as well as of the unhelpful ones are negative (63%).

4. Corpus Evaluation

In order to observe and evaluate the characteristics of the corpora on classification tasks we performed experiments in both subsets of UTLcorpus, Movie Review corpus (MR) and App Review corpus (AR), using a baseline and machine learning classifiers for comparison purposes. Helpfulness Prediction can be seen as a task very similar to polarity classification thus we defined a baseline and also used three machine learning classifiers (Support Vector Machines⁸, Multi-layer Perceptron⁹ and Random Forest¹⁰) following the work of [Brum and Nunes 2018], originally proposed for polarity classification of sentences. The work of [Brum and Nunes 2018] used a grid search technique to set the hyper parameters.

For baseline purposes we represented each sentence using a 2-dimensional vector with the number of positive and negative terms using a Brazilian Portuguese sentiment lexicon – Sentilex [Silva et al. 2012], which contains Portuguese terms (eg. *bom*, *ruim*, *péssimo*) and their respective polarity label. We trained a SVM model using this feature representation and evaluated the data sets in a 10-fold cross validation scheme. Furthermore, we used three classifiers trained and evaluated on a 10-fold cross validation. To avoid the skewing of the majority class, the data were balanced randomly (by the minority class), thus reducing the data sets considerably. The final sizes of the corpora are

⁷<https://play.google.com/store>. Visited in March 19th, 2019.

⁸Hyper parameters – C: 1; alpha: 0.1; linear kernel.

⁹Hyper parameters – Activation: $\tanh(x)$; learning rate: 0.001; alpha: 0.0001; neurons: 200; layers: 2.

¹⁰Hyper parameters – number of estimators: 200.

762.078 documents in the MR corpus and 100.322 in the AR corpus, nevertheless, the final corpus is still larger than the Brazilian Portuguese corpus *Buscapé*.

To the other three classifiers, differently from the baseline method, the data was represented using pre-trained 600-dimensional word2vec embeddings trained in more than 1 billion Portuguese Brazilian user-generated content (tweets and forums). The representation is described in [Corrêa et al. 2017] and has also been used for polarity classification in [Brum and Nunes 2018].

Classifier	Movie Review			App Review			Buscapé		
	F1-Help	F1-No-Help	F1-Measure	F1-Help	F1-No-Help	F1-Measure	F1-Help	F1-No-Help	F1-Measure
Baseline	0.4499	0.6493	0.5496	0.5896	0.6617	0.6256	0.4967	0.6343	0.5655
Linear SVM	0.6341	0.6039	0.6189	0.7115	0.6436	0.6775	0.6072	0.6142	0.6107
MLP	0.6387	0.6118	0.6252	0.7082	0.6516	0.6799	0.6114	0.5983	0.6048
Random Forest	0.6220	0.5920	0.6072	0.7267	0.7182	0.7224	0.6361	0.6436	0.6398

Table 3. Helpfulness detection results

The results obtained in the classification are shown in Table 3. The F1 values presented in the table are acquired with 10-fold cross-validation technique (Mean of 10 executions). Before classifying the data the class distribution was balanced by using *undersampling*, in other words, we removed samples of the majority class before performing the cross-validation. Experiments with the unbalanced corpora resulted in F1 far below the ones presented in Table 3, even the best results had the minority class F1 below 0.1.

The baseline worked pretty well in relation to F1-Measure, but one can see that it does not handle well the positive class. The best results for helpfulness detection in both corpora were obtained using Random Forest classifier which predicts the class based on several estimators (Decision Trees). It is still uncertain if our method for defining the helpful class is reliable enough, but with this methodology it is still possible to predict the correct label in 60% of the time for movie reviews and 70% of the time for app reviews (*std.dev.* = 0.004). We believe that one possible explanation for the results is that the prediction of the utility does not depend on text only. We understand that helpfulness may be affected by the context (domain, category, website, etc.) in which the comment is inserted, as well as by the intention with which the reader is reading the comment.

Since UTLcorpus also has polarity labels we were able to perform experiments using them. The main difference was the baseline used: for polarity classification we represented the data similarly (positive and negative term frequency) but predicted as positive any sentence with more positive terms than negative ones. In Table 4 we can see the results for polarity classification in the corpora.

Classifier	Movie Review			App Review			Buscapé		
	F1-Pos	F1-Neg	F1-Measure	F1-Pos	F1-Neg	F1-Measure	F1-Pos	F1-Neg	F1-Measure
Baseline	0.6167	0.1675	0.3920	0.6378	0.1541	0.3959	0.5467	0.2031	0.3748
Linear SVM	0.6878	0.6517	0.6697	0.7687	0.7710	0.7698	0.8106	0.8243	0.8174
MLP	0.6602	0.6843	0.6722	0.7818	0.7814	0.7815	0.8146	0.8121	0.8133
Random Forest	0.6528	0.6644	0.6586	0.7588	0.7786	0.7686	0.8310	0.8128	0.8218

Table 4. Polarity classification results

Finally, we merged the corpora and perform experiments using the two domains. The choices of candidates reviews are the same for each domain and all selected by the criteria above mentioned are used this time. The results are presented on Table 5.

Classifier	Helpfulness Prediction			Polarity Classification		
	F1-Help	F1-No-Help	F1-Measure	F1-Pos	F1-Neg	F1-Measure
Baseline	0.6708	0.4583	0.5645	0.1570	0.6179	0.3874
Linear SVM	0.6299	0.6759	0.6529	0.7011	0.7578	0.7294
MLP	0.6383	0.6863	0.6623	0.7304	0.7646	0.7475
Random Forest	0.6398	0.6834	0.6616	—	—	—

Table 5. Helpfulness prediction and Polarity classification results using the whole UTLcorpus

For polarity classification we are able to better compare the results than for helpfulness since the literature contains more works relating that task. The results of Table 4 reached almost 0.8 F1 and one of the reasons may be that polarities are easier to separate from each other – usually people use different expressions and different words when evaluating positively or negatively a movie or app, the same does not always apply for helpfulness. Even though the best result obtained in *Buscapé* corpus was 0.8174 in F1, other authors achieved 0.8935% in the same corpus [Avanço et al. 2016].

One of the reasons for the low results is that the representation used (pre-trained word embeddings) usually works well with neural models since they basically rearrange the data using weights. It may explain why the best results for the whole corpus were obtained using MLP (Table 5), which follows the same principle with less layers. Another limitation was the size of the dataset. Linguistic approaches (which use n-grams for example) demand more resources for storage and processing. The *t-value* between results was measured in order to calculate the significance of the differences and all of them were significant at $p < 0.05$.

5. Related Work

This paper relates to several other works both in helpfulness detection and sentiment analysis due to its similarities between fields.

For helpfulness prediction as a classification task, [Krishnamoorthy 2015] examines the impact of some specific linguistic features based on a model named Linguistic Category Model (LCM) [Semin 2011], on helpfulness prediction task. The author builds three machine learning methods for helpfulness binary classification, using a threshold $h = 0.60$.

Using a corpus extracted from Amazon.com (MDSD), the Random Forest method achieved the best result reaching an average of 84% of F-measure using all features. Individually the LCM features obtained the best results.

[Zeng et al. 2014] addressed the helpfulness prediction problem as a three-class classification problem. The classes are (1) Helpful positive reviews (star rating $\epsilon [4,5]$ and helpfulness score $h > \text{threshold}$); (2) Helpful negative reviews (star rating $\epsilon [1,2]$

and helpfulness score $h > threshold$), and (3) Unhelpful reviews (helpfulness score $h < threshold$). They collected 8.690 reviews from Amazon.com. The experimentation included an empirical test to decide the helpfulness score threshold. The best value obtained 72.82% of accuracy on ten-fold cross-validation. Specifically, regarding each class, the helpful positives reached 69% in macro-f1; the helpful negatives, 79,5% in macro-f1 and the unhelpful ones, 80% in macro-f1.

[Malik and Hussain 2017] used an emotion score of reviews (confidence, surprise, anger, etc.) as a feature to predict helpfulness. The authors modeled and evaluated a set of learning methods on Amazon.com corpus (MDSD), and they achieved 89% of f-measure, using emotion features as input for a deep neural network method.

In [Hartmann et al. 2014] the authors introduce *Buscapé*, a corpus for user-generated content research constructed using product reviews from an e-commerce website in Brazilian Portuguese. The authors extracted 85.910 documents and annotated typos and Internet slangs for normalisation task. The corpus also contains a 5-star-based score and “like” votes that we used in this paper (section 4) for comparison with our own results in UTLcorpus.

This data set has been used several times in literature [Avanço et al. 2016, Brum and Nunes 2018, Bertaglia 2017]. We emphasize the work of [Avanço et al. 2016] because the authors classified the data using machine learning classifiers (SVM and Naive Bayes), lexical-based classifiers and ensemble of classifiers (both machine learning-based and lexical-based) and achieved the state-of-the-art for the corpus, 0.8935 in f1. The main difference of this paper with ours is that we only used machine learning classifiers and we used embeddings for data representation, whilst those authors used a combination of *bag-of-words* and linguistic features such as number of sentiment words and PoS tags.

We can also compare our work with [Corrêa et al. 2017] since they also annotated a large corpus for semantic purposes (polarity classification). In this paper the authors crawled Twitter for Brazilian Portuguese posts and used Distant Supervision, automatically labeling documents based on semantic clues, to form a large corpora for sentiment analysis. Pelesent is composed of 980.067 tweets that contained emojis and/or emoticons indicating negative or positive polarity (eg. “:)” for positive and “:(” for negative).

6. Discussion and future work

In this paper we presented UTLcorpus, a review corpus of two domains (movies and apps) with automatic labels for helpfulness detection and polarity classification. We proposed an automatic label methodology for helpfulness using “like” votes and used a literature inspired method for attaching a polarity (positive or negative) to 2.755.959 forming two corpora – one of Movie Reviews (1.834.702 documents) and one of App Reviews (921.257). The methodology was replicated in a similar corpus (*Buscapé*) in order for comparing sizes and results obtained in classification experiments.

UTLcorpus is one of the first data sets for helpfulness detection in Brazilian Portuguese, but it can also be used for sentiment analysis (polarity classification, aspect extraction or else) and for others NLP tasks such as language modeling, normalization, discourse analysis or semantic parsing, for example.

We evaluated the corpus using machine learning methods from the literature and

obtained results up for replication and comparison with other models. The dataset is available in the github github.com/RogerFig/UTLCorpus already pre-processed, normalized with Enelvo and anonymised in order to be used for research purposes.

One of the future works to be conduct is the exploration of state-of-the-art models for classification such as convolutional neural networks [Kim 2014] and Long-short Term Memory architectures as well as the investigation of different representation models – morphological-based embeddings [Bojanowski et al. 2017] or context embeddings such as Elmo [Gardner et al. 2017]. Another future work is to expand the usefulness prediction for handling comment rating so that a list of best comments can be presented to users. Finally, manual evaluation of helpfulness can be performed on reviews, although we believe that a reader who is not interested in a product will handle a review differently from an interested one.

References

- Anchiête, R., Sousa, R. F., Moura, R., and Pardo, T. (2017). Improving opinion summarization by assessing sentence importance in on-line reviews. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 32–36.
- Avanço, L. V. (2015). Sobre normalização e classificação de polaridade de textos opinativos na web.
- Avanço, L. V., Brum, H. B., and Nunes, M. d. G. V. (2016). Improving opinion classifiers by combining different methods and resources. *XIII Encontro Nacional de Inteligência Artificial e Computacional*.
- Bertaglia, T. F. C. (2017). Normalização textual de conteúdo gerado por usuário.
- Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Brum, H. B. and Nunes, M. d. G. V. (2018). Building a sentiment corpus of tweets in brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resources Association.
- Corrêa, E. A., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., and Brum, H. B. (2017). Pelesent: Cross-domain polarity classification using distant supervision. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 49–54. IEEE.
- Diaz, G. O. and Ng, V. (2018). Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 698–708.

- Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). AllenNLP: A deep semantic natural language processing platform.
- Hartmann, N., Avanço, L., Balage Filho, P. P., Duran, M. S., Nunes, M. D. G. V., Pardo, T. A. S., Aluísio, S. M., et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *LREC*, pages 3865–3871.
- He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW ’16*, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.
- Malik, M. and Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, 73:290–302.
- McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’15*, pages 43–52, New York, NY, USA. ACM.
- Semin, G. R. (2011). The linguistic category model. *Handbook of theories of social psychology*, 1:309–326.
- Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. *International Conference on Computational Processing of the Portuguese Language*, pages 218–228.
- Tang, J., Gao, H., Hu, X., and Liu, H. (2013). Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys ’13*, pages 1–8, New York, NY, USA. ACM.
- Zeng, Y.-C., Ku, T., Wu, S.-H., Chen, L.-P., and Chen, G.-D. (2014). Modeling the helpful opinion mining of online consumer reviews as a classification problem. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*, 19(2).

Um Corpus de Notícias Falsas do Twitter e Verificação Automática de Rumores em Língua Portuguesa

Paulo Roberto Cordeiro¹, Vladia Pinheiro¹

¹Programa de Pós-Graduação em Informática Aplicada, Av Washington Soares 1322,
Fortaleza, Ceará, Brazil

paulocordeiro@gmail.com, vladiacelia@unifor.br

Abstract. Due to the increased reach of fake news sharing, automated approaches to classifying what is true or false become urgent, especially for short news on social networks such as Twitter. A recent survey showed that 66% of Brazilians who answered the survey use social media as a news source. This article proposes a reference corpus of Fake News in Portuguese, collected from Twitter. We evaluated different machine learning algorithms for this task and the results showed that the Twitter rumors verification task in Portuguese supplanted the related results in English.

Resumo. Devido ao aumento no alcance do compartilhamento de notícias falsas, abordagens automatizadas para classificar o que é verdade ou mentira se tornam urgentes, principalmente para notícias curtas veiculadas em redes sociais como o Twitter. Uma pesquisa recente mostrou que 66% dos Brasileiros que responderam a pesquisa usam as redes sociais como fonte de notícias. Este artigo propõe um corpus de referência de Fake News em língua portuguesa, coletado do Twitter. Avaliamos diferentes algoritmos de machine learning para essa tarefa e os resultados mostraram que a tarefa de verificação de rumores do Twitter, em português, suplantou os resultados relacionados em língua inglesa.

1. Introdução

Há mais de um século que o termo *fake news* é usado para qualificar notícias ou rumores como falsos. Em outubro de 1925 a revista Harper publicou um artigo [McKernon, E. 1925] com o seguinte tema ``Fake news and the public: How the press combats rumor, the market rigger, and the propagandist''. Nesse artigo, o autor discute o papel da imprensa na disseminação de notícias falsas e incorretas. Hoje em dia, a forma como notícias falsas são propagadas nas mídias sociais se tornou um problema de ordem mundial. De acordo com pesquisa descrita em [Reuters Institute. Digital News Report 2018], quando se pensa em notícias on-line, mais da metade da amostra (54%) expressa preocupação média ou forte sobre o que é verdadeiro ou não. Existem variações significativas entre os países. No Brasil esse percentual chega a 85%, enquanto que na Espanha, Estados Unidos e França, em média, 65% dos entrevistados se preocupam com a veracidade de notícias. Um importante fato é que nestes países eleições recentes ou campanhas de referendos foram afetadas por desinformação ou má informação. Por outro lado, tem-se países onde a preocupação parece ser menor, como na Alemanha (37%) e na Holanda (30%). Nestes países, a política tende a ser menos

polarizada e as mídias sociais tem um papel de menor importância como fonte de notícias. Em outra pesquisa, [Grinberg et al. 2019] examinou a exposição e o compartilhamento de *fake news* por eleitores americanos registrados no Twitter e encontrou que o engajamento foi bastante concentrado durante o período das eleições presidenciais de 2016.

O uso do termo “*fake news*” tem sido criticado por causa do seu uso político (figuras públicas usam para desacreditar notícias que não lhe favorecem [S. Vosoughi, D. Roy, S. Aral 2018]), e por não descrever a complexidade dos diferentes tipos de má informação (*misinformation*) ou desinformação (*disinformation*). Má informação é a notícia compartilhada de forma equivocada que se acredita estar correta, e Desinformação é a criação deliberada de informação falsa [Wardle, C. 2017]. [Lazer et al. 2018] defende que o uso do termo *fake news* possui um efeito benéfico, pois chama a atenção para o problema mais geral que é o compartilhamento disseminado de notícias falsas. [Zubiaga et al. 2017] definem rumor como “a circulação de uma história de veracidade questionável, que aparentemente é verdadeira, mas que possui difícil verificação e ainda produz ceticismo ou ansiedade suficientes para motivar a busca pela verdade dos fatos”. Neste trabalho usamos de forma intercambiável os termos rumor e *fake news*. Conforme [Dale, 2017], vivemos no mundo da pós-verdade. Hoje as pessoas se importam mais com a sua opinião sobre um assunto do que com a verdade dos fatos. Rumores são compartilhados o tempo todo e afetam a percepção das pessoas e o seu comportamento. Abordagens automatizadas oferecem um potencial para lidar com esse número cada vez maior de informações incorretas, entretanto ainda estão em estágio de desenvolvimento. A verdade factual é, na maioria das vezes, estabelecida manualmente por jornalistas ou especialistas que pesquisam afirmações e buscam fontes confiáveis de evidências.

Desde o trabalho seminal de [Qazvinian et al. 2011], várias tarefas compartilhadas foram propostas a fim melhorar o desempenho dos sistemas para a avaliação dos rumores (RumourEval 2017 [10], RumourEval 2019¹, FEVER2). Em todas estas competições, percebe-se o foco em notícias falsas disseminadas via Twitter e em língua inglesa. Twitter é a terceira rede social mais usada para compartilhar notícias [Reuters Institute. Digital News Report 2018], e tem a como característica o serviço de *microblogging para compartilhamento* de textos curtos. Atualmente, o estado da arte da tarefa de verificação de rumores em microtextos do Twitter, em língua inglesa, é F1-Score=0,53, em 2017, e F1-Score=0,57, do RumourEval 2019.

Para a língua portuguesa, o primeiro trabalho a criar um corpus para análise de notícias falsas foi o de [Monteiro et al. 2018], que propuseram o corpus Fake.Br, contendo 7200 notícias de blogs e jornais. Importante notar que este primeiro corpus para a tarefa de verificação de rumores, em português, consiste de textos longos. O trabalho de Monteiro, usa o corpus Fake.Br junto com uma abordagem tradicional de Machine Learning alcançando um F1-Score = 0.89.

Percebe-se, portanto, uma diferença significativa entre o desempenho do melhor sistema para o inglês (F1-Score = 0,57) e do melhor sistema para o português (F1-Score

¹ <https://competitions.codalab.org/competitions/19938>

² <http://fever.ai>

= 0.89). Nossa argumento é que esta diferença se deve à natureza distinta dos *corpora* usados como *benchmarking*. Como dito, o foco das competições internacionais é em notícias falsas compartilhadas pelo Twitter, e o único corpus de *fake news* em português consiste de notícias longas publicadas em jornais e blogs. Não se tem notícia de corpus de *tweets* em português para a tarefa de verificação de rumores.

Neste trabalho, propomos o corpus FakeTweet.Br, construído de acordo com metodologia de corpus similares em língua inglesa, contém *tweets*, em português, divididos entre notícias falsas e verdadeiras, de 24 assuntos distintos. Aplicamos ainda abordagens tradicionais de *Machine Learning* na tarefa de verificação de rumores, com o objetivo de avaliar o desempenho da tarefa para língua portuguesa. Os resultados mostraram que classificadores tradicionais como o *Naive Bayes* e *Gradient Descent*, usando o corpus FakeTweet.Br, suplantaram sistemas do estado da arte para língua inglesa para corpus de tamanhos similares.

2. Trabalhos relacionados

Para língua inglesa, a tarefa de verificação de rumores foi proposta pela primeira vez no RumorEval 2017, como parte da conferência SemEval 2017, e foi reeditada em 2019. Em 2017, o sistema IKM [Chen et. al. 2017] foi o ganhador com F1-Score = 0,53 e usou um modelo de Rede Neural Convolucionar com pré-treinamento usando GloVe³. Em 2019, o sistema ganhador obteve F1-Score=0.89. As competições RumourEval 2017 e RumourEval 2019 propuseram dois conjuntos de dados para benchmarking contendo *threads* do Twitter - um post original e os comentários destes posts realizados por usuários. A Tabela 1 apresenta a estatística dos conjuntos de dados nas classes “Falso” (tweet com notícia falsa), “Verdadeiro” (tweet com notícia verdadeira), e “Não-verificável” (tweet com notícia cuja veracidade não pode ser verificada).

Tabela 1 Estatística dos Conjuntos de Dados para a tarefa de Verificação de Rumores, propostos em 2017 e 2019.

Conjunto de dados de 2017		Conjunto de dados de 2019		% incremento 2019/2017
Classe de Rumores	# exemplos	Classe de Rumores	# exemplos	% incremento 2019/2017
Falso	50	Falso	79	58.00%
Verdadeiro	127	Verdadeiro	144	13.39%
Não Verificável	95	Não Verificável	103	8.42%
	272		326	19.85%

Para língua portuguesa, o primeiro trabalho a criar um corpus para análise de notícias falsas foi descrito em [Monteiro et al. 2018]. Neste trabalho, foi proposto o corpus balanceado Fake.Br que contempla 7200 notícias, classificadas como falsas ou verdadeiras. As notícias falsas foram selecionadas dos sites Diário do Brasil, A Folha do Brasil, The Jornal Brasil e Top Five TV. As notícias verdadeiras foram recuperadas dos seguintes portais e jornais na Web: G1, Folha de São Paulo e Estadão. O corpus Fake.Br se propõe a ser um conjunto de dados de referência para testes de algoritmos

³ <https://nlp.stanford.edu/projects/glove/>

para identificação de fake News, em artigos publicados na web através de portais, blogs e veículos de comunicação on-line.

O trabalho de [Monteiro et al. 2018] apresenta uma série de experimentos no corpus Fake.Br com o objetivo de analisar quais atributos poderiam ser utilizados para classificação de notícias verdadeiras ou falsas. O melhor resultado ($F1\text{-Score} = 0.89$) foi apresentado pelo classificador Linear SVM com os atributos Bag of Words (BOW) e *Emotiveness*, a qual expressa a intenção do autor em provocar emoção no leitor. No entanto, a diferença para configuração que usou apenas os atributos BOW ($F1\text{-Score} = 0.88$) não é significativa. Observa-se uma diferença significativa entre os resultados dos sistemas do estado da arte para a língua inglesa e portuguesa. Nossa hipótese é que textos de *microblogs* como o Twitter, a terceira maior rede social usada para compartilhamento de notícias, são mais difíceis de serem verificados, pelo tamanho reduzido e linguagem específica.

3. Corpus FakeTweet.Br

Com o objetivo de criar um corpus de tweets (microtextos), em português do Brasil, foram coletados tweets de perfis notadamente de boa reputação na tarefa de verificação de notícias: @agencialupa (123.000 seguidores, 10.600 tweets e está no Twitter desde 2015), @aosfatos (176.000 seguidores, 5.289 tweets e está no Twitter desde 2015) e @boatos.org (17.700 seguidores, 4.376 tweets e está no twitter desde 2013). Em seguida, foram recuperados os posts originais que divulgavam a notícia considerada falsa.

Uma premissa importante é que os tweets pudessem ser rastreados por seu respectivo ID e que a possível notícia falsa tivesse sido confirmada por pelo menos duas dentre as três agências de checagem de fatos. Um dificuldade inicial foi que não poderíamos somente coletar o post da agencia de checagem de noticias, pois os mesmos não trazem em seu texto a noticia original, ao contrário, trazem expressões que já afirmam a falsidade da noticia, tais como “É Falso que...” e “Ao contrário do que aponta a mensagem ...”. Portanto, foi necessário desenvolver um Twitter’s crawler para minerar e recuperar os tweets com os textos originais das notícias falseadas por duas ou mais agencias. A Figura 1 apresenta um exemplo de tweet do perfil @boatos.org, que afirma a falsidade de uma notícia; a referida notícia falsa, originalmente postada por um usuário, e outra notícia sobre o mesmo fato, considerada verdadeira.

Tweet publicado no perfil @boatos.org

Ao contrário do que aponta a mensagem, MST não foi ocupar "terras prometidas" pela Vale às vítimas. Entenda o caso

Postagem classificada como falsa

Graças a Deus as coisas estão entrando nos trilhos!! O medo de se deparar com um agricultor armado é maior que a desejo de saquear. Parabéns presidente. Agora o MST está perseguindo as famílias das vítimas em Brumadinho, parasitas.

Postagem classificada como verdadeira

Deputado repercute notícia falsa sobre o MST ter invadido Brumadinho para tomar terras de atingidos pela barragem", o que já foi desmentido

Figura 1 Exemplo de tweet de uma agência de checagem; a respectiva notícia original; e uma notícia relacionada, considerada verdadeira.

A metodologia de criação do corpus FakeTweet.Br consiste nas seguintes etapas:

1) Seleção de posts das agências de checagem de notícias (@agencialupa, @aosfatos e @boatos.org) – nesta etapa são recuperados os tweets que afirmam que uma certa notícia é falsa e identificado se a mesma notícia é considerada falsa por mais de uma das agências. Importante notar que durante a pesquisa não foi encontrada divergência entre as classificações das agências, apenas notícias que eram reportadas apenas por uma delas.

2) Desenvolvimento de Twitter's Crawler para recuperação dos tweets – para recuperar as mensagens originais de notícias consideradas falsas (tweets da etapa anterior), não seria eficaz usar a API original do Twitter, visto que a mesma limita o período de dias para recuperação dos posts e, além disso, não existiria garantia de que a mensagem original da notícia falsa seria retornada pela API. Uma ideia foi desenvolver um Twitter Crawler capaz de recuperar posts que mencionam determinadas palavras-chave. Para isso, foram identificados os assuntos mencionados nos tweets selecionados na etapa anterior e definido, para cada um deles, um conjunto de palavras-chave que servisse como critério de busca. A Tabela 3 apresenta exemplos dos 24 assuntos mencionados nos posts das agências de checagem e as palavras-chaves definidas para cada assunto. Por exemplo, para o assunto “Gleise Hoffman falou em entrevista que o piloto brasileiro Ayrton Senna não foi um brasileiro importante.”, foram definidas as palavras-chave “Gleisi” e “Senna”. Como parâmetro, tem-se um limite de 1000 tweets a serem recuperados para cada critério de busca (palavras-chave).

3) Anotação dos tweets originais contendo mensagens falsas ou verdadeiras – esta etapa é a mais dispendiosa em termos de tempo, pois cada tweet recuperado na etapa anterior (em torno de 20000 tweets) foi manualmente analisado e rotulado como “Falso”, caso o texto do post relatassem informações consideradas falsas (pela agência de checagem – etapa 1); ou “Verdadeiro”, caso o texto mencionasse o assunto, mas não divulgasse informação falsa. As principais dificuldades do processo de anotação deveram-se ao grande número de posts que mencionavam as palavras-chave, mas não tratavam o assunto em foco, a posts que apenas questionavam a veracidade do assunto, e, finalmente, a posts ou estavam escritos em outro idioma. Por exemplo, o tweet “*Mujer asegura que se curó del cáncer gracias al Aranto*” foi recuperado pelas palavras-chave “aranto” e “câncer”, relativas ao assunto “A planta brasileira aranto cura câncer”, porém estava escrito no idioma espanhol. A Tabela 3 apresenta, nos exemplos selecionados, a quantidade de posts rotulados como falsos e verdadeiros, em cada assunto. Importante notar que o número de tweets que relatam o que de fato ocorreu ou relatam fatos verdadeiros sobre o assunto (anotados como verdadeiros) é bem menor que o número de tweets que divulgam informações falsas. Pelo volume de tweets a serem analisados, o processo de anotação foi realizado por única pessoa. Ao final, o conjunto de tweets anotados (no total de 279) formam o conjunto de treinamento do corpus FakeTweet.Br.

4) Montagem do conjunto de teste do corpus FakeTweet.Br – visando definir um conjunto de tweets para testar sistemas de verificação de rumor, coletamos um conjunto reduzido de posts de assuntos diferentes dos assuntos do conjunto de treinamento. Para isso, realizamos as etapas 1 a 3 para outros assuntos, os quais são apresentados na Tabela IV.

A Tabela 2 apresenta um extrato dos assuntos do corpus FakeTweet.Br e a distribuição de tweets falsos e verdadeiros em cada assunto. No total, o conjunto de treinamento do corpus é composto por 279 tweets distribuídos em 24 assuntos. É bem notória a predominância de assuntos ligados a política, o que é esperado dado o momento polarizado em que estamos vivendo no Brasil e em vários países do mundo. O conjunto de teste é composto de 20 tweets distribuídos em 7 assuntos. Neste caso, os assuntos estão平衡ados entre assuntos relacionados à política e assuntos gerais.

Tabela 2 Exemplos de Assuntos e respectivas Palavras-chave abordados no corpus FakeTweet Br – Conjunto de Treinamento e Teste

Corpus FakeTweet Br (conjunto de treinamento)			
Assunto	Palavras chaves	#Falsos	# Verdadeiros
MST invadindo cidade de Brumadinho	mst brumadinho	35	14
O verdadeiro assassino da vereadora Marielle é um negro apelidado de macaco	macaco marielle	26	9
Cantor John Hocker que havia chamado Jesus de Travesti é contratado para se apresentar no programa Criança Esperança	travesti criança esperança	27	
Israel doou helicópteros para o Governo Brasileiro	israel helicopteros	13	6
A planta brasileira aranto cura câncer	aranto câncer	14	3
Lula está listado na forbes como um dos políticos mais ricos do mundo	lula forbes	12	8
O presidente Jair Bolsonaro foi eleito a personalidade do ano pela revista norte americana TIME	jair time personalidade	3	5
Ministra damares revoga lei maria da penha	Lei maria da penha damares	6	1
Marina Silva impede que radares sejam desativados pelo governo federal	Marina silva radares	4	1
Corpus FakeTweet Br (conjunto de teste)			
Assunto	Palavras chaves	#Falsos	# Verdadeiros
Comercial do banco do brasil vetado pelo Presidente Jair Bolsonaro custou 17 milhões de reais	banco brasil vetado bolsonaro	3	1
Pintor Gustave Fraipoint fez uma pintura chamada Premonição que previa o incêndio da Catedral de Notre Dame	gustave notre dame	1	2

A Tabela 3 apresenta uma visão geral do corpus FakeTweet.Br. Importante salientar que o tamanho do corpus é compatível com *corpora* similares em língua inglesa (vide Tabela 1). O corpus FakeTweet.Br se propõe a ser um corpus de referência para a tarefa de verificação de rumores disseminados em língua portuguesa.

Tabela 3 Estatística do Corpus FakeTweet.Br em português

Corpus FakeTweet br			
Conjunto de Treinamento		Conjunto de Teste	
Classificação da Notícia	# exemplos	Classificação da Notícia	# exemplos
Falsa	194	Falsa	12
Verdadeira	85	Verdadeira	8
Qtd Assuntos diferentes - 24	279	Qtd Assuntos diferentes - 7	20

A Figura 2 apresenta como o corpus FakeTweet.Br está formalizado. Cada tweet é identificado por seu ID, palavras-chave, texto original, classificação (falso ou verdadeiro), data da publicação, número de retweets, número de marcações como favorito, e o link para mensagem original.

ID: 1112860378936100000	Palavras Chaves: stf lula salvo conduto
Texto: @CarlosBolsonaro É URGENTE que antecipemos a ida às ruas antes do dia 7 de abril ! O STF resolveu dar salvo conduto a Lula no dia 04 de abril.E estará SOLTO Vamos dia 3 de abril !!!!!	
Classificação: Fake	Data Publicação: 2019-04-01 20:32
Retweets: 0	Favoritos: 0
Link Mensagem Original:	
https://twitter.com/Jane_GRodrigues/status/1112860378936107008	

Figura 2 Estrutura lógica do corpus FakeTweet.Br

4. Experimentos e Resultados

Inspirados no trabalho de [Monteiro et al., 2018] para a língua portuguesa, nós definimos os seguintes cenários de aprendizagem para a tarefa de verificação de rumores, usando o conjunto de treinamento do corpus de tweets FakeTweet.Br.

- CENÁRIO 1 – foram usados vários algoritmos ML (Logistic Regression, Stochastic Gradient Descent, Complement Naive Bayes, Random Forest e Linear Support Vector), com os parâmetros padrões do pacote SktLearn. e com os atributos BOW-unígrama e bigrama. Neste cenário, o conjunto de treinamento foi usado desbalanceado.
- CENÁRIO 2 – mesmos algoritmos e atributos do CENARIO 1, mas, devido ao desequilíbrio das classes, neste cenário as mesmas foram balanceadas usando o algoritmo SMOTE [Chawla, N.V., Bowyer, K. SMOTE 2002].

Tabela 4 Resultados dos Cenários 1 e 2 - fase de Treinamento (sem e com balanceamento usando SMOTE)

Algoritmo	Corpus FakeTweet.Br									
	F1 Médio		F1 Falsa		F1 Verdade		Precisão Média		Cobertura Média	
	sem bal	com bal	sem bal	com bal	sem bal	com bal	sem bal	com bal	sem bal	com bal
Logistic Regression	0.81	0.83	0.89	0.83	0.73	0.84	0.87	0.84	0.78	0.84
Stochastic Gradient Descent	0.80	0.86	0.86	0.86	0.73	0.86	0.80	0.86	0.79	0.86
Complement Naive Bayes	0.77	0.78	0.81	0.76	0.74	0.79	0.78	0.78	0.81	0.78
Random forest	0.66	0.76	0.84	0.74	0.49	0.79	0.85	0.78	0.66	0.77
Linear Support Vector	0.81	0.81	0.89	0.79	0.73	0.82	0.86	0.82	0.79	0.81

Em todos os cenários da fase de treinamento, foi usada a validação cruzada com k=5. A Tabela 4 apresenta os resultados desta fase, onde o melhor resultado foi

alcançado pelo algoritmo Stochastic Gradient Descent, em ambos os cenários. O CENARIO 2 (com balanceamento) apresentou F1-Score = 0.86. Na fase de teste, foi usado o conjunto de teste do corpus FakeTweet.Br com os mesmos algoritmos do CENARIO 1, e os resultados estão relatados na Tabela 5. Para o conjunto de teste, os melhores algoritmos foram o *Complement Naive Bayes* e o *Stochastic Gradient Descent* com F1-Score=0.74 e F1-Score=0.73 respectivamente, os quais suplantam o estado da arte para língua inglesa (F1-Score = 0.57).

Tabela 5 Resultados dos testes de algoritmos de classificação - (com balanceamento usando SMOTE)

Corpus FakeTweet.Br					
Algoritmo	F1 Médio	F1 Falsa	F1 Verdade	Precisão Média	Cobertura Média
Logistic Regression	0.41	0.26	0.56	0.53	0.52
Stochastic Gradient Descent	0.73	0.80	0.66	0.74	0.72
Complement Naive Bayes	0.74	0.78	0.70	0.74	0.75
Random forest	0.50	0.50	0.50	0.52	0.52
Linear Support Vector	0.47	0.37	0.58	0.59	0.56

5. Conclusão

Este artigo propõe um novo corpus de referência – FakeTweet.Br - para a tarefa de verificação de rumores, disseminados via microblogs como Twitter, em língua portuguesa. O corpus proposto foi construído seguindo a metodologia de corpus similares e é compatível em tamanho com corpus da língua inglesa, propostos em competições internacionais (RumourEval 2017 e 2019). A diversidade de assuntos do *corpus*, aqui proposto, é uma característica desejada em um trabalho de referência. Avaliamos o desempenho da tarefa de verificação de rumores disseminados via Twitter, em língua portuguesa, em vários classificadores tradicionais, usando o corpus FakeTweet.Br. Os resultados dos testes indicaram que o melhor algoritmo foi o Complement Naive Bayes, com F1-Score=0.74, que suplanta o estado da arte para língua inglesa. Apesar do resultado ser bem interessante, temos ciência de que os *corpus* são diferentes e portanto a comparação é limitada. Uma hipótese para essa diferença está exatamente no fato de estarmos lidando com um corpus relativamente pequeno, e para os trabalhos em língua inglesa a técnica mais abordada é a utilização de redes neurais, que necessitam de corpus maiores para ter um bom desempenho.

Uma análise da diferença do desempenho da tarefa para textos longos e curtos indica que é mais difícil verificar rumores divulgados em textos curtos, pois para textos de notícias longas (jornais e blogs) mais elementos textuais podem ser aprendidos e a linguagem é mais formal. Como trabalhos futuros, acredita-se que para microtextos seja importante a coleta de informações contextuais como perfis dos usuários, quantidade de seguidores, tempo em que a conta está ativa, perfil das postagens anteriores, e conhecimento de senso comum. Outra tendência é investigar o impacto de informações sobre a reação de usuários às notícias falsas, especificamente, investigar o uso de sistemas de classificação de postura (*stance classification*), expressas nos comentários e respostas dos posts originais.

References

- McKernon, E. Fake News and the Public: How the Press Combats Rumor, The Market Rigger, and The Propagandist. Harper's Monthly (1925)
- Reuters Institute. Digital News Report 2018. at: < <http://www.digitalnewsreport.org> > Access in 28 april 2019
- Grinberg, N., Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. Science 25 Jan 2019: Vol. 363, Issue 6425, pp. 374-378 DOI: 10.1126/science.aau2706
- S. Vosoughi, D. Roy, S. Aral, The Spread of True and False News Online. Science 359, 1146–1151 (2018).
- Wardle, C. Fake News. It's Complicated. (2017), (available at <https://firstdraftnews.com/fake-news-complicated/>)
- Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, Jonathan L. Zittrain. The science of fake news. Science 09 Mar 2018: Vol. 359, Issue 6380, pp. 1094-1096. DOI: 10.1126/science.aoa2998
- Zubiaga, Arkaitz, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. In Proceedings of the AAAI Workshop on AI for Cities.
- Dale, Robert. NLP in a post-truth world. February 2017. Natural Language Engineering 23(02):319-324. DOI: 10.1017/S1351324917000018
- Qazvinian, V., Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 1589–1599.
- Derczynski, L., Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 69–76. <http://www.aclweb.org/anthology/S17-2006>.
- Monteiro R.A., Santos R.L.S., Pardo T.A.S., de Almeida T.A., Ruiz E.E.S., Vale O.A. (2018) Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In: Villavicencio A. et al. (eds) Computational Processing of the Portuguese Language. PROPOR 2018. Lecture Notes in Computer Science, vol 11122. Springer, Cham
- Chen, Yi-Chin & Liu, Zhao-Yang & Kao, Hung-Yu. (2017). IKM at SemEval-2017 Task 8: Convolutional Neural Networks for stance detection and rumor verification. 465-469. 10.18653/v1/S17-2081.

Chawla, N.V., Bowyer, K. SMOTE: Synthetic Minority Over-sampling Technique.
Article in Journal of Artificial Intelligence Research · January 2002 DOI:
[10.1613/jair.953](https://doi.org/10.1613/jair.953)

New developments on processing European Portuguese verbal idioms

Ana Galvão^{1,3}, Jorge Baptista^{2,3}, Nuno J. Mamede^{1,3}

¹Instituto Superior Técnico, Universidade de Lisboa
Av. Rovisco Pais 1, P-1049-001 Lisboa, Portugal

²Universidade do Algarve - Faculdade de Ciências Humanas e Sociais
Campus de Gambelas, P-2005-139 Faro, Portugal

³Instituto de Engenharia de Sistemas e Computação - INESC-ID Lisboa
L2F - Spoken Language Laboratory
R. Alves Redol 8, P-1000-029 Lisboa, Portugal

{a.s.galvao, Nuno.Mamede}@tecnico.ulisboa.pt, jbaptis@ualg.pt

Abstract. This paper presents recent developments in processing verbal idioms within a rule-based grammar of European Portuguese. It describes the automatic construction of parsing rules directly from a lexicon-grammar matrix with about 2,500 idioms and about 100 structural, distributional, and transformational properties. Transformations (passive, pronominalization, etc.) of idioms' base sentences are now taken into account within the automatic rule generation process. An intrinsic evaluation achieves 95% recall.

1. Introduction

Verbal idioms are a type of frozen sentences [Gross 1982, Gross 1996], where the verb and at least one of its arguments (subject or complement) are distributionally frozen together; and the global interpretation of the sentence is non-compositional, that is, it can not be calculated from the meaning the components of the idiom when they are used independently, e.g. *O Rui agarrou o touro pelos cornos* , lit.: ‘Rui took/grasped the bull by the horns’. ‘to take definite and determined action in order to deal with a difficult situation’. Parsing multiword expressions (MWE) such as verbal idioms is a challenging task for many Natural Language Processing (NLP) systems [Sag et al. 2002, Constant et al. 2017], since, for the most part, they have an internal structure identical to that of ordinary sentences, including one or more distributionally free arguments, and can undergo several, very general transformations, such as pronominalisation, passive, nominalisation, etc. Taking MWE into consideration, especially sub-sentential lexical units, can significantly improve the quality of several NLP tasks, like part-of-speech tagging [Constant and Sigogne 2011] or parsing [Constant et al. 2017]. Obviously, identifying MWE may lead to a more adequate representation of the meaning of a text. Most previous work deals with the identification of idioms and other MWE in texts [Ramisch et al. 2018, Ramisch et al. (eds.) 2018], since the low frequency of many verbal idioms in corpora makes spotting them a difficult task in lexicographic studies [Manning 1999, Pecina 2010]. The focus of this paper, however, will be on processing verbal idioms once they have already been integrated in a computational lexicon.

This paper presents new developments in the processing of European Portuguese (EP) verbal idioms, within the framework of a pipeline NLP system, STRING¹ [Mamede et al. 2012]. The paper’s main contribution is the processing of the most common syntactic transformations (passive, pronominalization, *etc.*) accepted by these idioms. The paper is organized as follows: First, the lexicon-grammar of EP verbal idioms is presented, along with the parsing strategy adopted in STRING. Next, the parsing of transformations is outlined. The paper, then, reports the results obtained in an *intrinsic* evaluation of the system, and concludes by pointing the challenges ahead.

2. Processing verbal idioms: current state

Previous work on European Portuguese verbal idioms [Baptista et al. 2004, Baptista et al. 2014, Baptista et al. 2016] done within Lexicon-Grammar framework [Gross 1982, Gross 1996], produced a lexicon-grammar matrix, currently with around 2,500 entries (one verbal idiom per line) along with the corresponding linguistic description. This description uses about 100 features (columns) to account for the structural, distributional and transformational properties of the idioms. The idioms are organized into 13 major classes (Table 1), according to the number of arguments selected by the verb in the frozen construction, and which arguments are distributionally free or frozen with the verb (the subject or one or more complements). For lack of space, the classification procedure is not provided in further detail here (see [Baptista et al. 2016, Galvão 2019]). The verb and the frozen elements of the idiom are explicitly encoded: the complements’ preposition Prep (if any), the determiner Det, the frozen head noun C, and its left or right modifier Modif). The human/non-human nature of distributionally free arguments is represented by binary features (‘+/-’), as well as several transformational properties. The transformations considered so far are all very general: passive, pronominalisation, symmetry, and dative restructuring (see below). Finally, all entries are illustrated by a manually produced example. These examples, while being perfectly natural and acceptable utterances, are ‘artificial’, almost ‘laboratorial’, since they contain all essential arguments (subject and complements) of the verb, and have been stripped of any spurious lexical material not relevant for the interpretation of the idiom. Furthermore, whenever necessary, the verb is provided in an non-ambiguous inflected form, in order to prevent errors in previous stages of the processing, particularly in PoS tagging and disambiguation. These examples are also used for the intrinsic evaluation of the system.

Table 1. Lexicon-Grammar of European Portuguese verbal idioms.

Class	Structure	Example	Translation/gloss	Count	%
C0	C0 V w	O azar bateu à porta do Rui	Bad luck knocked on Rui’s door (have bad luck)	25	0.010
C1	N0 V C1	O Rui bateu a bota	Rui kicked the boot (died)	506	0.198
C1P2	N0 V C1 Prep2 C2	O Rui comeu gato por lebre	Rui ate cat for hare (was cheated)	284	0.111
C1PN	N0 V C1 Prep2 N2	O Rui acertou agulhas com o Pedro	Rui matched needles with Pedro (are in accord)	255	0.100
CADV	N0 V ADV	O Rui vai longe	Rui will go far (will be successful)	70	0.027
CAN	N0 V (C de N)1 = C1 a N2	O Rui partiu os olhos de/a/a Ana	Rui open the eyes of/to Ana (make understand)	182	0.071
CDN	N0 V (C de N)1	O Rui veste a camisola da empresa	Rui dones the t-shirt of the company (dedicate/loyal)	47	0.018
CNP2	N0 V NI Prep2 C2	O Rui conhece a Ana de nome	Rui knows Ana by name (id)	175	0.068
CPI	N0 V Prep1 C1	O Rui foi aos arames	Rui went to the strings (be mad)	598	0.233
CPN	N0 V Prep1 (N de C)1	O Rui foi na cantiga do Pedro	Rui went in Pedro’s song (be dupped)	103	0.040
CPIP2	N0 V Prep1 C1 Prep2 C2	O Rui foi destu para melhor	Rui went from this [one] to a better [one] (die)	170	0.066
CPP	N0 V Prep1 C1 Prep2 N2	O Rui foi de Caifás para Pilatos	Rui went from Caifas to Pilates (get in a worst situation)	77	0.030
CPPN	N0 V C1 Prep1 C2 Prep3 C3	Isso deu água pela barba à Ana	That gave water by the beard to Ana (very complicated)	51	0.020
CPPP	N0 V Prep1 C1 Prep2 C2 Prep3 C3	O Rui contava com o ovo no cu da galinha	Rui counted with the egg in the chicken’s ass (be too confident)	5	0.002
CV	N0 V Inf w	Esta rua vai dar à praça	This street goes give to the square (lead to)	13	0.005
				Total	2,562

¹<https://string.12f.inesc-id.pt/> (last access: 06/08/2019)

Since the construction of the lexicon-grammar matrix is not only a complex process but it is also carried out manually, it is thus a very error-prone task. To reduce the human error in this process, an *Automatic Validator* has been built, written in Perl, to check the formal consistency of the matrix. This validator takes as input the CSV-converted lexicon-grammar matrix and performs the following checks, outputting the corresponding error messages: (i) *cell* content validation: checks if the content of the cell in a given column is consistent with the predefined values for that column; (ii) *class* consistency cross-validation: depending on the class of the idiom, the number of relevant columns and the values therein can vary; and (iii) related properties cross-validation: consistency among related properties, represented in different columns, is checked. The validator resorts to a set of several dozens of manually crafted rules. Based on the error messages outputted by the validator, it is possible to detect most input errors, which are then manually corrected. When no formal errors are found, the matrix is ready to be processed.

The processing of verbal idioms is done within the framework of the rule-based, Xerox Incremental Parser (XIP) [Ait-Mokhtar et al. 2002], which is the parsing module of the NLP system pipeline STRING [Mamede et al. 2012], developed for Portuguese. This system performs all the basic text processing tasks: (i) text segmentation and tokenisation; (ii) part-of-speech (PoS) tagging; (iii) rule-based and statistical PoS disambiguation; and (iv) parsing. The later includes both *chunking* and dependency parsing. The first forms the elementary constituents (e.g. *chunks*: noun phrase, NP; prepositional phrase, PP; *etc.*). The second extracts the relations between the chunks' heads, e.g. subject (SUBJ). STRING also performs other, common NLP tasks, such as named entity recognition (identification and classification, NER), anaphora resolution, event detection, among others. Fig. 1 illustrates the parse tree of the example above, with the chunks and some dependencies calculated by the parser.

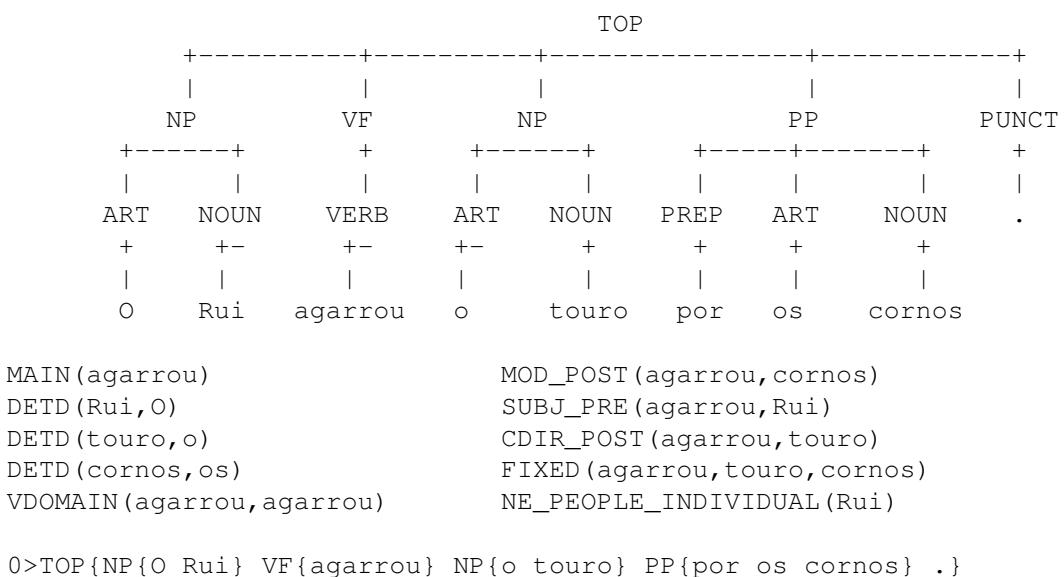


Figure 1. Parse tree of a verbal idiom.

Figure 1 shows some auxiliary dependencies, such as the determiner DETD, linking the articles to the head nouns of the NP and PP chunks. There is also a PREPD dependency (not shown) linking the preposition *por* 'by' to the head noun of the PP. The

VDOMAIN dependency links the first verb of a chain of auxiliary verbs to the main verb [Baptista et al. 2010]. The main dependencies, v.g. subject (SUBJ), modifier (MOD), and direct complement (CDIR), are also calculated, linking the verb to its arguments. The named entity *Rui* is also captured by the unary dependency NE, which takes the features _PEOPLE and _INDIVIDUAL corresponding to the entity type [Hagège et al. 2008].

Since, for the most part, verbal idioms comply with the general rules governing the structure of well-formed sentences in the language, including word order, it has been deemed more appropriate [Rassi et al. 2014, Baptista et al. 2014] to detect them only at a later stage of the parsing process, after the main syntactic dependencies between the sentences's constituents have been calculated. These dependencies are then used to detect idioms. If an idiom is detected, the system produces a FIXED dependency, whose arguments are the the main verb and its frozen arguments. The following rule, shown in Fig. 2), was fired and captured the idiom of this example:

```
if ( VDOMAIN (#?, #2[lemma:agarrar]) &
    CDIR[post] (#2, #3[surface:touro]) &
    DETD (#3, ?[surface:o]) &
    MOD[post] (#2, #4[surface:cornos]) &
    PREPD (#4, ?[surface:por]) &
    DETD (#4, ?[surface:os])
)
FIXED (#2, #3, #4)
```

Figure 2. Parsing rule for the verbal idiom *agarrar o touro pelos cornos*
‘take/grab the bull by the horns’.

This parsing rule has two parts: the first is a `if ()` structure of conditions that must be satisfied so that the consequent of the rule is triggered; the consequent writes the FIXED dependency and its arguments. The VDOMAIN is used to capture the main verb, even if this is construed with a chain of auxiliary verbs. The next dependencies verify if the heads of direct complement CDIR and the prepositional phrase modifier MOD are the same as encoded in the lexicon-grammar matrix. The same applies to the auxiliary dependencies DETD within the complements and the PREPD dependency for the MOD. All conditions are joined by ‘&’ (disjunction ‘||’ may also be used) The variables (signaled by '#') are then used to produce the FIXED dependency.

The process of automatically generating the parsing rules directly form the linguistic information encoded in the matrix is quite complex, so it will only be sketched here. First, each column of the matriz is associated with a XIP rule. Each relevant column value contributes with a condition to that rule (relevant columns depend on the verbal idiom class); and each main constituent’s head is associated to a variable (by convention, the subject is associated with variable #1, the verb to #2, and so forth). The system systematically explores the properties encoded in all columns of the matrix, adding the corresponding conditions to the `if ()` structure of the parsing rule. Finally, it writes the FIXED dependency with its arguments. The output of this module consists of the parsing rules, the corresponding manually produced example, and the expected output of the FIXED dependency.

Already at a first attempt to integrate verbal idioms [Baptista et al. 2016] into the STRING, a special module had been built to automatically generate the parsing rules for

the detection of idioms, based on the information directly extracted from the lexicon-grammar matrix. However, the method proved to be very rigid, as it depended on the column number and order. Also, only the passive transformation was considered. Furthermore, the evaluation was carried out sentence by sentence, and each time the system had to be initialised. This took too much time to be practical. Besides, this evaluation only reported whether the **FIXED** dependency had been extracted or not, thus providing limited feedback for the further development of STRING. The new developments, presented below, address all these issues.

3. Processing verbal idioms: new developments

Several improvements were introduced in the processing of idioms within STRING. These new developments are presented next. Foremost, the system is now able to process several sentence transformations: Different types of *pronominalisation* of the distributionally free complements are considered, named after the case/type of the pronoun involved: (i) *accusative* [PronA]: *O Pedro lançou o João às feras* (class CNP2) ‘Pedro threw João to the wolves’ = *O Pedro lançou-o às feras* ‘Pedro threw him to the wolves’; (ii) *dative* [PronR]: *O Pedro lançou a escada à Ana* (CNP2) lit.: ‘Pedro threw the stairs to Ana’ ‘Pedro tried to seduce Ana’ = *O Pedro lançou-lhe a escada* lit.: ‘Pedro threw to-her the stairs’; (iii) *reflexive* [PronR]: *O Pedro reduziu o João ao silêncio* (CNP2) lit.: ‘Pedro reduced João to silence’ cp. *O Pedro reduziu-se ao silêncio* lit.: ‘Pedro reduced himself to silence’; (iv) *possessive* [PronPos]: *O Pedro abriu os horizontes da Ana* (CAN) lit: ‘Pedro opened the horizons of Ana’ = *O Pedro abriu os seus horizontes* ‘Pedro open her horizons’. Two types of passive sentences, with different auxiliaries: (v) *Passive* with auxiliary *ser* ‘be’; in this type of passive, the subject of the active sentence becomes a prepositional complement *por N* ‘by N’; only the verb *ser* ‘be’ is considered in this case: *O João foi reduzido ao silêncio pelo Pedro* (CNP2) ‘João was reduced to silence by Pedro’; (vi) *Passive* with auxiliary *estar* ‘be’; in this passive, the subject of the active sentence is usually zeroed; any other copula verb, including *ficar* ‘become’, can also be captured in this rule: *O João estava/ficou reduzido ao silêncio* (CNP2) ‘João was reduced to silence’. And, finally, the dative restructuring: (vii) *dative restructuring* [Rdat] [Leclère 1995]: this type of transformation splits a complex complement ($N_a de N_b$)₁ ‘N of N’ into two constituents, (N_a)₁ ($a N_b$)₂ ‘N to N’, the noun’s complement becoming a dative (indirect) complement, more closely attached to the verb: *O Pedro abriu os horizontes da Ana* (CAN) lit: ‘Pedro opened the horizons of Ana’ = *O Pedro abriu os horizontes à Ana* ‘Pedro open the horizons to Ana’, which can now undergo the dative pronominalisation: = *O Pedro abriu-lhe os horizontes* ‘Pedro open her the horizons’. Symmetric constructions [Borillo 1971, Baptista 2005] involve the coordination of two constituents, e.g. *O Rui juntou os trapinhos com a Ana* (C1PN), lit.: ‘Rui got his rags together with Ana’ = *O Rui e a Ana juntaram os trapinhos*, lit.: ‘Rui and Ana got his rags together’, ‘Rui and Ana got married/together’. They were described by manually crafted rules , not only for the small number of symmetric verbal idioms found so far, but also for the complexity involved in capturing the coordinated arguments and the (facultative) presence of an echo complement, *um Prep outro* [Baptista and Mamede 2013].

When a verbal idiom accepts one of these transformations, the rule generator produces a disjunction ‘||’ in the `if()` structure. For example, the first example *O Pedro lançou o João às feras* ‘Pedro threw João to the wolves’ accepts both the accusative and the reflexive pronominalization of the direct complement CDIR, so this line becomes:

```
( CDIR[post] (#2,#3[UMB-Human]) ) || CLITIC(#2,?[ref]) || CLITIC(#2,#3[acc]) ) &...
```

For passive transformations, a new rule is produced because of the changes in the set of dependencies and their arguments that such structurally different sentences entail.

A *configuration file* enables the user to define which restrictions are to be applied to generate the transformation rules. Controllable restrictions apply to determinants, prepositions and both left and right modifiers of the frozen head noun; the distributional constraints to any of the free constituents, both the subject and/or the complements, can also be taken into account or ignored. In the barest configuration, only the major dependencies between the verb and the head nouns of the frozen constituents are included in the rules.

Secondly, a rule-based *Automatic Example Generator* was build from scratch, which produces a simple example for each transformation that can be applied to a given idiom, based on the linguistic information encoded in the matrix. These ‘artificial’ examples allow the linguist to better perceive the adequacy of his/her (theoretical) description, and are also used to evaluate the system. Sentences are generated along with the correct **FIXED** dependency, used for reference. For a better perception of the system’s performance, the sentences produced for each transformation are kept apart. The examples produced by this rule-based generator were manually checked by a linguist, who signalled the grammatical errors (e.g. missing contractions or prepositions) or inconsistencies produced (missing pronouns, complements); the code was, then, revised to correct those errors and a new set of sentences was generated, in a iterative way, until a ‘cleaner’ output was generated. In this process, several inconsistencies in the linguistic data could also be resolved. In all, 1,170 transformationally-derived sentences were automatically generated. For lack of space, the break down of transformations per class can not be provided here. Please refer to [Galvão 2019] for further details.

Finally, an *Evaluation Module* was build anew, which performs an intrinsic evaluation of the system. It also now takes into consideration 3 levels of granularity in the results: (i) whether the **FIXED** dependency was captured or not; (ii) if the number of its arguments is correct (**NB-ARG**); and (iii) if the arguments are the same as those in the reference (**ARG**). To this end, this module takes as input the two sets of examples, those manually produced along the verbal idioms’ entries; and those automatically generated for the transformations. The sentences are processed through **STRING** in a single batch, and the output is then compared with the reference.

In order to achieve a more comprehensive evaluation of the system, 20% of the base sentences that constitute the examples in the matrix were randomly selected from each class of idioms, totalling 511 sentences, and they were then subject to different types of modifications. These, manually produced, modifications aimed at creating incorrect/unacceptable (“*”) or non-idiomatic (“o”), but still similar, examples in order to test whether **STRING** still incorrectly extracts the **FIXED** feature in spite of them. Examples of these changes are, for the idiom *O Rui agarrou o touro pelos cornos* ‘Rui grabbed the bull thy the horns’ (C1P2): the human/non-human nature of the different distributionally free arguments; changing or zeroing the preposition (/**O Rui agarrou no touro aos cornos*), the determinant or the modifier of the frozen head noun of an argument; removing one or more frozen constituents (^*O Rui agarrou o touro*); changing the case of the pronominalized constituent (in a way unacceptable by the idiom); etc.

4. Results

**Table 2. Results for verbal idioms identification:
Manually produced sentences.**

Class	Total	#FIXED	%	#NB-ARG	%	#ARG	%
CADV	16	7	43.8	7	43.8	5	31.3
C0	21	15	71.4	15	71.4	12	57.1
C1	503	484	96.2	481	95.6	448	89.1
CAN	182	156	85.7	156	85.7	153	84.1
CDN	46	37	80.4	37	80.4	36	78.3
C1P2	291	274	94.2	266	91.4	228	78.4
C1PN	259	224	86.5	216	83.4	206	79.5
CNP2	176	152	86.4	151	85.8	149	84.7
CP1	718	635	88.4	628	87.5	558	77.7
CPN	106	74	69.8	71	67.0	63	59.4
CPP	195	130	66.7	126	64.6	115	59.0
CPPN	36	28	77.8	27	75.0	26	72.2
CV	12	6	50.0	4	33.3	4	33.3
Total	2,561	2,222	86.8	2,185	85.3	2,003	78.2

Table 2 shows the STRING results from the manually produced sentences in a first run. Overall, recall varies from 86.8% with the more relaxed criterion of just capturing the FIXED dependency; to 85.3, when the number or the dependency's arguments (NB-ARG) is considered; down to 78.2% for a complete match of the dependency's arguments (ARG). Many of these errors are due to previous processing steps in the pipeline, especially POS-tagging and disambiguation.

**Table 3. Results for verbal idioms identification:
Automatically generated, transformation-derived sentences.**

Transformation	Total	#FIXED	%	#NB-ARG	%	#ARG	%
PronA	187	170	90.9	169	90.4	165	88.2
PronD	178	131	73.6	130	73.0	129	72.5
PronPos	324	268	82.7	266	82.1	265	81.8
Rdat	192	107	55.7	106	55.2	106	55.2
PassSer	185	142	76.8	141	76.2	139	75.1
PassEstar	83	70	84.3	69	83.1	68	81.9
Total	1,170	909	77.7	902	77.1	884	75.6

Table 3 shows the results of the evaluation on the set of transformation-derived, automatically generated sentences. Notice that only the idiom accepting each transformation were considered for each class, so that the total number of idioms per class varies depending on the transformation being evaluated. Global results, even if somewhat inferior (77.7% for FIXED), are similar to those found for the manually produced sentences, especially in the strictest criterion (75.6%).

For a second run, several duplicate entries of the matrix were either removed or corrected. Some of these duplicates resulted from indicating the lemma instead of the surface form of a given frozen element. Obvious input errors, like the verb in the matrix being different from the verb in the example, were also correct. The transformation-derived sentences were automatically generated again and integrated in this run. Results, shown in Table

4, indicate an overall 95.1% recall when only the FIXED dependency is considered, and 92.5% when there is a perfect match, including the arguments of the dependency.

Table 4. Results for the 2nd run evaluation.

Sentences	Count	FIXED	%	NB-ARG	%	ARG	%
base	2,542	2,429	0.956	2,400	0.944	2,337	0.919
transformed	1,157	1,088	0.940	1,083	0.936	1,083	0.936
Total	3,699	3,517	0.951	3,483	0.942	3,420	0.925

Finally, the system was also run over the 511 modified based sentences, with the system configured to consider only the verb and frozen constituents' head nouns, and to ignore the distributional properties of the free arguments, the determiners and modifiers of the frozen head nouns. The FIXED dependency was not extracted from 481 sentences (94.1%). However, in spite of the changes introduced, for 30 sentences, the system still extracts the FIXED dependency incorrectly. Some of these wrong cases are due, as expected, to the settings defined in the configuration file, e.g. determiners being ignored, as in *O Rui engoliu esses sapos* lit: ‘Rui swallowed those frogs’ ‘eat crow’. Also, the absence of an obligatory complement, as in *O João fechou [algo] a sete chaves* ‘João closed [something] with seven keys’ ‘safely locked/under lock and key’ is not enough to preclude the extraction of the FIXED dependency. In the future, we intend to produce a new example generator module, to systematically explore all the variations considered in this small sample file and test it against the lexicon-grammar using all the XIP rule-generator configurations envisaged.

5. Conclusion and future work

This paper presented the new developments in the parsing of verbal idioms in European Portuguese. Several improvements were introduced in the STRING processing chain, namely in the automatic parsing rules' generator, which is now able to produce rules that capture several transformation-derived sentences. Also, an automatic example generator was produced anew, which builds these transformed sentences from the information encoded in the lexicon-grammar matrix. A new automatic evaluator was built, featuring a more granular assessment of the system's performance, including not only the extraction of the FIXED dependency, but also its correct number of arguments, and the correspondence to the arguments stated in the reference. At a first run, results were already very promising, reaching 78.2% recall in the strictest evaluation (exact match) and 86.8% in the relaxed mode (only the dependency). Similar, though lower results were obtained for the transformed sentences: 75.6% recall (for exact match) and 77.7% (in the relaxed mode). After error analysis and correction, it was possible to improve these results, both for the base and the transformed sentences. A second run of the system produced 92.5% recall in the exact match scenario, and 95.1% in the relaxed mode. In the near future, a similar procedure to integrate the lexicon-grammar of verbal idioms from Brazilian Portuguese [Vale 2001], as well as an *extrinsic* evaluation of the system are being envisaged, using the data sets of [Baptista et al. 2014] and [Ramisch et al. (eds.) 2018, Ramisch et al. 2018].

Acknowledgments

Research for this paper was partially supported with public funds by Fundação para a Ciência e a Tecnologia (FCT) through program ref. UID/CEC/50021/2019.

References

- Ait-Mokhtar, S., Chanod, J., and Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121–144.
- Baptista, J. (2005). Construções simétricas: argumentos e complementos. In *Estudos de Homenagem a Mário Vilela*, pages 353–367. Campo das Letras, Porto.
- Baptista, J., Correia, A., and Fernandes, G. (2004). Frozen sentences of portuguese: Formal descriptions for NLP. In *Workshop on Multiword Expressions: Integrating Processing*, pages 72–79. ACL.
- Baptista, J., Fernandes, G., Talhadas, R., Dias, F., and Mamede, N. (2016). Implementing European Portuguese Verbal Idioms in a Natural Language Processing System. In Corpas Pastor, G., editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 102–115. Proceedings of EUOPHRAS 2015.
- Baptista, J. and Mamede, N. (2013). Reciprocal Echo Complements in Portuguese: Linguistic Description in view of Rule-based Parsing. In Baptista, J. and Monteleone, M., editors, *Proceedings of the 32nd International Conference on Lexis and Grammar (CLG'2013)*, pages 33–40, Faro, Portugal. CLG'2103, Universidade do Algarve – FCHS.
- Baptista, J., Mamede, N., and Gomes, F. (2010). Auxiliary verbs and verbal chains in European Portuguese. In *Computational Processing of the Portuguese Language (PROPOR 2010)*, number 6001 in LNAI/LNCS, pages 110–119.
- Baptista, J., Mamede, N., and Markov, I. (2014). Integrating verbal idioms into an NLP system. In *Computational Processing of the Portuguese Language (PROPOR 2014)*, volume 8775 of LNAI/LNCS, pages 251–256.
- Borillo, A. (1971). Remarques sur les verbes symétriques. *Langue Française*, 11(1):17–31.
- Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, (837–892).
- Constant, M. and Sigogne, A. (2011). Mwu-aware part-of-speech tagging with a crf model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56, Portland, Oregon, USA. Association for Computational Linguistics.
- Galvão, A. (2019). Processar expressões fixas em português: Geração automática de regras e exemplos a partir de um léxico-gramática. Master's thesis, Universidade de Lisboa – Instituto Superior Técnico, Lisboa.
- Gross, M. (1982). Une classification des phrases «figées» du français. *Revue Québécoise de Linguistique*, 11-2:151–185.
- Gross, M. (1996). Lexicon-grammar. In Brown, K. and Miller, J., editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge.
- Hagège, C., Baptista, J., and Mamede, N. J. (2008). Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre o INESC-L2F e a Xerox. In Mota, C. and

- Santos, D., editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 261–274. Linguateca.
- Leclère, C. (1995). Sur une restructuration dative. *Language Research*, (31-1):179–198.
- Mamede, N., Baptista, J., Diniz, C., and Cabarrão, V. (2012). STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In Abad, A., editor, *International Conference on Computational Processing of Portuguese (PROPOR 2012) - Demo Session*, Coimbra, Portugal. <http://www.propor2012.org/demos/DemoSTRING.pdf>.
- Manning, Chris; Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1st edition.
- Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.
- Ramisch, C., Ramisch, R., Zilio, L., Villavicencio, A., and Cordeiro, S. (2018). A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese. In *Computational Processing of the Portuguese Language (PROPOR 2018)*, volume 11122 of *LNAI/LNCS*, pages 24—34.
- Ramisch et al. (eds.), C. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Rassi, A., Santos-Turati, C., Baptista, J., Mamede, N., and Vale, O. (2014). The fuzzy boundaries of operator verb and support verb constructions with *dar* “give” and *ter* “have” in Brazilian Portuguese. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing, COLING 2014*, pages 92–101. ACL.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, volume 2276 of *LNAI/LNCS*, pages 1–15, Berlin. 3rd International Conference CICLing-2002, Springer.
- Vale, O. A. (2001). *Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia*. Tese de Doutorado, Universidade Estadual Paulista, Araraquara.

Enriquecendo o *corpus* CM2News: Construção e Anotação de Coleções Bilíngues de Notícias

Yasmin V. Camargo^{1,2}, Ariani Di-Felippo^{1,2}

¹Núcleo Interinstitucional de Linguística Computacional (NILC)

²Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

yvizeu@gmail.com, arianidf@gmail.com

Abstract. We present the extension of CM2News, a multi-document bilingual (Portuguese-English) corpus for Multilingual Multi-document Summarization (MMDS). We added 10 bilingual clusters of news texts to the original set of 20 clusters, and performed a lexical-conceptual annotation of the 20 new source-texts based on WordNet.

Resumo. Apresenta-se a extensão do CM2News, corpus multidocumento bilíngue (português-inglês) para pesquisas em Sumarização Automática Multidocumento Multilingue (SAMM). A extensão consistiu na adição de 10 coleções bilíngues de notícias às 20 pré-existentes e na anotação léxico-conceitual dos 20 novos textos-fonte baseada na WordNet de Princeton.

1. Introdução

Na Sumarização Automática Multidocumento Multilingue (SAMM), busca-se, por exemplo, construir métodos que partem de um conjunto de ao menos 2 textos que abordam o mesmo assunto, sendo 1 texto em uma L_x e 1 em uma língua L_y , e geram um sumário em uma das línguas-fonte (L_x ou L_y) [Evans *et al* 2004]. Na SAMM, os *corpora* [Berber Sardinha 2004], são recursos centrais ao permitirem modelar computacionalmente a sumarização, além de treinar e avaliar tais modelos/sistemas.

O CM2News [Di-Felippo 2016] é um *corpus* multidocumento bilíngue (português (pt) e inglês (in)) de textos jornalísticos. Sua primeira versão engloba 20 *cluster*, distribuídos em 6 categorias (mundo (8), política (3), saúde (4), ciência (3), entretenimento (1) e meio ambiente (1)). Cada *cluster* contém (i) 2 textos-fonte (1-pt e 1-in), (ii) 1 sumário multidocumento de referência em pt e (iii) anotação manual dos textos-fonte em nível léxico-conceitual. Totalizando 40 textos e 19.983 palavras, o CM2News subsidiou o desenvolvimento de 2 métodos de SAMM envolvendo o português [Tosta 2014, Di-Felippo *et al.* 2016] e o estudo de métricas conceituais (p.ex.: *concept frequency (cf)* e *cf-idf*) que capturaram a relevância das sentenças em *clusters* multidocumento multilingue [Chaud 2015, Chaud e Di-Felippo 2018].

Dada sua relevância/potencial, tem-se focado na expansão do *corpus*, que engloba: (i) construção e anotação léxico-conceitual de novos *clusters* bilíngues, (ii) inclusão de 1 notícia em outra língua estrangeira a todos os *clusters*, e (iii) produção de novos sumários de referência. Na Seção 2, apresenta-se a construção de 10 novos *clusters* bilíngues e, na Seção 3, descreve-se a anotação léxico-conceitual dos novos 20 textos-fonte. Na seção 4, tecem-se algumas observações finais sobre o *corpus*, destacando pesquisas de SAMM em andamento que utilizam sua versão estendida.

2. A construção dos novos *clusters* bilíngue

Os novos *clusters* foram construídos com base nas diretrizes de Tosta [2014] e Di-Felippo [2016], a saber: (i) compilação manual dos textos, (ii) seleção de fontes jornalísticas confiáveis, (iii) compilação de notícias atuais cujos assuntos sejam variados e (iv) seleção de notícias bilíngues de tamanho similar. Assim, as notícias constitutivas dos 10 *clusters* adicionais foram manualmente compiladas das fontes: (i) jornal *A Folha de São Paulo* e portal UOL para os textos em português e (ii) portais *BBC News* e CNN para os textos em inglês. Selecionaram-se notícias publicadas entre abril e outubro de 2018. Com relação à variedade de assuntos, os novos *clusters* distribuem-se nas categorias: saúde (2), poder (1), meio ambiente (3), ciência (1) e entretenimento (3). Sobre a extensão dos textos, ressalta-se que, no geral, os textos-fonte têm tamanho relativamente similares. Os novos *clusters* do CM2News estão descritos na Tabela 1.

Tabela 1. Descrição dos 10 novos *clusters* do CM2News.

Cluster	Domínio	Assunto	Documento	Publicação (data/hora)	Qt. pal./doc	Qt. pal./cluster
C21	Poder	Encontro de líderes das Coreias	D1_C21_folha	27/04/18 - 00:34	386	770
			D2_C21_bbc	27/04/18 - 08:06 (GMT)	384	
C22	Ciência	Reprodução de camundongos	D1_C22_folha	11/10/18 - 12:00	578	1.240
			D2_C22_bbc	11/10/18 - 9:46 (GMT)	662	
C23	Entreten.	Kanye West na política	D1_C23_uol	30/10/18 - 19:49	328	782
			D2_C23_bbc	31/10/18 - 7:57	454	
C24	Entreten.	Bebê de Hilary Duff	D1_C24_folha	30/10/18 - 11:00	182	285
			D2_C24_cnn	30/10/18 - 15:21 (GTM)	103	
C25	Entreten.	Acusações a Stallone	D1_C25_uol	31/10/18 - 05:05	150	280
			D2_C25_bbc	31/10/18 - 1:45 (GTM)	130	
C26	Meio ambiente	Oleoduto EUA-Canadá	D1_C26_folha	09/11/18 - 17:55	428	973
			D2_C26_bbc	09/11/18 - 14:32 (GTM)	545	
C27	Meio ambiente	Ataque de leoa em zoológico	D1_C27_folha	22/10/18 - 16:15	220	419
			D2_C27_bbc	22/10/18 - 10:11 (GMT)	199	
C28	Meio ambiente	Baleia morta na Indonésia	D1_C28_uol	21/11/18 - 11:21	287	646
			D2_C28_cnn	21/11/18 - 16:57 (GMT)	359	
C29	Saúde	EUA poliomielite	D1_C29_folha	17/10/18 - 8:00	390	782
			D2_C29_cnn	23/10/18 - 12:27 (GMT)	392	
C30	Saúde	Camisinha autolubrificante	D1_C30_uol	18/10/18 - 12:17	522	1.056
			D2_C30_cnn	19/10/18 - 15:20 (GMT)	434	
Total						7.233

3. A anotação léxico-conceitual dos novos textos-fonte

A anotação foi feita por um linguista computacional e durou 20 dias, em sessões diárias de 60 a 90 minutos. Seguindo Di-Felippo [2016], explicitaram-se os conceitos nominais via MulSen¹ (versão multilíngue do NASP [Nóbrega 2013]), que (i) identifica os nomes via *tagging* (etiquetação morfossintática) e (ii) recupera os conceitos da WordNet de Princeton [Fellbaum 1998] em função do nome a ser anotado. A anotação seguiu 4 regras gerais: (i) anotar de início o texto em inglês do *cluster*, pois a recuperação dos possíveis conceitos da WordNet é direta, (ii) anotar os nomes não

¹ <http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/files/MulSEN.zip>.

detectados pelo *tagger*, (iii) ignorar as palavras erroneamente detectadas como nome pelo *tagger*, e (iv) anotar as ocorrências de um conceito com o mesmo *synset*².

A anotação léxico-conceitual de um nome *n* em inglês inicia com a recuperação automática de todos os *synsets* que contêm *n* e a sugestão do *synset* mais adequado por um algoritmo de desambiguação [Nóbrega 2013], que pode (ou não) ser validado.

Caso a sugestão não seja adequada, pode-se identificar outro *synset* entre os recuperados pelo MulSen. A anotação de um nome *n* em português segue basicamente os mesmos passos. A exceção é um processo adicional de tradução automática (pt-in) (via WordReference®³) para que o MulSen recupere os conceitos/*synsets* na WordNet em função do nome que se quer anotar.

A anotação conceitual seguiu 4 regras específicas: (i) uma vez o *tagger* identifica apenas lexias simples (e.ex: [gás_N de pimenta]), anotar todo nome que é núcleo de uma expressão multipalavra com o *synset* que expressa o significado da expressão (p.ex.: [gás<{pepper spray}> de pimenta]; (ii) analisar todas as traduções recuperadas do WordReference® antes de selecionar a mais adequada, posto que a melhor tradução não necessariamente é a primeira da lista recuperada pelo editor; o mesmo procedimento se aplica à seleção do *synset*; (iii) caso necessário, procurar traduções mais adequadas em repositórios externos, inserindo-as manualmente no MulSen, e analisar todos os *synsets* recuperados em função dessas traduções, e (iv) caso a WordNet não contiver certo conceito, selecionar um *synset* hiperônimo, pois os conceitos nominais estão organizados hierarquicamente na base de dados. No total, anotaram-se 1.593 nomes, distribuídos nos *clusters* como indicado na Tabela 2.

Tabela 2. Distribuição da anotação léxico-conceitual nos novos *clusters* bilíngues do CM2News.

Cluster	Qt. Nomes anotados	Cluster	Qt. Nomes anotados
C21	202	C26	163
C22	255	C27	87
C23	143	C28	134
C24	68	C29	250
C25	51	C30	240

4. Considerações finais

Por meio da expansão aqui descrita, o CM2News passou a ter 30 coleções bilíngues (pt-en) que totalizam 27.270 palavras, aumentando o volume de dados para as pesquisas em SAMM que envolvem o português. Atualmente, tem-se conduzido a produção de sumários (abstrativos) de referência em português para os novos *clusters*, seguindo os critérios de Tosta [2014]. Com base na anotação léxico-conceitual dos nomes nas 30 coleções, Camargo [2018] tem conduzido o desenvolvimento de métodos de seleção de conteúdo para a SAMM que consideram (i) pontuação ou peso diferenciado para os conceitos mais genéricos (hiperônimos) de um *cluster*, os quais são relevantes para a construção de extratos do tipo informativo/genérico, e (ii) identificação da redundância baseada na medida *concept overlap*, que considera a ocorrência de expressões distintas de um mesmo conceito (sinônima e equivalência) no *cluster* para calcular a similaridade entre sentenças. Nascimento [2018], por sua vez, conduz uma pesquisa que visa refinar a avaliação de métodos extrativos de SAMM, variando (i) a taxa de

² Conjunto de sinônimos que representa um conceito lexicalizado; p.ex: o conceito “veículo que se move por motor próprio” é representado pelo *synset* {car, auto, automobile, machine, motorcar}.

³ <http://www.wordreference.com/>

compressão (isto é, tamanho ou extensão em número de palavras) dos extratos automáticos e (ii) a língua materna dos produtores dos sumários de referência. Para tanto, tem-se adicionado 1 notícia em alemão a cada um dos 30 *clusters*, transformando-os em coleções trilíngues, e construído sumários de referência (em português) produzidos por falantes do português e do inglês a partir da leitura dos 3 textos-fonte.

Agradecimento. À CAPES, pelo suporte financeiro.

Referências bibliográficas

- Berber Sardinha, T. B. (2004). Lingüística de corpus. São Paulo, Manole, 410 p.
- Camargo, Y.V. (2018). Multilingual Multi-Document Summarization: content selection and redundancy treatment based on lexical-conceptual knowledge. In the Proceedings of the Student Research Workshop (SRW) of the 13th International Conference on the Computational Processing of Portuguese, pp. 1-4. September, 24, Canela/RS, Brazil.
- Chaud, M.R. (2015). Investigação de estratégias de seleção de conteúdo baseadas na UNL (Universal Networking Language). 2015. 157f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, SP.
- Chaud, M.R.; Di-Felippo, A. (2018) Exploring content selection strategies for Multilingual Multi-Document Summarization based on the Universal Network Language (UNL). Revista de Estudos da Linguagem, v. 26 (1), p. 45-71.
- Di-Felippo, A. (2016). CM2News: Towards a Corpus for Multilingual Multi-document Summarization. In the Proceedings of the Workshop on Corpora and Tools for Processing Corpora (CTPC), Collocated with PROPOR 2016 (The 12th International Conference on the Computational Processing of Portuguese), Tomar, Portugal, p.1-8.
- Di-Felippo, A. Tosta, F. E. S., Pardo, T. A. S. (2016). Applying Lexical-Conceptual Knowledge for Multilingual Multi-Document Summarization. In the Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR). Lecture Notes in Computer Science, Vol 9727, Springer, pp. 38-49, July, 13-15. Tomar, Portugal. ISBN 978-3-319-41552-9
- Evans, D.K.; Klavans, J.L.; McKeown, K.R. (2004). Columbia NewsBlaster: multilingual news summarization on the web. In the Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Boston, p.1-4
- Fellbaum, C. (1998): Wordnet: an electronic lexical database (Language, speech and communication). Massachusetts: MIT Press.
- Nascimento, D.X. Exploring the evaluation of automatic multilingual multi-document summaries. In the Proceedings of the Student Research Workshop (SRW) of the 13th International Conference on the Computational Processing of Portuguese, pp. 1-4. September, 24, Canela/RS, Brazil.
- Nóbrega, F.A.A. (2013). Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - ICMC, USP, São Carlos.
- Tosta, F.E.S. (2014). Aplicação de conhecimento léxico-conceitual na Sumarização Multidocumento Multilíngue. 116p. Dissertação (Mestrado) - Universidade Federal de São Carlos - UFSCar.

Violações linguísticas em referências a entidades do tipo “pessoa” em extratos automáticos multidocumento

Luana Fonseca Cristini^{1,2} and Ariani Di-Felippo^{1,3}

¹ Interinstitutional Center for Computational Linguistics (NILC), São Carlos/SP, Brazil

² College of Letters and Sciences (FCL), São Paulo State University (UNESP)

Rodovia Araraquara-Jaú Km 1, Araraquara, 14800-901, Brazil

³ Language and Literature Department (DL), Federal University of São Carlos (UFSCar)

Rodovia Washington Luís, km 235 - SP 310, São Carlos, 13565-905, Brazil

{luanafcristini;arianidf}@gmail.com

Abstract. We present the typification of linguistic violations’ in references to people occurred in multi-document summaries generated by the RSumm and GistSumm summarizers based on the CSTNews corpus. This task allows us to evaluate the impact of the rewrite rules on the overall quality of automatic summaries and to develop automatic methods of violations detection.

Resumo. Descreve-se a tipificação de violações linguísticas em referências a “pessoa” ocorridas em sumários multidocumento gerados pelos sistemas RSumm e GistSumm a partir do corpus CSTNews. Essa tarefa permite avaliar o impacto da reescrita de referências na qualidade dos sumários automáticos e na criação de métodos automáticos de detecção das violações.

1. Introdução

Os sumarizadores automáticos multidocumento comumente geram um sumário a partir de uma coleção de notícias que tratam de um mesmo assunto [Mani 2001]. Nos métodos extractivos, os sumários (extratos) são compostos pela justaposição das sentenças mais centrais da coleção extraídas integralmente dos textos-fonte, gerando vários problemas de coesão e coerência [Nenkova e McKeown, 2011].

Alguns deles ocorrem no nível das entidades nomeadas: (i) primeira menção sem explicação (1M-EXP), (ii) menção subsequente com explicação (nM+EXP), (iii) acrônimo sem explicação (ACR-EXP), (iv) sintagma nominal (SN) definido com referência a menção anterior (SNdef-REF), (vi) SN indefinido com referência a menção anterior (SNind+REF), (v) pronome sem antecedente (PRO-ANT) e (iv) pronome com antecedente enganoso (PRO-ENG) [Kaspersson *et al.* 2012, Friedrich *et al.* 2014 e Dias 2016].

Neste artigo, apresenta-se a tipificação das violações linguísticas específicas das referências ou menções a entidades do tipo “pessoa” que ocorrem em extratos automáticos multidocumento em português. Tal tarefa pode contribuir para: (i) avaliação do impacto da reescrita de menções na informatividade e na qualidade linguística dos extratos, posto que a reescrita tem se mostrado uma alternativa de pós-edição bastante viável [Nenkova e McKeown 2003a,b e Siddharthan *et al.* 2011] e (ii) na criação de métodos automáticos de detecção das violações em menções desse tipo.

2. Tipificação das Violações nos Extratos Automáticos

Neste trabalho, utilizaram-se dois sumarizadores: GistSumm [Pardo 2005] e RSumm RSumm [Ribaldo *et al* 2012, 2016]. O GistSumm [Pardo 2005] é um sistema superficial que seleciona as sentenças para compor um extrato pela frequência das palavras das sentenças na coleção e similaridade lexical entre a sentença que possui as palavras mais frequentes (*gist sentence*) e as demais da coleção. O RSumm [Ribaldo *et al* 2012, 2016] é um sistema híbrido, pois une medidas estatísticas aplicadas a uma modelagem em grafo dos textos-fonte e informação de subtópicos. Devido à sofisticação do método, o RSumm gera extratos com maior informatividade e qualidade linguística que o GistSumm. Assim, tais sumarizadores foram escolhidos devido ao objetivo de se observar o impacto da reescrita na qualidade linguística e informatividade em extratos gerados por sistemas de desempenho bem diferentes.

Cada um dos sistemas gerou 1 extrato com aproximadamente 100 palavras para cada um dos 50 *clusters* do CSTNews [Cardoso *et al.* 2011]. Tais *clusters* são compostos por 2 ou 3 notícias em português sobre mesmo assunto, provenientes de diferentes fontes jornalísticas, e estão englobam notícias de diferentes domínios: esporte (10 *clusters*), mundo (14 *clusters*), dinheiro (1 *clusters*), política (10 *clusters*), ciência (1 *clusters*) e “cotidiano” (14 *clusters*).

As menções problemáticas nos extratos gerados pelos GistSumm e RSumm foram manualmente identificadas e tipificadas com uma anotação no seguinte formato: <e TYPE=(Error Type)>(Text Passage)</e>. Para preencher Error Type, os anotadores dispunham das 7 etiquetas de Dias (2016) (isto é, 1M-EXP, nM+EXP, ACR-EXP, SNdef-REF, SNind+REF, PRO-ANT e PRO_ENG) e de outras 6 etiquetas secundárias que foram propostas para especificar as violações que tipicamente ocorrem em primeiras menções e menções subsequentes a pessoas.

Tendo em vista que as entidades do tipo pessoa tendem a ser introduzidas nas notícias por uma menção com núcleo *full name* e um *pre-modifier* [Di-Felippo 2016], propuseram-se as 3 etiquetas secundárias –*FullName*, –*PreMod* e –*FullName/-PreMod* para explicitar especificamente o(s) elemento(s) ausente(s), que deveria(m) compor a estrutura prevista para as primeiras menções. Para as menções subsequentes, que tendem a ter somente um *first name* ou *noun* como núcleo [Di-Felippo 2016], as 3 etiquetas secundárias +*PreMod*, +*PostMod* e +*PreMod/+PostMod* foram propostas para explicitar especificamente o(s) elemento(s) que não deveria(m) estar presentes na estrutura desse tipo de menção.

Na Tabela 1, apresentam-se as combinações de etiquetas genéricas e secundárias empregadas na anotação dos 100 extratos automáticos.

Tabela 1. Etiquetas para anotação de violações em referências a “pessoa”.

Violão	Etiqueta
Acrônimo sem explicação	ACR-EXP
SN definido sem referência a menção anterior	SNdef-REF
Primeira menção sem “explicação” (nome completo)	1M-EXP [-FullName]
Primeira menção sem “explicação” (pré-modificador)	1M-EXP [-PreMod]
Primeira menção sem “explicação” (nome completo e pré-mod)	1M-EXP [-FullName/-PreMod]
Menção subsequente com “explicação” (pré-modificador)	nM+EXP[+PreMod]
Menção subsequente com “explicação” (pós-modificador)	nM+EXP[+PostMod]
Menção subsequente com “explicação” (pré- e pós-modificador)	nM+EXP [+PreMod/+PostMod]

Assim, a primeira violação ocorrida no extrato automático da Figura 2, por exemplo, é <e TYPE=1M-EXP[-PreMod/-FullName]>Cahe</e>. Nessa anotação, explicita-se que a “primeira menção” (1M) (Cahe) “não possui explicação” (-EXP) adequada para a identificação do referente, a qual, no caso, refere-se à “ausência de pré-modificador”¹ e núcleo do tipo *full name* (-PreMod/-FullName).

Nas Tabela 2 e 3, apresentam-se, respectivamente, os resultados da anotação das violações nos extratos gerados pelo GistSumm e RSumm. Nessas tabelas, as violações estão organizadas por extrato/*cluster*.

Tabela 2. Distribuição das violações nos extratos do GistSumm.

Extrato/Cluster	Violação	Qt
C05	nM+EXP [+PreMod/+PostMod]	1
	nM+EXP [+PreMod]	1
	1M-EXP [-PreMod/-FullName]	2
	1M-EXP [-PreMod]	2
C07	1M-EXP [-PreMod/-FullName]	1
C08	1M-EXP [-PreMod/-FullName]	1
C14	nM+EXP [+PreMod]	1
C17	1M-EXP [-PreMod/-FullName]	1
C18	1M-EXP [-PreMod/-FullName]	1
C19	1M-EXP [-PreMod/-FullName]	2
C21	nM+EXP [+PreMod]	1
C24	1M-EXP [-PreMod]	1
C25	1M-EXP [-PreMod/-FullName]	1
C27	1M-EXP [-PreMod]	1
	1M-EXP [-PreMod/-FullName]	2
C28	1M-EXP [-PreMod/-FullName]	1
C31	nM+EXP [+PreMod]	1
C33	1M-EXP [-PreMod/-FullName]	1
C35	1M-EXP [-PreMod/-FullName]	1
C38	1M-EXP [-PreMod]	4
C40	1M-EXP [-PreMod/-FullName]	2
	nM+EXP [+PostMod]	1
C41	1M-EXP [-PreMod]	5
C42	SNdef-REF [-PreMod/-FullName]	1
	1M-EXP [-PreMod/-FullName]	1
C44	1M-EXP [-PreMod/-FullName]	2
C45	1M-EXP [-FullName]	1
C48	SNdef-REF [-PreMod/-FullName]	1
	1M-EXP [-PreMod/-FullName]	2
	nM+EXP [+PreMod]	1
C50	1M-EXP [-PreMod/-FullName]	1
Total		45

¹ Os pré-modificadores são nomes ou adjetivos que, antepostos ao núcleo da menção, informam o leitor sobre afiliação, cargo ou função exercido pela entidade “pessoa”. Os pós-modificadores, pospostos ao núcleo, podem ser do tipo oposto, sintagma preposicional, sintagma adjetival ou oração relativa (ou ainda observações parentéticas).

Tabela 3. Distribuição das violações nos extratos do RSumm.

Extrato/Cluster	Violação	Qt
C02	1M-EXP [-FullName]	2
	1M-EXP [-PreMod]	1
C07	1M-EXP [-PreMod/-FullName]	1
C08	1M-EXP [-FullName]	2
	1M-EXP [-PreMod/-FullName]	1
C19	1M-EXP [-PreMod/-FullName]	2
	nM+EXP [+PostMod]	1
C21	nM+EXP [+PreMod]	1
C24	SNdef-REF [-PreMod/-FullName]	1
C25	1M-EXP [-PreMod/-FullName]	2
	1M-EXP [-PreMod]	1
C33	nM+EXP [+PreMod]	1
	SNdef-REF [-FullName]	1
C34	nM+EXP [+PreMod]	1
C35	1M-EXP [-PreMod/-FullName]	1
	nM+EXP [+PreMod/+PostMod]	1
C36	ACR-EXP	1
C38	1M-EXP [-PreMod]	4
C43	1M-EXP [-PreMod/-FullName]	1
	1M-EXP [-PreMod]	1
	nM+EXP [+PreMod]	1
C44	1M-EXP [-PreMod/-FullName]	2
C45	1M-EXP [-PreMod/-FullName]	1
C47	nM+EXP [+PreMod]	1
C48	1M-EXP [-PreMod/-FullName]	5
C50	1M-EXP [-PreMod/-FullName]	1
Total		38

Nas Tabela 4 e 5, apresentam-se as violações organizadas pelo tipo de menção.

Tabela 4. Distribuição das violações nas 1^{as} menções por sistema.

Primeira menção		
Violação	GistSumm	RSumm
1M-EXP [-PreMod/-FullName]	22	17
1M-EXP [-PreMod]	13	7
1M-EXP [-FullName]	1	4
SNdef-REF [-PreMod/-FullName]	2	1
SNdef-REF [-FullName]	-	1
ACR-EXP	-	1
Total	38	31

Tabela 5. Distribuição das violações nas menções subsequentes por sistema.

Menção subsequente		
Violação	GistSumm	RSumm
nM+EXP [+PreMod]	5	5
nM+EXP [+PreMod/+PostMod]	1	1
nM+EXP [+PostMod]	1	1
Total	7	7

De acordo com as Tabelas 2 e 3, identificaram-se 45 menções problemáticas nos extratos gerados pelo GisSumm e 38 nos do RSumm. Embora haja mais casos nos extratos do GistSum em termos absolutos, a quantidade média de violações é praticamente a mesma para os dois sistemas. Ambos têm média aproximada de 2 violações por extrato, já que as 45 do GistSumm se distribuem em 23 extratos e as 38 do RSumm, em 18 extratos.

Com base nas Tabela 4 e 5, as violações nas 1^{as} menções são mais frequentes que nas subsequentes. Tais violações resultam do fato de que a 1^a menção do extrato automático ocorreu como menção subsequente no texto-fonte do qual foi extraída, não sendo suficientemente informativa. O tipo mais frequente de violação é de 1^a menção sem núcleo *full name* e sem *pre-modifier* (1M-EXP[-PreMod/-FullName]), ilustrado na Figura 2, por exemplo, pela 1^a menção excessivamente curta a “Cahe”. O segundo tipo mais frequente é de 1^a menção sem *pre-modifier* (1M-EXP[-PreMod]).

Quanto às menções subsequentes, identificaram-se 7 casos de violações nos extratos do GistSumm e do RSumm. A distribuição dos casos entre os 3 tipos é a mesma nos dois conjuntos de extratos: (i) 5 casos de nM+EXP [+PreMod], (ii) 1 caso de nM+EXP[+PreMod/+PostMod], e (iii) 1 caso de nM+EXP [+PostMod]. A violação mais comum pode ser ilustrada pela menção subsequente “Maradona, 46” da Figura 2, que veicula a idade da entidade “pessoa”. Esse tipo de violação resulta do fato de que a menção subsequente de um extrato automático ocorreu como 1^a menção no texto-fonte de origem.

3. Potencialidades da anotação

A anotação aqui descrita permite que se aprofunde o conhecimento sobre os problemas gerados pelos sumarizadores extrativos e que se investigue o impacto da reescrita das menções a pessoas na qualidade linguística e informatividade dos extratos automáticos.

Uma vez que as violações tenham sido anotadas e tipificadas, estas podem ser reescritas de tal maneira que atendam às preferências identificadas em sumários humanos multidocumento. Diz-se isso porque a revisão², entendida como um processo de pós-edição dos extratos, é uma estratégia de abstração relativamente mais barata que as demais (p.ex.: compressão sentencial e fusão de informação) reconhecidamente útil para a melhoria dos extratos automáticos [Nenkova e McKeown 2011].

Para tal reescrita das referências, destaca-se que Di-Felippo (2016), com base na descrição manual das cadeias de correferência em 50 sumários humanos (de 100 palavras) produzidos a partir dos *clusters* do CSTNews, identificou preferências quanto à forma e à sequência das menções, o que resultou em um conjunto de regras de reescrita para menções a entidades do tipo “pessoa” (Figura 1).

Tais regras, no entanto, ainda não foram testadas ou avaliadas e, para isso, a produção de versões reescritas de extratos automáticos é necessária. Uma vez que as violações nos extratos gerados pelo GistSumm e RSumm para as 50 coleções do *corpus* CTSNews foram anotadas e que Cristini e Di-Felippo (2018) já haviam anotado as cadeias de correferência de entidades do tipo pessoa nos textos-fonte do CSTNews, pode-se proceder à reescrita das referências como ilustrado na sequência.

² A revisão consiste em qualquer modificação realizada nos extratos como eliminação, combinação e/ou substituição de expressões e/ou sentenças.

Regra de reescrita para primeira menção (*discourse-new*) a person

1. IF o núcleo da referência não for *full name* THEN:
 - (a) Analisar todas as primeiras menções do *input*³ com o objetivo de identificar *full name*
 - (b) IF *full name* for encontrado no *input* THEN:
 - i. Reescrever a menção original por *full name*.
 - (c) ELSE IF nenhum *full name* for encontrado no *input* THEN:
 - i. Não reescrever o núcleo da menção.
2. IF a primeira menção não for acompanhada de *pre-modifier*
 - (a) Analisar todas as primeiras menções do *input* com o objetivo de identificar *pre-modifier*
 - (b) IF qualquer *pre-modifier* for encontrado no *input* THEN:
 - i. Inserir o *pre-modifier* mais longo no SN.
 - (c) ELSE IF nenhum *pre-modifier* for no *input* THEN:
 - i. Analisar todas as primeiras menções do *input* com o objetivo de identificar um pós-modificador do tipo *appositional phrase*
 - ii. Selecionar o *appositional phrase* mais longo e incluí-lo no SN da primeira menção.
 - (d) ELSE IF nenhum (*pre-* ou *post-*) *modifier* for encontrado no *input* THEN:
 - i. Manter o núcleo já reescrito ou a menção original.

Regra de reescrita para menção subsequente (*discourse-old*) a person

1. IF o núcleo da referência não for *first name* ou *noun* THEN:
 - (a) Analisar todas as menções do *input* com o objetivo de identificar *first name*
 - (b) IF *first name* for encontrado no *input* THEN:
 - i. Reescrever a menção original por *first name* e remover os *pre-* e *post-modifiers*
(a não ser que seja um acrônimo parentético)
 - (c) ELSE IF nenhum *first name* for encontrado no *input* THEN:
 - i. Analisar todas as menções do *input* com o objetivo de identificar *noun*.
 - ii. IF *noun* for encontrado no *input* THEN:
 - I. Reescrever a menção original por *noun* e remover os *pre-* e *post-modifiers*
 - iii. ELSE IF *noun* não foi encontrado no *input* THEN:
 - I. Não reescrever o núcleo da menção.

Figura 1. Algoritmo de reescrita para referências a pessoas [Di-Felippo 2016].

Apenas para ilustrar o processo de reescrita, aplicam-se as regras de Di Felippo às menções problemáticas do extrato da Figura 2. O extrato de C19 apresenta 2 casos de 1M-EXP[-PreMod/-FullName] e 1 caso de nM+EXP[+PostMod].

<e TYPE=1M-EXP[-PreMod/-FullName]>Cahe</e> disse ainda que <e TYPE=1M-EXP [-PreMod/-FullName]>Maradona</e> não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador - Cahe descartou pancreatite ou úlcera. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <e TYPE=nM+EXP[+PostMod]>Maradona, 46,</e> desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.

Figura 2. Erros anotados no extrato gerado pelo RSumm para C19.

³ Entende-se *input* como a coleção de textos-fonte a ser sumarizada.

Diante dessas violações, que se referem às entidades “Alfredo Cahe” (E1) e Diego Maradona (E2), recuperam-se de forma manual todas as menções dessas entidades anotadas nos textos-fonte da coleção, juntamente com sua caracterização linguística. Na Figura 3, vê-se que, para a reescrita da 1^a menção à E1, a menção 1 (M1) do texto 1 (D1) (negrito) é a opção mais adequada, pois possui núcleo *full name* (“Alfredo Cahe”) e o *pre-modifier* mais longo (“O médico pessoal do argentino Diego Maradona”).

Tendo em vista que o pré-modificador da menção reescrita de E1 engloba uma primeira menção à E2 com estrutura *Pre-modifier + Full name* (ou seja, “o argentino Diego Maradona”), a primeira menção original a E2 (“Maradona”) passou a ser *uma* menção subsequente. Com núcleo do tipo *last name*, essa menção, agora subsequente, não satisfaz a regra da Figura 1, tendo de ser reescrita por *first name* ou *noun*. Dessa duas opções, observa-se na Figura 3 que somente menções subsequentes com núcleo *noun* ocorrem nos textos-fonte da coleção (em negrito). Em D1, tem-se “o ex-craque” (M2) e “o ex-jogador” (M4) (em negrito). Em D2, por sua vez, ocorrem “o ex-jogador” (M4) e “o ídolo argentino” (M8) (em negrito). Ao descartar “o ídolo argentino” devido à ocorrência de pós-modificação (“argentino”), o que não é desejável em menções subsequentes, restaram duas opções, “o ex-craque” e “o ex-jogador”. No caso, selecionou-se “o ex-jogador”, pois, embora sendo mais longa, foi considerada mais informativa que “o ex-craque”. A menção “o ex-jogador”, aliás, também foi utilizada para a reescrita da menção subsequente “Maradona, 46,”, cujo problema foi anotado como nM+EXP[+PostMod] devido à presença de um *pós-modificador* (“46”).

Entidade/Menção/Doc	Texto da Menção	Headedness	Definiteness	PreMod	PostMod
E1_M1_D1	O médico pessoal do argentino Diego Maradona, Alfredo Cahe,	FullName	DefArt	Any	None
E1_M2_D1	Cahe	LastName	None	None	None
E1_M3_D1	Cahe	LastName	None	None	None
E1_M4_D1	Cahe	LastName	None	None	None
E1_M1_D2	seu médico pessoal, Alfredo Cahe	FullName	Possessive	Any	None
E1_M2_D2	o médico	Noun	DefArt	None	None
E1_M3_D2	Cahe	LastName	None	None	None
E2_M1_D1	o argentino Diego Maradona	FullName	DefArt	Any	None
E2_M2_D1	o ex-craque	Noun	DefArt	None	None
E2_M3_D1	Maradona	LastName	None	None	None
E2_M4_D1	o ex-jogador	Noun	DefArt	None	None
E2_M5_D1	Maradona	LastName	None	None	None
E2_M6_D1	Maradona, 46	LastName	None	None	Other
E2_M7_D1	Maradona	LastName	None	None	None
E2_M1_D2	Maradona	LastName	None	None	None
E2_M2_D2	ele	Pronoun	None	None	None
E2_M3_D2	ele	Pronoun	None	None	None
E2_M4_D2	o ex-jogador	Noun	DefArt	None	None
E2_M5_D2	Maradona	LastName	None	None	None
E2_M6_D2	ele	Pronoun	None	None	None
E2_M7_D2	Maradona	LastName	None	None	None
E2_M8_D2	o ídolo argentino	Noun	DefArt	None	AdjP

Figura 3. Cadeias de correferência dos textos-fonte de C19.

Ao final da aplicação das regras, tem-se a versão reescrita do extrato (Figura 4), a qual não apresenta mais os problemas destacados na Figura 2.

O médico pessoal do argentino Diego Maradona, Alfredo Cahe, disse ainda que o ex-jogador não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador --Cahe descartou pancreatite ou úlcera. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. **O ex-jogador** desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.

Figura 4. Versão reescrita do extrato de C19 gerado pelo RSumm.

3. Considerações finais

Uma vez que se tenha gerado as versões reescritas dos extratos gerados pelo GistSumm e RSumm que apresentam violações em menções a pessoas, pretende-se avaliar o impacto da reescrita na qualidade linguística e na informatividade dos extratos automáticos. A avaliação da qualidade poderá ser feita de duas formas distintas. Uma das avaliações pode consistir na análise dos extratos automáticos (original e versão reescrita) por meio do julgamento humano quanto aos 5 parâmetros proposta na *Document Understanding Conference* (DUC) de 2005 (DANG, 2005): (i) gramaticalidade, (ii) não-redundância, (iii) clareza referencial, (iv) foco (temático), e (v) estrutura/coerência. Na outra avaliação, pretende-se aplicar o mesmo procedimento realizado por Siddharthan *et al.* (2011), que consistiu no julgamento das versões reescritas em comparação às suas versões originais. No caso, um extrato automático original e a sua respectiva versão com as referências reescritas são submetidos à avaliação de um humano. Quanto à avaliação do impacto das reescritas na informatividade dos extratos automáticos multidocumento, poder-se-á utilizar o tradicional pacote de medida ROUGE [Lin 2004], que calcula a informatividade pela coocorrência de n-gramas entre sumários automáticos e humanos (ou de referência) e a expressa pelas medidas “precisão”, “cobertura” e “medida-f”.

Agradecimento. À FAPESP, pelo suporte financeiro (Proc. Nº 2017/15344-8)

Referências bibliográficas

- Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di-Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, Cuiabá/MT, Brazil.
- Cristrini, L.F.; Di-Felippo, A. (2018) Source Texts Annotation for Rewriting References to People in Automatic Multi-Document Extracts. In the Proceedings of the PROPOR Student Research Workshop (Tilic), pp. 1-5. September, 24. Canela, RS/Brazil.

- Dang, H.T. (2005). Overview of DUC 2005. In the Proceedings of the Document Understanding Conference (HLT/EMNLP Workshop on Text Summarization), 2005.
- Di-Felippo, A. (2016). “Revisão de sumários baseada em conhecimento: transformando extratos multidocumento em *abstracts*”. Relatório de Bolsa de Pesquisa no Exterior (FAPESP #2015/01450-5). <http://www.nilc.icmc.usp.br-nilc/index.php/team?id=23>.
- Friedrich, A., Valeeva, M., Palmer, A. (2014). “LQVSumm: a corpus of linguistic quality violations in multi-document summarization”. In the Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik/ISL, pp. 1591-1599.
- Kaspersson, T.; Smith, C.; Danielsson, H.; Jönsson, A. (2012). This also affects the context - Errors in extraction based summaries. In the Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul/TU, pp.173-8.
- Lin, C-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In the Proceedings of the ACL Workshop on Text Summarization Branches, p. 74-81.
- Mani, I. (2001). Automatic summarization. Amsterdam: John Benjamins Publishing Co.
- Nenkova, A.; McKeown, K. (2003a). Improving the Coherence of Multi-document Summaries: a Corpus Study for Modeling the Syntactic Realization of Entities, Columbia University, CS Department Technical Report, CUCS-001-03.
- Nenkova, A.; McKeown, K. (2003b). Improving the Coherence of Multi-document Summaries: a Corpus Study for Modeling the Syntactic Realization of Entities, Columbia University, CS Department Technical Report, CUCS-001-03.
- Nenkova A.; McKeown. K. (2011). Automatic summarization. In *Foundations and Trends in Information Retrieval*, 5(2-3), pages 103–233.
- Pardo, T. A. S. (2005) GistSumm - GIST SUMMarizer: Extensões e novas funcionalidades, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.
- Ribaldo, R.; Akabane, A. T.; Rino, L. H. M.; Pardo, T. A. S. (2012). Graph-based Methods for Multidocument Summarization: Exploring Relationship Maps. Complex Networks and Discourse Information. In the Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243), Coimbra/Portugal, pp. 260–271.
- Ribaldo, R. (2013). Investigação de Mapas de Relacionamento para Sumarização Multidocumento. Monografia de Conclusão de Curso. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 61p.
- Siddharthan, A., Nenkova, A., McKeown, K. (2011). Information status distinctions and referring expressions: An empirical study of references to people in news summaries. In *Computational Linguistics* 37(4), pages 811–842.

Anotação de unidades de informação em transcrições de fala na tarefa de reconto de narrativas em português

Leandro Borges dos Santos¹, Lilian Cristine Hübner²,
Letícia Lessa Mansur³, Anderson Smidarle², Sandra Aluisio¹

¹ Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

²Escola de Humanidades
Pontifícia Universidade Católica do
Rio Grande do Sul (PUCRS) – Porto Alegre, RS – Brasil

³Faculdade de Medicina
Universidade de São Paulo (USP) – São Paulo, SP – Brasil

{leandrobs, lmansur}@usp.br, lilian.hubner@pucrs.br

anderson.smidarle@acad.pucrs.br, sandra@icmc.usp.br

Abstract. Several batteries utilize narrative recall as a subtest to identify cognitive impairment which characterizes Alzheimer's Disease and Mild Cognitive Impairment. Text retelling analyses include the identification of which of the most important textual elements could be retrieved by the participants. In this study, we present the annotation process of the information units of two sets of data originated by two batteries, the level of recall of each unit, besides automatically characterizing the sets with nine linguistic metrics. Thus, we enable the future application of automated techniques to identify important elements in stories.

Resumo. Diversas baterias utilizam o reconto de narrativas como subteste para identificação de déficits cognitivos caracterizando a Doença de Alzheimer e o Comprometimento Cognitivo Leve. A análise do reconto inclui a identificação de quais elementos textuais importantes foram relembradas pelos participantes. Neste estudo, apresentamos o processo da anotação das unidades informações de dois conjuntos de dados de duas baterias, a taxa de recordação de cada unidade, além de caracterizar os conjuntos com nove métricas linguísticas, automaticamente. Desse modo, possibilitamos a aplicação futura de técnicas automáticas para a identificação dos elementos importantes das histórias.

1. Introdução

O envelhecimento da população é uma tendência social conhecida em países desenvolvidos e que tem se tornado cada vez mais pronunciada também nos países em desenvolvimento, como o Brasil, que vem apresentando um grande crescimento da população com mais de 60 anos [Fichman et al. 2011]. O envelhecimento pode ser acompanhado de doenças neurodegenerativas, como as demências, dentre as quais a Doença de Alzheimer (DA) é a mais proeminente, correspondendo a 50 – 75% dos casos. Outra enfermidade que tem recebido atenção nos últimos anos é o Comprometimento Cognitivo Leve (CCL),

que ocasiona declínio em funções cognitivas e, em certos casos, progride para um quadro demencial [Clemente and Ribeiro-Filho 2008].

O diagnóstico das demências e síndromes relacionadas pode ser feito por exames de neuroimagem, mas comumente baseia-se na análise das funções cognitivas do paciente, pela administração de baterias de avaliação cognitiva e neuropsicológica. Investigam-se as funções que são mais afetadas como diferentes tipos de memória, orientação, linguagem e resolução de problemas. Estas baterias são usadas antes, durante e depois de tratamentos [de Abreu et al. 2005], como diagnóstico, acompanhamento e direcionamento de tratamento. Alguns testes e baterias utilizam como subtarefa o reconto de narrativas, que pode se dar como reconto imediato ou tardio. Narrativas têm sido o tipo textual mais empregado, devido à clara estrutura de uma história narrativa, que compreende uma situação inicial, uma complicação, um desenrolar, uma situação final, e uma conclusão, além de serem empregadas no dia a dia dos falantes. As tarefas de reconto de narrativas utilizam uma história curta que é contada ao paciente, a quem se solicita que relate a história ouvida com o máximo de detalhes, para posterior transcrição e análise. A análise do reconto inclui o número de unidades de informação que podem ser relembradas pelo participante imediatamente após ouvir a história e 30 minutos depois. Durante os 2 testes há a aplicação de outros testes de natureza diferente, incluindo testes linguísticos, neuropsicológicos e cognitivos.

Um conjunto de dados anotado com proposições (ou unidades de informação recordadas) possibilita a aplicação futura de técnicas automáticas para a identificação dos elementos importantes das histórias, automatizando a aplicação destes tipos de testes neurolingüísticos e auxiliando na tarefa de classificação de narrativas [Prud'hommeaux and Roark 2012, Yancheva and Rudzicz 2016, Fraser et al. 2019].

Neste trabalho, descrevemos a proposta de uma anotação manual (Seção 3) e uma análise da quantidade das unidades de informação recordadas pelos participantes à luz do modelo de análise de discurso de [Kintsch and van Dijk 1978] (Seção 4) em dois testes de reconto diferentes, descritos na Seção 2. Também apresentamos na Seção 4 uma caracterização linguística automática das narrativas produzidas pelos dois grupos clínicos (CCL e DA) comparada com as narrativas de um grupo de controle idoso saudável, usando nove métricas publicamente disponíveis dos projetos Coh-Metrix-Port [Scarton and Aluísio 2010], Coh-Metrix-Dementia [Aluísio et al. 2016b] e Simpligo¹.

2. Caracterização dos conjuntos de dados de reconto

Utilizamos dois conjuntos de dados de reconto. A Tabela 1 apresenta as estatísticas e contrastes dos dois conjuntos de dados, com uma média do tamanho de sentenças bem próxima entre CCLs e Controles na Bateria Arizona para Desordens de Comunicação e Demência (ABCD) [Bayles and Tomoeda 1993]² (diferença de 0,5). Entretanto, tabela mostra uma diferença maior na média do tamanho de sentenças para os grupos DA e CCL com o grupo de controle da Bateria de Avaliação da Linguagem no Envelhecimento (BALE) [Hübner et al. 2019] (diferença de aproximadamente 1,6). O mesmo padrão se repete para a média das palavras por sentenças, nas duas baterias.

¹<http://simpligo.sidle.al>

²A tradução e adaptação para o português foi realizada por Danielle Rüegg, Isabel Maranhão de Carvalho, Letícia Lessa Mansur e Márcia Radanovic.

Tabela 1. Estatísticas dos Conjuntos de Dados

Bateria	Grupo	Sujeitos	Média Sentenças (Desvio Padrão)	Média de palavras por sentença (Desvio Padrão)
ABCD	CCL	23	8,17 (1,92)	60,76 (17,39)
	Controle	12	7,67 (2,06)	58,96 (14,73)
BALE	DA	11	6,09 (2,63)	36,18 (17,10)
	CCL	5	6,00 (1,00)	36,40 (5,68)
	Controle	53	7,68 (2,67)	52,06 (19,18)

O primeiro conjunto de dados é formado por transcrições da ABCD que é composta de 17 subtestes que compreendem os seguintes domínios: estado mental, memória episódica, compreensão da linguagem, produção da linguagem e construção visuoespacial. Nos interessa neste trabalho o domínio da memória episódica, que é composto pelos subtestes de reconto imediato e tardio de estória, além de outros testes, não avaliados aqui. O teste do reconto foi aplicado em 23 idosos com CCL e 12 adultos com envelhecimento saudável, na Faculdade de Medicina da USP. Este teste possui 17 unidades de informação, apresentadas na Figura 1, com possíveis alternativas entre parênteses, sendo 17 a sua pontuação máxima.

Senhora (mulher) // estava fazendo compras (na loja, foi às compras, foi ao mercado) // Sua carteira (seu porta-notas, sua moedreira) // carteira caiu (derrubou a carteira, perdeu a carteira, perdeu a bolsa) // da sua bolsa (da sua mochila, de sua pasta) // Ela não viu a carteira cair (ela não notou) // No caixa (quando ela foi pagar, guichê) // não tem como pagar (ela não tinha dinheiro, não tinha sua carteira) // Coloca as mercadorias de lado (coloca as mercadorias de volta) // foi para sua casa (voltou para sua casa) // Quando ela abriu a porta (quando ela chegou em casa, assim que ela entrou) // telefone tocou (fone tocou, ela recebeu uma ligação) // Pequena (jovem) // menina (garota) // lhe disse (falou, contou) // **ela achou a carteira (achou sua moedreira, achou o porta-notas) // Senhora aliviada (senhora estava feliz, senhora estava radiante, senhora estava agradecida)**

Figura 1. Narrativa utilizada na ABCD, separada em unidades de informação; as nove unidades da macroestrutura são marcadas em negrito

O segundo conjunto de dados é formado por transcrições da BALE, que inclui tarefas de nomeação, avaliação da memória verbal episódica, julgamento semântico, categorização semântica no nível da palavra, compreensão e conclusão da metáfora no nível da sentença, bem como quatro tarefas discursivas. As tarefas discursivas são: produção de narrativa baseada em uma sequência de sete cenas, produção de narrativa livre sobre um tema atual, produção de história engraçada não sendo uma piada, reconto e compreensão de texto de uma história apresentada oralmente (História da Lúcia), que possui originalmente 24 unidades de informação que foram reagrupadas neste trabalho (cf. Seção 3), resultando em 21 unidades (Figura 2). A bateria objetiva abordar algumas das deficiências de linguagem geralmente associadas ao CCL e à DA. Além disso, as tarefas foram desenvolvidas de modo a possibilitarem a administração junto a analfabetos e pessoas com menor escolaridade, amostras populacionais muito comuns no sistema público de saúde brasileiro. O teste do reconto foi aplicado em 11 idosos com Alzheimer, 5 idosos com CCL e 53 adultos com envelhecimento saudável.

Lúcia // mora // interior // do Paraná // Numa manhã de 2a feira // ela saiu de casa // para buscar emprego (foi para uma uma entrevista, foi buscar trabalho) // na capital do estado (em Curitiba) // Foi para rodoviária // foi de carona (pegou carona) // com amigo Pedro (com Pedro) // Estava chovendo // naquela manhã // O carro // passou (caiu) // por um buraco // o pneu furou // Pensou que ia perder (achou que ia perder) // o ônibus // Pegou um táxi // conseguiu chegar chegou a tempo (chegou a tempo)

Figura 2. Narrativa utilizada na BALE, separada em unidades de informação; as onze unidades da macroestrutura são marcadas em negrito

Nas Figuras 1 e 2, anotamos as unidades da macroestrutura em negrito, seguindo o modelo de análise de [Kintsch and van Dijk 1978] em que as unidades de informação do texto são organizadas de forma hierárquica, sendo a macroestrutura correspondente às ideias principais e a microestrutura às ideias acessórias e detalhes.

3. Proposta de uma anotação manual das unidades de informação

Para cada conjunto de dados, o áudio do participante foi transscrito manualmente, seguindo os princípios do NURC / SP No 338 EF e 331 D [Preti 2005] e segmentado manualmente em orações por um anotador experiente, usando conhecimento prosódico (pausas), sintático e semântico. Chamamos essas duas etapas de pré-processamento.

Na segmentação em orações, foram mantidas as disfluências, uma vez que estas caracterizam fortemente os grupos clínicos, mas eliminadas as marcas de incompREENsão de palavras/segmentos, prolongamentos de vogais e consoantes, silabação, interrogação, pausas curtas e longas e comentários descritivos do transcritor. Primeiro foram segmentadas as orações bem formadas sintaticamente e segmentaram-se também as orações coordenadas, pois formam ideias isoladamente. Palavras com ortografia incorreta não representam um problema para a tarefa de segmentação sentencial. Em seguida, as orações que são mal formadas sintática e/ou semanticamente também foram delimitadas.

Para criarmos os conjuntos de dados anotados com as unidades de informação sobre as unidades de interesse (orações anotadas no pré-processamento), utilizamos o sistema de anotação brat (brat rapid annotation tool) [Stenetorp et al. 2012], realizando a anotação em duas fases.

Na primeira fase, cada sentença da transcrição foi classificada de acordo com a lista de unidades de informações de cada bateria por um único anotador; na segunda fase, outro anotador revisou a anotação e os casos discordantes foram discutidos, visando a uma anotação concordante (cf. Figura 3).

O reconto da ABCD foi mantido com as 17 unidades de informação originais, mas para as narrativas da BALE realizamos algumas modificações nas unidades de informações (ora separando, ora juntando) para termos uma anotação manual uniforme, sem discrepâncias e possibilitar a aplicação de métodos automáticos. A partir dessas modificações, finalizamos com 21 unidades de informações (Figura 2) ao invés das 24 unidades originais, com 11 delas sendo unidades macroestruturais. Dentre essas modificações, agrupamos as unidades que eram precedidas pelo verbo “ir” como “foi para a rodoviária” e “foi de carona” e removemos a unidade de informação “foi”. Também removemos as unidades que estavam repetidas (havia duas “Lúcia” e duas proposições relacionadas com “a rodoviária”, variando somente a preposição “para” e “até”), que dificultam a análise automática. Essas mudanças alteraram a pontuação máxima da narrativa de 24 para 21 pontos, com onze unidades macroestruturais. E se mostram uma limitação somente para anotação com repetições de trechos idênticos que usam diferentes categorizações na estrutura do texto. Mais especificamente, no caso da anotação de “Lúcia”, na primeira vez é classificada como macroestrutura e na segunda lembrança anotada como unidade da microestrutura. Esse esquema pode, entretanto, ser anotado com indexação (“Lúcia1”, “Lúcia2”) ou com refraseamentos, como, por exemplo: “(foi) para a rodoviária”, anotado como unidade macro e “(um táxi) até a rodoviária”, anotado como micro, por ser um reforço somente. Acreditamos, entretanto, que essa anotação com

	<input type="checkbox"/> PARANA <input type="checkbox"/> MORA <input type="checkbox"/> LUCIA <input type="checkbox"/> INTERIOR
1	Lucia mora no interior né do paraná . <u>BUSCAR_EMPREGO</u>
2	ela foi pra ir no pra procurar emprego . <u>FOI_RODOVIARIA</u>
3	teria que pegar um ônibus na rodoviária . <u>NUMA_MANHA_SEGUNDA</u>
4	era manhã de segunda-feira . <u>ESTAVA_CHOVENDO</u>
5	estava chovendo . <u>FOI_CARONA</u> <u>COM_PEDRO</u>
6	e um colega o pedro deu carona pra ela . <u>PASSOU_CAIU</u> <u>CARRO</u> <u>BURACO</u>
7	e quando estavam indo o carro bateu num buraco . <u>PNEU_FUROU</u>
8	furou o pneu . <u>PEGOU_TAXI</u>
9	ela teve que pegar um táxi . <u>PENSOU_ACHOU_PERDER</u> <u>ONIBUS</u>
10	e ela ficou insegura porque achou que não ia chegar em tempo de pegar o ônibus . <u>CONSEGUIU_CHEGAR_TEMPO</u>
11	e aí mais conseguiu . <u>NA_CAPITAL</u>
12	e chegou na capital .

Figura 3. Exemplo da anotação das unidades de informação em uma narrativa com 12 orações e 19 unidades, no brat.

rótulos muito similares sobrecarregue o anotador, dado que a anotação usa uma lista de entidades descontextualizada (cf. Figura 4), levando a possíveis erros de anotação. Esta foi, então, a razão para alterarmos a pontuação de 24 para 21 pontos, para avaliar o sucesso (ou não) da anotação com menos pontos com vistas à identificação de semelhanças e diferenças entre os grupos de interesse.

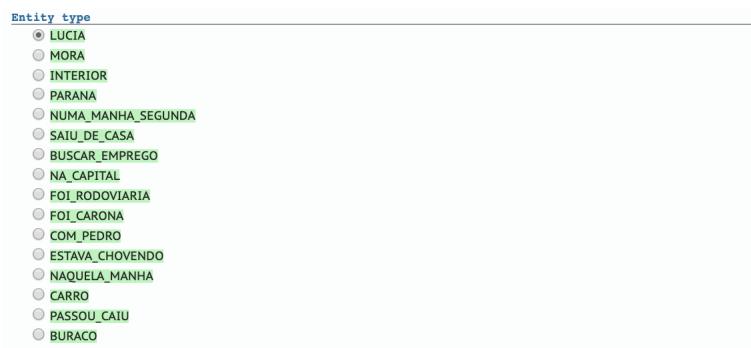


Figura 4. Exemplo do esquema de anotação com lista de entidades no brat

4. Resultados da Anotação Manual e Caracterização Automática

4.1. Análise da anotação manual

Na Tabela 2, apresentamos uma análise da quantidade das unidades de informação da ABCD. Identificamos que os idosos do grupo de controle apresentaram

uma porcentagem de unidades relembradas para componentes da microestrutura [Kintsch and van Dijk 1978] da narrativa muito mais marcante (diferença maior que 5,8 pontos) do que os do grupo CCL para as unidades “Sua carteira”, “Da sua bolsa” e “Pequena”. Mas o que é interessante é a grande discrepância para a unidade da macroestrutura “Senhora ficou aliviada”, que foi mais relembrada pelo grupo CCL (diferença de 38,4 pontos); o grupo CCL reembrou somente 1 elemento da microestrutura (“Quando abriu a porta”) com diferença marcante (7,3 pontos) quando comparado com o grupo de controle.

Tabela 2. Porcentagem, média e desvio padrão das unidades de informações recordadas por cada grupo da ABCD. Unidades em negrito são unidades da macroestrutura.

Unidades de informação	Controle		CCL	
	Unidades Recordadas %	Média	Unidades Recordadas %	Média
Senhora	91,67	1,00 (0,42)	93,48	0,96 (0,29)
estava fazendo compras	91,67	1,25 (0,61)	97,83	1,13 (0,40)
Sua carteira	62,5	0,63 (0,49)	50	0,54 (0,59)
carteira caiu	58,33	0,58 (0,50)	47,83	0,52 (0,59)
da sua bolsa	16,67	0,17 (0,38)	10,87	0,11 (0,31)
Ela não viu	33,33	0,38 (0,58)	41,3	0,48 (0,66)
No caixa	83,33	0,96 (0,55)	82,61	0,85 (0,42)
não tem como pagar	83,33	0,88 (0,45)	86,96	1,00 (0,56)
Colocou de lado	75,00	0,75 (0,44)	78,26	0,78 (0,42)
foi para casa	91,67	0,96 (0,36)	89,13	0,89 (0,31)
Quando abriu a porta	66,67	0,67 (0,48)	73,91	0,74 (0,44)
telefone tocou	91,67	0,92 (0,28)	91,3	0,91 (0,28)
Pequena	70,83	0,75 (0,53)	52,17	0,52 (0,51)
menina	87,5	0,92 (0,41)	82,61	0,83 (0,38)
lhe disse	83,33	0,83 (0,38)	84,78	0,85 (0,36)
achou carteira	95,83	1,04 (0,36)	93,48	0,98 (0,33)
Senhora ficou aliviada	33,33	0,38 (0,58)	71,74	0,74 (0,49)

Na Tabela 3, apresentamos uma análise da quantidade das unidades de informação da BALE. Diferentemente da ABCD, os idosos do grupo de controle apresentaram um número de unidades relembradas maior (com diferença marcante) do que os pacientes do grupo CCL para várias unidades de informações da microestrutura como “Paraná”, “Numa manhã de segunda-feira”, “na capital”, “estava chovendo”, “passou”, “pensou que ia perder”; já para os idosos do grupo Alzheimer podemos elencar as unidades: “Paraná”, “Numa manhã de segunda-feira”, “na capital”, “estava chovendo”, “carro”, “passou”, “buraco”, “pensou que ia perder”.

Em geral, os idosos do grupo CCL apresentaram uma taxa de recordação maior que o grupo com Alzheimer, exceto para as unidades de informação “Lúcia”, “Foi para rodoviária” e “Pegou um táxi”.

4.2. Caracterização automática das narrativas

Para descrevermos automaticamente os conjuntos de dados, selecionamos métricas comumente utilizadas na tarefa de classificação de narrativas [Roark et al. 2011, Aluísio et al. 2016a, Santos et al. 2017, Fraser et al. 2019] ou na análise de narrativas [Toledo et al. 2018].

As métricas selecionadas se dividem em: (i) contagens básicas (média de palavras por sentença, média de sentenças da narrativa, razão de substantivos por palavras do

Tabela 3. Porcentagem, média e desvio padrão das unidades de informações recordadas por cada grupo da BALE. Unidades em negrito são unidades da macroestrutura.

Unidades de informação	Controle		CCL		Alzheimer	
	Unidades Recordadas %	Média	Unidades Recordadas %	Média	Unidades Recordadas %	Média
Lucia	96,23	1,04(0,34)	80	0,8(0,45)	90,91	1,18(0,60)
mora	66,04	0,68(0,51)	20	0,2(0,45)	9,09	0,09(0,30)
interior	54,72	0,58(0,57)	20	0,2(0,45)	9,09	0,09(0,30)
Paraná	66,04	0,68(0,51)	20	0,20(0,45)	9,09	0,09(0,30)
Numa manhã de segunda-feira	13,21	0,13(0,34)	0	—	0	—
saiu de casa	5,66	0,06(0,23)	20	0,20(0,45)	0	—
buscar emprego	56,6	0,62(0,60)	20	0,20(0,45)	18,18	0,18(0,40)
na capital	13,21	0,13(0,34)	0	—	0	—
Foi para rodoviária	54,72	0,64(0,65)	20	0,20(0,45)	36,36	0,45(0,69)
foi de carona	54,72	0,58(0,57)	40	0,60(0,89)	27,27	0,27(0,47)
com o Pedro	43,4	0,45(0,54)	20	0,20(0,45)	0	—
Estava chovendo	28,3	0,38(0,69)	20	0,20(0,45)	9,09	0,09(0,30)
naquela manhã	3,77	0,04(0,19)	0	—	0	—
Carro	62,26	0,64(0,52)	60	0,80(0,84)	18,18	0,18(0,40)
passou	35,85	0,36(0,48)	20	0,20(0,45)	18,18	0,18(0,40)
buraco	43,4	0,43(0,50)	40	0,40(0,55)	27,27	0,27(0,47)
pneu furou	71,7	0,75(0,52)	60	0,80(0,84)	45,45	0,55(0,69)
Pensou que iria perder	49,06	0,49(0,50)	40	0,40(0,55)	18,18	0,18(0,40)
ônibus	30,19	0,30(0,46)	60	0,80(0,84)	27,27	0,27(0,47)
Pegou um táxi	64,15	0,72(0,60)	40	0,60(0,89)	45,45	0,45(0,52)
conseguiu chegar a tempo	56,6	0,58(0,53)	60	0,60(0,55)	9,09	0,09(0,30)

texto, razão de verbos por palavras do texto); (ii) métricas baseadas na análise sintática (distância de dependência, complexidade de Yngve [Yngve 1960], complexidade de Frazier [Fraser et al. 2019], quantidade média de orações por sentenças e a média dos tamanhos médios dos sintagmas nominais nas sentenças). Não realizamos nenhum tratamento para remover disfluências automaticamente porque visamos à construção de um dataset *gold standard*, embora haja um sistema que extrai automaticamente as disfluências (cf. [Treviso and Aluísio 2018]). O DeepBonD remove as disfluências do tipo pausas preenchidas e marcadores do discurso com bastante precisão, embora os tipos de disfluências mais complexos (repetições e revisões) não tenham a mesma precisão.

Na Tabela 4 apresentamos os resultados da aplicação das nove métricas. Na ABCD, os valores das métricas são muito próximos para os dois grupos analisados; utilizamos o teste de Mann-Whitney com um intervalo de confiança de 95% e não encontramos diferença estatística entre os grupos. Para a BALE, utilizamos o teste estatístico Kruskal-Wallis e o pós-teste de Dunn com um intervalo de confiança de 95%; encontramos resultados estatisticamente relevantes entre os idosos dos grupos Controle vs CCL e CLL vs Doença de Alzheimer para uma métrica do grupo morfossintáticas (**Razão de substantivos por palavras do texto**) com p-valor de 0.0192 e 0.0170, respectivamente; e entre os idosos do grupo de Controle e Doença de Alzheimer para a métrica sintática **Complexidade de Yngve** com p-valor de 0.0128.

5. Conclusões e Trabalhos Futuros

Neste trabalho, apresentamos a proposta de uma anotação manual, suportada por ferramentas automáticas de anotação, para criação de dois conjuntos de dados com as unidades de informação macro e micro identificadas, e também uma análise quantitativa das unidades de informações recordadas e a extração de nove métricas linguísticas automáticas.

Tabela 4. O valores médios (desvio padrão) das métricas por cada grupo clínico

Métricas	ABCD		BALE		
	Controle	CCL	Controle	CCL	Alzheimer
Complexidade de Yngve	1.78 (0.13)	1.78 (0.12)	1.82 (0.17)	1.72 (0.13)	1.64 (0.22)
Complexidade de Frazier	6.79 (0.48)	6.64 (0.40)	6.59 (0.52)	6.67 (0.26)	6.31 (0.48)
Distância de dependência	11.35 (3.34)	10.38 (3.11)	8.66 (3.25)	7.74 (1.33)	7.33 (2.35)
Número de sentenças	7.67 (2.06)	8.17 (1.92)	7.68 (2.67)	6.00 (1.00)	6.09 (2.63)
Média de Palavras por Sentença	7.77 (1.30)	7.41(1.41)	6.81 (1.58)	6.11 (0.73)	5.85 (1.46)
Quantidade média de orações por sentença	3.09 (0.60)	2.82 (0.60)	2.31 (0.86)	2.57 (0.60)	2.10 (1.09)
Média dos tamanhos médios dos sintagmas nominais nas sentenças	2.52 (0.59)	2.43 (0.77)	2.84 (0.93)	2.92 (0.96)	2.47 (0.69)
Razão de substantivos por palavras do texto	0.24 (0.03)	0.24 (0.03)	0.30 (0.06)	0.24 (0.04)	0.31 (0.05)
Razão de verbos por palavras do texto	0.29 (0.04)	0.29 (0.04)	0.23 (0.04)	0.26 (0.05)	0.23 (0.04)

Como trabalhos futuros, pretendemos utilizar as unidades micro e macro e estender o número de métricas automáticas para a tarefa de classificação automática de narrativas dos grupos de DA, CCL e Controles. Além disso, pretendemos aplicar métodos para automatizar a tarefa de identificação de unidades de informações, para facilitar a aplicação de testes neurolinguísticos e neuropsicológicos para um número maior de pessoas, auxiliando assim a detecção precoce de demências para a intervenção e o tratamento do declínio cognitivo, incluindo o linguístico.

Referências

- Aluísio, S., Cunha, A., and Scarton, C. (2016a). Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 109–114. Springer.
- Aluísio, S. M., Cunha, A., Toledo, C., and Scarton, C. (2016b). Computational tool for automated language production analysis aimed at dementia diagnosis. In *International Conference on Computational Processing of the Portuguese Language, Demonstration Session*.
- Bayles, K. and Tomoeda, C. (1993). *ABCD: Arizona Battery for Communication Disorders of Dementia*. Tucson, AZ: Canyonlands Publishing.
- Clemente, R. S. and Ribeiro-Filho, S. T. (2008). Comprometimento cognitivo leve: Aspectos conceituais, abordagem clínica e diagnóstica. *Revista Hospital Universitário Pedro Ernesto*, 7(1):68–77.
- de Abreu, I. D., Forlenza, O. V., and de Barros, H. L. (2005). Demência de alzheimer: correlação entre memória e autonomia. *Revista de Psiquiatria Clínica*, 32:131–136.
- Fichman, H. C., Oliveira, R. M., and Fernandes, C. S. (2011). Neuropsychological and neurobiological markers of the preclinical stage of alzheimer's disease. *Psychology & Neuroscience*, 4(2):245–253.
- Fraser, K. C., Fors, K. L., and Kokkinakis, D. (2019). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139.
- Hübner, L. C., LOUREIRO, F., TESSARO, B., SIQUEIRA, E. C. G., JERÔNIMO, G. M., and SMIDARLE, A. (2019). Bale: Bateria de avaliação da linguagem no envelhe-

- cimento. In Zimmermann, N., Delaere, F., and Fonseca, R. P., editors, *Tarefas de avaliação neuropsicológica para adultos: memória e linguagem*, volume 3. Memnon, Rio de Janeiro, 1 edition.
- Kintsch, W. and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.
- Preti, D., editor (2005). *O discurso oral culto*. Associação Editorial Humanitas, São Paulo, 3 edition. Projetos Paralelos. V.2.
- Prud'hommeaux, E. T. and Roark, B. (2012). Graph-based alignment of narratives for automated neurological assessment. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP 12, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.
- Santos, L., Corrêa Júnior, E. A., Oliveira Jr, O., Amancio, D., Mansur, L., and Aluísio, S. (2017). Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1284–1296, Vancouver, Canada. Association for Computational Linguistics.
- Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.
- Toledo, C. M., Aluísio, S. M., dos Santos, L. B., Brucki, S. M. D., Trés, E. S., de Oliveira, M. O., and Mansur, L. L. (2018). Analysis of macrolinguistic aspects of narratives from individuals with alzheimer’s disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:31–40.
- Treviso, M. V. and Aluísio, S. M. (2018). Sentence segmentation and disfluency detection in narrative transcripts from neuropsychological tests. In *Computational Processing of the Portuguese Language (PROPOR)*, pages 409–418. Springer International Publishing.
- Yancheva, M. and Rudzicz, F. (2016). Vector-space topic models for detecting alzheimer’s disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2337–2346.
- Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Caracterização de desvios sintáticos em um *corpus* de redações O processo de anotação

Renata Ramisch¹, Ariani Di Felippo¹

¹Departamento de Letras
Universidade Federal de São Carlos (UFSCar)
Núcleo Interinstitucional de Linguística Computacional (NILC)
São Carlos – SP – Brasil

{renata.ramisch, arianidf}@gmail.com

Abstract. This article describes the annotation of syntactic errors in essays written by High School students in Brazilian Portuguese (BP). Using a typology of syntactic errors based on the formal written style of BP, we annotated a set of sentences from the essays, which reveals that absence of punctuation marks and verbal agreement are the most common errors. Such annotation may contribute to refine the computational treatment of these violations.

Resumo. Descreve-se a anotação de desvios sintáticos em redações nos moldes do ENEM, escritas por estudantes do ensino médio. A partir de uma tipologia de desvios da modalidade escrita formal do português brasileiro, um conjunto de sentenças foi anotado, relevando que a ausência de pontuação e a concordância verbal são os desvios mais comuns. Tal anotação pode refinar o tratamento computacional dessas violações.

1. Introdução

Escrever redações é um processo inerente à trajetória educacional. A redação também é frequentemente utilizada como mecanismo de avaliação dos conhecimentos de português e de produção textual em vestibulares e exames de seleção, como o Exame Nacional do Ensino Médio¹ (ENEM). Assim, um bom desempenho nessa tarefa garante melhores notas e, por consequência, aumenta as chances na disputa pelas vagas mais concorridas para o ensino superior. Porém, textos escritos por estudantes, mesmo na etapa final da educação básica, ainda apresentam diversos desvios de ortografia e gramática quanto à modalidade escrita esperada pelos avaliadores desses exames de seleção [Castaldo 2009].

Nesse sentido, algumas aplicações computacionais podem ser úteis no processo de correção e avaliação dos textos por professores ou avaliadores, assim como no aperfeiçoamento das habilidades de produção textual pelos próprios alunos. Exemplos de ferramentas que podem ser usadas pelos estudantes são os corretores ou revisores gramaticais (isto é, sistemas que detectam desvios gramaticais em um texto e sugerem correções [Soni e Thakur 2018]) e as ferramentas de auxílio à escrita (FAE), que dão suporte a todo o processo de escrita, seja no agrupamento de ideias ou na composição do texto. Um exemplo de FAE que auxilia na composição de textos acadêmicos em português é o SciPo

¹O ENEM é uma prova realizada pelo INEP/MEC para avaliar a qualidade do ensino médio no país e dar acesso ao ensino superior em universidades públicas brasileiras e em algumas universidades estrangeiras.

[Feltrim 2004], um ambiente na *web* composto por um conjunto de ferramentas integradas para auxiliar estudantes a escreverem resumos e introduções de textos da área da computação².

Além disso, como o processo de avaliação e correção manual dessas redações costuma ser longo e caro, tem crescido o interesse pelo desenvolvimento de aplicações computacionais que possam agilizar também a correção e/ou a avaliação humana. Um exemplo de trabalhos nessa área é o de [Santos et al. 2016], que propõem um analisador léxico-sintático para a avaliação automática de atividades escritas em português. No experimento, eles utilizaram 20 textos e identificaram desvios que não haviam sido marcados pelos corretores humanos.

Sistemas que buscam realizar o processamento automático de uma língua natural requerem uma série de ferramentas, como os *part-of-speech (POS) taggers* (etiquetadores morfossintáticos) e os *parsers* (analisadores sintáticos). Para que seja possível utilizar tais ferramentas e aplicações também para analisar redações escritas por estudantes, é necessário que elas sejam capazes de lidar com textos que apresentam desvios de escrita. Essa tarefa constitui um desafio para a área de processamento de língua natural (PLN), uma vez que esses desvios são de vários tipos (p.ex., pontuação, concordância, regência, crase, etc.), e a ocorrência de alguns desses tipos costuma ser pouco frequente.

Para subsidiar o desenvolvimento de tais aplicações, os *corpora* anotados são recursos importantes, pois permitem modelar computacionalmente os fenômenos e/ou as tarefas linguísticas, além de treinar e avaliar tais modelagens. Esses *corpora* podem então ser utilizados como base para estudos linguísticos desses fenômenos, bem como para a construção de ferramentas como *POS taggers* e *parsers* e para aplicações de apoio à escrita e de correção/avaliação automática de textos.

Neste artigo, descreve-se a anotação manual de desvios sintáticos em um *corpus* de redações nos moldes do ENEM, escritas por estudantes do ensino médio. O objetivo é caracterizar os desvios sintáticos desses textos de forma a gerar descrições linguísticas que possam subsidiar o refinamento do tratamento computacional dessas violações.

2. A construção do *corpus* de redações

A construção de *corpora* de aprendizes é útil para a análise da linguagem utilizada, a avaliação de ferramentas de PLN e o desenvolvimento de sistemas de correção de desvios gramaticais [Köhn e Köhn 2018]. Neste estudo, focam-se os desvios sintáticos em redações de falantes nativos de português do Brasil, mas aprendizes da modalidade escrita formal da língua. Para definir o conceito de “desvio sintático”, utiliza-se essa mesma modalidade, que também é adotada como critério de avaliação no ENEM, conforme consta na Cartilha do Participante [Brasil 2018]. Segundo a Cartilha, a avaliação das redações se divide em cinco competências, sendo que a Competência 1 avalia o domínio das convenções de escrita e a estrutura sintática, que deve estar adequada às regras gramaticais e à fluidez de leitura.

Da mesma maneira, as ferramentas de PLN também são desenvolvidas com base na modalidade escrita, a partir de uma perspectiva mais tradicional da língua. Logo, as análises linguísticas automáticas se apoiam, em grande medida, na gramática tradicional.

²<http://www.nilc.icmc.usp.br/scipo/>

Como esta tarefa de anotação leva em conta tanto as noções estabelecidas pelo ENEM quanto a abordagem das ferramentas de PLN, um desvio sintático é definido aqui como todo aquele relacionado a problemas na organização das palavras e suas combinações, de acordo com a modalidade escrita formal do português, podendo ser de ordem, de concordância e de dependência entre palavras.

Para analisar os desvios sintáticos presentes em textos de estudantes do ensino médio, construiu-se um *corpus* de 1.045 redações dissertativo-argumentativas (já em formato digital) nos moldes do ENEM, fornecidas pela empresa *Letrus*³, um centro de tecnologia e letramento que desenvolve ferramentas de escrita e avaliação de textos para escolas. O *corpus* possui 10.653 sentenças, totalizando 325.111 palavras e 184.967 *types* (palavras únicas).

A construção do *corpus* seguiu as etapas de [Aluísio e Almeida 2006]: i) projeto do *corpus* (seleção dos textos); ii) compilação, manipulação, nomeação dos arquivos; iii) anotação. Antes da anotação, procedeu-se à (i) limpeza do *corpus*, (ii) segmentação das redações em sentenças, visto que a anotação se deu em nível sentencial, e (iii) correção ortográfica (via MS Word[®]), a fim de manter o foco de atenção do anotador nos desvios sintáticos (foram corrigidos apenas os desvios ortográficos identificados automaticamente por essa ferramenta). O arquivo original foi preservado, efetuando-se todas as alterações em arquivo específico. Na sequência, selecionou-se a parcela do *corpus* que seria anotada (6.000 sentenças) e construiu-se uma diretriz de anotação, que engloba uma tipologia de desvios sintáticos, exemplos e orientações gerais e específicas⁴.

3. Metodologia de anotação de desvios sintáticos

Observando-se as questões de [Hovy e Lavid 2010] sobre anotação linguística, a tarefa aqui descrita teve início com a adaptação da tipologia de desvios gramaticais de [Pinheiro 2008] aos dados do *corpus*, resultando em 11 categorias e 27 subcategorias, organizadas conforme a Tabela 1.

Para evitar a sobreposição de categorias, estabeleceu-se uma regra hierárquica de anotação. Assim, a categoria crase, por exemplo, é hierarquicamente superior à regência, o que determinou que problemas de crase relacionados a regência fossem anotados apenas em uma das subcategorias de crase.

A escolha das categorias a serem anotadas se deu em função da noção de desvio e das alterações nas estruturas das árvores sintáticas das sentenças causadas pela presença de desvios. Nesse sentido, por exemplo, os desvios que alteram a classe morfossintática das sentenças (ou a interface entre a ortografia e a sintaxe) tornam a etiquetação automática via *POS taggers* difícil, já que essas ferramentas em geral consideram o contexto em que as palavras ocorrem para atribuir etiquetas. Assim, se a classe morfossintática de uma palavra é identificada equivocadamente devido a problemas de ortografia (p. ex. a ausência ou o excesso de acento no par *esta/está*), as etiquetações das palavras que a circundam provavelmente também terão problemas.

Da mesma forma, a presença ou ausência de determinadas palavras (tanto grama-

³<https://www.letrus.com.br/>

⁴Tanto o *corpus* como a diretriz de anotação poderão ser disponibilizados mediante contato com as autoras.

Tabela 1. Tipologia de desvios sintáticos

Categoría	Subcategoria	Descrição
01 - Pontuação	Ausência (pont-aus)	Ausência de pontuação em casos obrigatórios (p. ex. em adjuntos adverbiais deslocados).
	Excesso (pont-exc)	Ocorrência de pontuação em lugares não permitidos, como separação de sujeito e verbo com vírgula.
	Uso inadequado (pont-desv)	Uso inadequado de um sinal de pontuação no lugar de outro (p. ex. aglutiñações de sentenças por vírgulas).
02 - Crase	Ausência (crase-aus)	Falta de crase quando a sua ocorrência é obrigatória.
	Excesso (crase-exc)	Uso da crase quando ela não é permitida.
03 - Regência	Verbal (rege-verb)	Ausência, excesso ou uso inadequado de preposições quando o termo regente é um verbo.
	Nominal (rege-nom)	Ausência, excesso ou uso inadequado de preposições quando o termo regente é um substantivo, adjetivo ou advérbio.
04 - Concordância	Verbal (concor-verb)	Problemas de concordância entre sujeito e verbo.
	Nominal (concor-nom)	Problemas de concordância entre adjetivo, artigo, etc. e os termos a que se referem (substantivo ou pronome).
	Anafórica (concor-anaf)	Retomada equivocada de elementos citados na sentença, mas cujo retomador não concorda com o retomado.
05 - Pronomes	Colocação (pronom-col)	Colocação irregular dos pronomes em termos de posição na sentença (uso de próclise/ênclise).
	Ausência (pronom-aus)	Casos de ausência de qualquer tipo de pronome quando a sua ocorrência é obrigatória.
	Excesso (pronom-exc)	Ocorrência excessiva de pronome (p. ex. retomada do sujeito por meio de pronome pessoal).
	Uso inadequado (pronom-desv)	Utilização inadequada de pronomes, como uso de <i>cujo/cuja</i> sem valor de retomada.
06 - Preposições	Ausência (prepo-aus)	Ausência de preposição obrigatória não ligada a regência (p. ex. ausência de preposições em locução).
	Excesso (prepo-exc)	Excesso ou repetição de preposições (p. ex. <i>mediante a</i> ou <i>muitas das vezes</i>).
	Uso inadequado (determ-desv)	Uso inadequado de preposição ou contração (p. ex. uso de contração quando a estrutura exige a forma não contraída).
07 - Determinantes	Ausência (determ-aus)	Falta de determinante (p. ex. em casos de paralelismo obrigatório).
	Excesso (determ-exc)	Uso duplicado/excessivo de determinantes (p. ex. <i>cujo o</i>).
	Uso inadequado (determ-desv)	Ocorrência inadequada de determinantes (pouco frequente).
08 - Conjunções	Uso inadequado (conjunc)	Ausência, excesso ou uso inadequado de conjunções (p. ex. uso inadequado dos porquês, <i>mas porém</i>).
09 - Formas verbais	Uso equivocado de formas verbais (verbo-mod)	Ocorrência de desvios de formas, tempos, modos verbais (p. ex. uso de indicativo em vez de subjuntivo).
	Uso equivocado de formas nominais (verbo-nom)	Uso equivocado das formas nominais gerúndio, particípio e infinitivo.
10 - Segmentação	Segmentação inadequada de sentenças (segment)	Sentenças que foram segmentadas, mas deveriam estar ligadas à anterior (p. ex. que começem por <i>assim como</i>).
11 - Outros	Ordem (ordem)	Ordem equivocada, ausência ou excesso de palavras ou grupos de palavras de conteúdo (não abarcadas pelas demais categorias).
	Interface ortografia-sintaxe (orto-sin)	Problemas de ortografia que alterem a classe morfossintática da palavra, influenciando na sintaxe.
	Sem especificação (sem-espec)	Desvios que não se encaixem em nenhuma das categorias anteriores.

ticiais quanto de conteúdo) e os desvios ligados às formas verbais geram dificuldades para um *parser* encontrar as relações corretas entre os elementos de uma sentença. Portanto, uma vez que a anotação tem como objetivo gerar subsídios para o desenvolvimento e o aprimoramento de tais ferramentas, é importante que esses fenômenos façam parte do esquema de anotação.

Estabelecida a tipologia, a tarefa de anotação se deu em duas fases: classificação das sentenças em “com desvio” e “sem desvio”; e tipificação dos desvios presentes em parte das sentenças com desvio. Na primeira fase, classificaram-se 6.000 sentenças (56,3% do *corpus*) em “sem” e “com desvio”. Em um arquivo *xls* composto por três colunas, a classificação consistiu em apenas identificar se cada sentença possuía (ou não) ao menos um desvio sintático. As duas primeiras colunas do arquivo codificavam o ID de uma sentença e o respectivo texto, e a terceira coluna era a da anotação, que foi realizada por meio da atribuição de uma das *tags*: N (= sem desvio) e D (= com desvio).

Na segunda fase da anotação, 2.500 sentenças classificadas como D tiveram seus desvios categorizados conforme a tipologia, o que foi feito por meio da plataforma de anotação FLAT (*FoLia Linguistic Annotation Tool*) [Gompel e Reynaert 2013]. A caracterização consistiu em delimitar o segmento sentencial relativo ao desvio e anotá-lo com a etiqueta correspondente à sua subcategoria, seguindo as diretrizes de anotação. Os desvios caracterizados pela ausência de um elemento foram ancorados no *token* imediatamente anterior à posição em que o elemento ausente deveria ocorrer, com exceção dos casos de regência, cuja anotação foi associada ao termo regente.

4. Resultados da anotação

A Tabela 2 apresenta os resultados da primeira fase, contendo o número de sentenças com e sem desvio nas 6.000 sentenças anotadas (56,3% do *corpus*), e os respectivos percentuais.

Tabela 2. Número de sentenças com e sem desvio

	Nº sentenças	Percentual (%)
Contém desvio	4.409	73,48
Não contém desvio	1.591	26,52

A presença significativa de sentenças com desvio justifica a necessidade de tais descrições linguísticas. A segunda fase identificou os tipos de desvios das categorias e subcategorias mais e menos frequentes, chegando a um total de 7.290 desvios. Os desvios por categoria se distribuem como mostra a Tabela 3, por ordem de frequência.

Os desvios mais frequentes são os de pontuação e de concordância. As categorias menos frequentes são as de uso de conjunções e de determinantes. Em função da estrutura da tipologia, é preciso analisar também as subcategorias para que seja possível estabelecer melhor os padrões de desvios encontrados. A Tabela 4 mostra a distribuição dos desvios por subcategoria, ordenados por frequência.

Analizando as subcategorias, vê-se que há maior ocorrência de desvios de ausência de pontuação do que de excesso ou de uso inadequado desses sinais. Já nas subcategorias de concordância, os desvios de concordância verbal são quase duas vezes mais frequentes

Tabela 3. Distribuição dos desvios sintáticos por tipo: categoria

Categoría	Nº desvios
01 - Pontuação	3.224
04 - Concordância	1.378
09 - Formas verbais	500
05 - Pronomes	422
06 - Preposições	418
02 - Crase	312
10 - Segmentação	303
03 - Regência	250
11 - Outros	168
08 - Conjunções	167
07 - Determinantes	148

Tabela 4. Distribuição dos desvios sintáticos por tipo: subcategoria

Subcategoria	Nº desvios
01.2-pont-exc	726
01.3-pont-desv	614
04.2-concor-nom	455
10.1-segment	303
09.1-verbo-mod	300
05.4-pronom-desv	246
02.1-crasis-aus	229
03.1-rege-verb	208
09.2-verbo-nom	200
08.1-conjunc	167
06.2-prepo-exc	160
06.3-prepo-desv	142
11.2-orto-sin	135
06.1-prepo-aus	116
05.3-pronom-exc	106
02.2-crasis-exc	83
07.1-determ-aus	81
07.2-determ-exc	61
04.3-concor-anaf	48
03.2-rege-nom	42
05.1-pronom-col	36
05.2-pronom-aus	34
11.1-ordem	33
07.3-determ-desv	6
11.3-sem-espec	-

que as outras subcategorias de concordância. Além disso, como era esperado, a subcategoria de uso inadequado de determinantes foi a menos frequente (excetuando-se a de desvios sem especificação, que deveria ser usada apenas em casos que realmente não se inseriam em nenhuma das outras subcategorias definidas, e que não foi aplicada a nenhum desvio).

A segunda menos frequente foi a subcategoria relacionada à ordem de palavras (isto é, a estruturas cuja ordenação das palavras está equivocada), repetição, ausência ou excesso de palavras de conteúdo. Essa pouca ocorrência de problemas de ordem provavelmente está associada ao fato de as redações terem sido escritas por falantes nativos, que já têm internalizada a ordem adequada de elementos na sentença.

Vê-se que a subcategoria de uso de conjunções é a décima mais frequente, sendo que a categoria referente a ela estava entre as menos frequentes. Isso se dá porque a categoria é composta de uma única subcategoria, isto é, todos os desvios da categoria “08 - Conjunções” também são os que aparecem na respectiva subcategoria. O mesmo ocorre com a categoria “10 - Segmentação”.

5. Considerações finais

Neste artigo, apresentou-se a tarefa de anotação de desvios sintáticos presentes em redações de estudantes do ensino médio nos moldes do ENEM. O *corpus* anotado busca servir de subsídio para o desenvolvimento de ferramentas de PLN capazes de lidar com textos que tenham como característica a presença de desvios sintáticos. A partir da tipologia proposta e da metodologia de anotação estabelecida, observou-se que os textos do *corpus* apresentam muitos desvios, sendo que os mais frequentes são de pontuação (principalmente a sua ausência) e de concordância, com ênfase para a concordância verbal.

Agradecimento. À CAPES, pelo suporte financeiro.

Referências

- Aluísio, S. M. e Almeida, G. M. d. B. (2006). O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calídoscópio*, 4(3):156–178.
- Brasil (2018). *Redação no ENEM 2018: Cartilha do Participante*. INEP/MEC, Brasília.
- Castaldo, M. M. (2009). *Redação no vestibular: a língua cindida*. Tese (doutorado em educação), Universidade de São Paulo.
- Feltrim, V. D. (2004). *Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes Web de auxílio à escrita acadêmica em português*. Tese (doutorado em computação), Universidade de São Paulo.
- Gompel, M. v. e Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.
- Hovy, E. e Lavid, J. (2010). Towards a ‘science’ of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22:13–36.

- Köhn, C. e Köhn, A. (2018). An annotated corpus of picture stories retold by language learners. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 121–132, Santa Fe. Association for Computational Linguistics.
- Pinheiro, G. M. (2008). *Redações do ENEM: estudo dos desvios da norma padrão sob a perspectiva de corpos*. Tese (doutorado em linguística), Universidade de São Paulo, São Paulo.
- Santos, J., Paiva, R., e Bittencourt, I. I. (2016). Avaliação Léxico-Sintática de Atividades Escritas em Algoritmo Genético e Processamento de Linguagem Natural: Um Experimento no ENEM. *Revista Brasileira de Informática na Educação*, 24(2):92–107.
- Soni, M. e Thakur, J. S. (2018). A systematic review of automated grammar checking in english language. *ArXiv (submitted to Computational Linguistics)*, arXiv:1804.00540:1–23.

Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português

Sidney Evaldo Leal¹, Sandra Maria Aluísio¹,
Erica dos Santos Rodrigues²,
João Marcos Munguba Vieira³, Elisângela Nogueira Teixeira³

¹Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)

² Departamento de Letras - Pontifícia Universidade Católica do Rio de Janeiro (PUC)

³ Departamento de Letras Vernáculas - Universidade Federal do Ceará (UFC)

¹sidleal@gmail.com, ¹sandra@icmc.usp.br, ²ericasr@puc-rio.br
³joaomvieira@gmail.com, ³elisteixeira@leturas.ufc.br

Abstract. This paper presents a method for automating the process of choosing a short passages subset of a large corpus to be used in psycholinguistic research that investigates reading using eye-tracking. To show the method effectiveness, a corpus with 100 short passages of 3 textual genres was used to choose a smaller corpus with 50 passages, using clustering methods and 58 metrics of several linguistic levels. The groups resulting from clustering were evaluated by similarity criteria and the method proved to be useful in supporting the selection of material to be used in psycholinguistic studies.

Resumo. Este trabalho apresenta um método para automatização do processo de escolha de um subconjunto de parágrafos de grandes corpora a ser utilizado em pesquisas psicolinguísticas que investigam a leitura usando rastreamento ocular. Para mostrar a efetividade do método, foi utilizado um corpus com 100 parágrafos de 3 gêneros textuais para a escolha de um corpus menor com 50 parágrafos, via métodos de clusterização, usando 58 métricas linguísticas. Os grupos resultantes da clusterização foram avaliados com base em critérios de similaridade e o método mostrou-se útil para apoiar a seleção de material para estudos psicolinguísticos.

1. Introdução

Atualmente, corpora de rastreamento ocular são frequentemente utilizados no estudo de custos de processamento de estruturas linguísticas para, por exemplo, (i) avaliar modelos e métricas de dificuldade sintática [González-Garduño and Søgaard 2017], (ii) para melhorar ou avaliar modelos computacionais de simplificação via compressão sentencial [Klerke et al. 2016] e (iii) avaliar a qualidade da tradução automática com métricas objetivas [Klerke et al. 2015]. No entanto, existem poucos destes recursos, para um pequeno número de idiomas, por exemplo, inglês [Luke and Christianson 2017, Cop et al. 2016], inglês e francês [Kennedy et al. 2003], alemão [Kliegl et al. 2004] e russo [Laurinavichyute et al. 2018].

Para o português do Brasil, o rastreamento ocular já é utilizado há algum tempo nas pesquisas da área de Psicolinguística. Por exemplo, [Maia et al. 2007]

utilizaram para investigar o papel do processamento morfológico na identificação de palavras; [Leitão et al. 2012] utilizaram na investigação do processamento anafórico; [da Silva e Forster 2013] investigou o processamento incremental de orações relativas restritivas de objeto; e [Teixeira et al. 2014] para evidenciar o custo de resolução de pronomes nulos e plenos. Entretanto, não há nenhum grande corpus do português, publicamente disponível, com dados de rastreamento ocular de jovens adultos e com normas de previsibilidade para a tarefa de leitura silenciosa. Essa é uma grande lacuna que restringe as possibilidades de pesquisa nas áreas de Psicologia Cognitiva, Psicolinguística e Processamento de Línguas Naturais (PLN).

Pesquisas na área de Psicolinguística, especificamente de processamento da sentença, podem se beneficiar de corpora de textos autênticos linguisticamente anotados, que permitem fazer uma correlação dos tempos de leitura com fenômenos linguísticos, por exemplo os elencados abaixo:

1. complexidade estrutural do período (períodos simples vs. compostos);
2. transitividade verbal;
3. animacidade do sujeito e do objeto;
4. tipos de sentenças (ativas/passivas/relativas);
5. mecanismos de construção de relações de correferência, entre outros.

Nos experimentos psicolinguísticos, estímulos são construídos para examinar o efeito de fatores/variáveis independentes no comportamento do participante e, assim, poder investigar hipóteses de trabalho. Uma crítica muitas vezes feita a esses trabalhos diz respeito ao nível de naturalidade dos estímulos experimentais, com consequências em termos do grau de validade ecológica das pesquisas. Assim, projetos atuais utilizam textos autênticos, envolvendo diferentes gêneros textuais (jornalísticos, científicos, literários, etc.), para permitir uma avaliação da influência conjugada de um conjunto de fatores linguístico-textuais que podem afetar o processamento linguístico durante a leitura, em condições menos artificiais de realização da tarefa. Esses corpora são compilados para trazerem uma rica diversidade de fenômenos linguísticos, como, por exemplo, os cinco tipos de descrição das estruturas sintáticas elencados acima, para que se possa correlacionar a diversidade destes fenômenos com tempos de leitura, o comportamento durante a leitura e a avaliação de modelos complexos de controle dos movimentos dos olhos durante a leitura (por exemplo, o E-Z reader - modelo de processamento lexical serial [Reichle et al. 2006] - e o Swift - modelo de processamento lexical paralelo [Engbert et al. 2002]), implementados por simulações computacionais.

Entretanto, uma dificuldade para a compilação desses corpora é a anotação manual destes fenômenos, que idealmente deveria usar mais de um anotador para a avaliação do nível de concordância entre eles [Carletta 1996]. De posse deste corpus anotado com os fenômenos, se pode escolher aquele subconjunto com os atributos variados dos fenômenos linguísticos para a adequação do estudo. Por exemplo, os parágrafos da Figura 1 são do gênero de divulgação científica e jornalístico, respectivamente, e apresentam o mesmo número de sentenças, mas eles diferem em vários níveis linguísticos, por exemplo, na complexidade de seu léxico, na complexidade sintática e tamanho das sentenças, no nível de formalidade.

Dada a disponibilização pública de várias métricas automáticas para avaliação da coesão e coerência de textos escritos ou falados para a língua portuguesa

([Scarton and Aluísio 2010]; [Aluísio et al. 2016]), várias métricas, além do número de sentenças, poderiam ser analisadas para que a escolha dos parágrafos seja adequada para uma dada pesquisa em psicolinguística.

Figura 1. Parágrafos de gêneros diferentes, com mesmo número de sentenças

- a** Nos últimos tempos, o crescente desenvolvimento da genética tem suscitado discussões acaloradas. Os críticos descrevem um cenário aterrador: a criação de uma sociedade homogênea, a perda de privacidade, a ameaça à própria condição humana. Os avanços em neurotecnologia - que permitem manipular o cérebro e modular as emoções - levantam questões éticas e legais da mesma natureza e gravidade que os da genética. Até recentemente, a maioria das experiências com cérebro humano não era considerada ética.
- b** Você já imaginou a sua vida sem ouvir nenhuma palavra? Vinte milhões de brasileiros têm alguma dificuldade para ouvir. Quanto mais cedo o problema for descoberto, maiores são as chances de cura. Para isso, o teste da orelhinha é fundamental.

Fontes: (a) Revista Pesquisa Fapesp¹ e (b) Globo Comunicação e Participações S.A.²

Esta pesquisa apresenta um método para automatização do processo de escolha de um subconjunto, tomado de um grande corpus de parágrafos para pesquisas que utilizam rastreamento ocular durante a leitura destes parágrafos. Ela é parte integrante do projeto RastrOS³.

Para mostrar a efetividade do método, utilizamos, como exemplo, um corpus com 100 parágrafos de três gêneros (jornalístico, divulgação científica e literário) (Seção 3), para a escolha de um subcorpus que traga 50 parágrafos, sendo 35 dos gêneros jornalístico e literário e 15 de divulgação científica, via métodos de clusterização, detalhados na Seção 2. O método proposto faz uso de um grande conjunto de métricas de vários níveis linguísticos (Seção 4), disponíveis publicamente na Plataforma Simpligo (<https://simpligo.sidle.al/nlcmatrixdoc>). Particularmente, foram escolhidas 58 métricas, agrupadas em quatro conjuntos; três destes conjuntos – tipos de sentenças (7 métricas), complexidade da estrutura sintática (22 métricas) e análise de correferência (8 métricas) foram escolhidos para modelar diretamente os três estudos de comparação dos tempos de leitura abaixo: (i) complexidade estrutural do período (períodos simples vs. compostos); (ii) tipos de sentenças (ativas/passivas/relativas); (iii) mecanismos de construção de relações de correferência, entre outros. E o conjunto denominado morfossintaxe (21 métricas) foi escolhido para modelar indiretamente os estudos sobre transitividade verbal e animacidade do sujeito e do objeto. Finalmente, a Seção 5 mostra o conjunto de agrupamentos resultante, juntamente com métodos para avaliar sua qualidade.

2. Aprendizado de Máquina e Métodos de Clusterização

Inicialmente, a área de Inteligência Artificial (IA) era considerada uma área teórica, mas nas últimas décadas com o crescimento do volume de dados e complexidade de problemas que necessitam de tratamento computacional, as técnicas de Aprendizagem de Máquina (AM) começaram a se destacar [Faceli et al. 2011]. Elas são boas ferramentas na criação

¹<https://revistapesquisa.fapesp.br/2002/07/01/manipuladores-de-cerebros/>

²<https://g1.globo.com/bemestar/noticia/mais-de-20-milhoes-de-brasileiros-tem-alguma-dificuldade-para-escutar.ghtml>

³Um grande corpus com medidas de RASTreamento Ocular e normas de previsibilidade durante a leitura de estudantes do ensino Superior no Brasil - <http://www.nilc.icmc.usp.br-nilc/index.php/rastros>

de hipóteses (ou funções) a partir da experiência passada, para predizer respostas ou descrever dados dos problemas que se deseja tratar. Hoje são utilizadas em tarefas tão diversas quanto reconhecimento de fala, detecção de fraudes financeiras, condução autônoma de automóveis, diagnóstico de doenças, dentre outras.

Dentro da AM, existem algoritmos que procuram identificar padrões ou tendências relevantes em conjuntos de dados sem necessidade de um elemento externo servindo de guia do aprendizado. Essas técnicas são chamadas de aprendizagem não supervisionada. Destas, as de clusterização (ou agrupamento) são de especial interesse deste trabalho, pois permitem analisar um grande número de métricas e dados, gerando sugestões de grupos por afinidade.

Os algoritmos dessas técnicas geralmente são classificados em [Faceli et al. 2011]:

- **Baseados em centróides:** Otimizam o critério de agrupamento de forma iterativa, procurando minimizar o erro quadrático ou variação dentro do *cluster*.
- **Hierárquicos:** Geram uma sequência de partições aninhadas a partir de uma matriz de proximidade. Podem ser do tipo **aglomerativo**, que começa com um grupo para cada objeto e vai combinando, ou **divisivo**, que começa com um único grupo e vai dividindo sucessivamente.
- **Baseados em densidade:** Assumem que cada *cluster* é uma região de alta densidade de objetos, separada das demais por regiões de baixa densidade.
- **Baseados em grafos:** Os dados são representados em um grafo de proximidade, no qual cada nó representa um objeto e as arestas, a similaridade ou distância.
- **Baseados em redes neurais:** Sistemas paralelos compostos de unidades simples de processamento; por exemplo o algoritmo SOM (*Self-Organizing Map*).
- **Baseados em Grid:** Define um *grid* (reticulado) para o espaço de dados. Muito eficiente para grandes conjuntos de objetos.

O algoritmo mais simples e mais utilizado é o K-Means⁴ que utiliza técnica baseada em centróides. Nesta pesquisa, além do K-means, também foram avaliados dois outros algoritmos – o AgglomerativeClustering⁵, do tipo hierárquico e o DBScan⁶, baseado em densidade. Este trabalho utilizou a implementação deles na biblioteca scikit-learn em python. O DBScan não teve bons resultados no nosso cenário devido ao tamanho e distribuição do conjunto de dados.

3. Conjunto de dados separados por gêneros de texto

Foram selecionados manualmente 100 parágrafos de três gêneros e várias fontes, procurando incluir uma boa amostra para abranger o máximo dos fenômenos do português brasileiro escrito. Os parágrafos do gênero jornalístico foram obtidos de portais de notícias bem conhecidos como G1, Metro, BBC, Reuters, Terra, Estadão, Folha de São Paulo, Jornal da USP, dentre outros. Os parágrafos do gênero literário vieram de romances em domínio público. Os parágrafos de divulgação científica vieram das fontes: Revista Pesquisa Fapesp, Galileu, Aventuras na História, Época, Exame, Isto é, caderno ciência e

⁴<https://scikit-learn.org/stable/modules/clustering.html#k-means>

⁵<https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>

⁶<https://scikit-learn.org/stable/modules/clustering.html#dbSCAN>

tecnologia do Jornal do Brasil, Mente e Cérebro, National Geographic Brasil, Piauí, Scientific American Brasil, dentre outras.

A distribuição dos parágrafos pode ser vista na Tabela 1. O objetivo deste trabalho foi selecionar, dentre os 100 parágrafos, um subconjunto, com 50 parágrafos, que mantivesse a maior variância possível dos fenômenos da língua, relacionados com os cinco estudos de comparação dos tempos de leitura, descritos na Seção 1.

Tabela 1. Distribuição dos parágrafos por gênero

Gênero	Quantidade disponível	Alvo da seleção
Jornalístico	43	35
Literário	9	
Divulgação científica	48	15
Total	100	50

4. Métricas Selecionadas

Para representar cada parágrafo do conjunto de dados, foram escolhidas 58 métricas calculadas com o apoio da ferramenta NILC-Metrix⁷. Essas métricas foram agrupadas em 4 conjuntos, resultando em 22 sobre complexidade estrutural/sintática (e.g. períodos simples vs compostos), 7 com tipos de orações (e.g.ativas/passivas, relativas), 8 com mecanismos de construção de relações de correferência e 21 relacionadas com a morfossintaxe (e.g. categorias gramaticais e flexão de substantivos e verbos); a lista completa pode ser vista na Tabela 2.

5. Método para Escolha de Subconjuntos via Clusterização e Avaliação

Após selecionar as métricas, os três conjuntos de parágrafos foram processados e foram executados diversos experimentos, buscando a melhor divisão de grupos, dentro de cada conjunto. Os melhores resultados foram obtidos utilizando a técnica chamada “Método do Cotovelo”⁸ (do inglês *Elbow Method*) para encontrar o número ideal de agrupamentos. Esta técnica simula diversas divisões em número crescente de grupos e calcula as variâncias internas de cada grupo, buscando o ponto de equilíbrio [Dangeti 2017]. O gráfico com o cálculo do “cotovelo” para o gênero jornalístico pode ser visto na Figura 2, com o título “Cotovelo Kmeans”, no exemplo ele indica 7 grupos ótimos, que foram plotados no gráfico com título “Grupos”, com números de 0 a 6.

5.1. Redução de Dimensionalidade via Análise de Componentes Principais

Outra técnica utilizada para melhorar os resultados dos experimentos foi a Análise de Componentes Principais ou PCA (do inglês *Principal Component Analysis*). PCA é um procedimento matemático que cria novas métricas (ou variáveis) que são uma combinação linear das métricas originais e é utilizado para reduzir a dimensionalidade dos dados, sendo aplicado como uma etapa de pré-processamento antes de métodos de clusterização, como os apresentados na Seção 2. Ele permite visualizar os dados no espaço (cf. na Figura 2, os gráficos com títulos “Gênero Jornalístico” e “Textos - por índice”) e também melhorar a generalização dos algoritmos [Dangeti 2017].

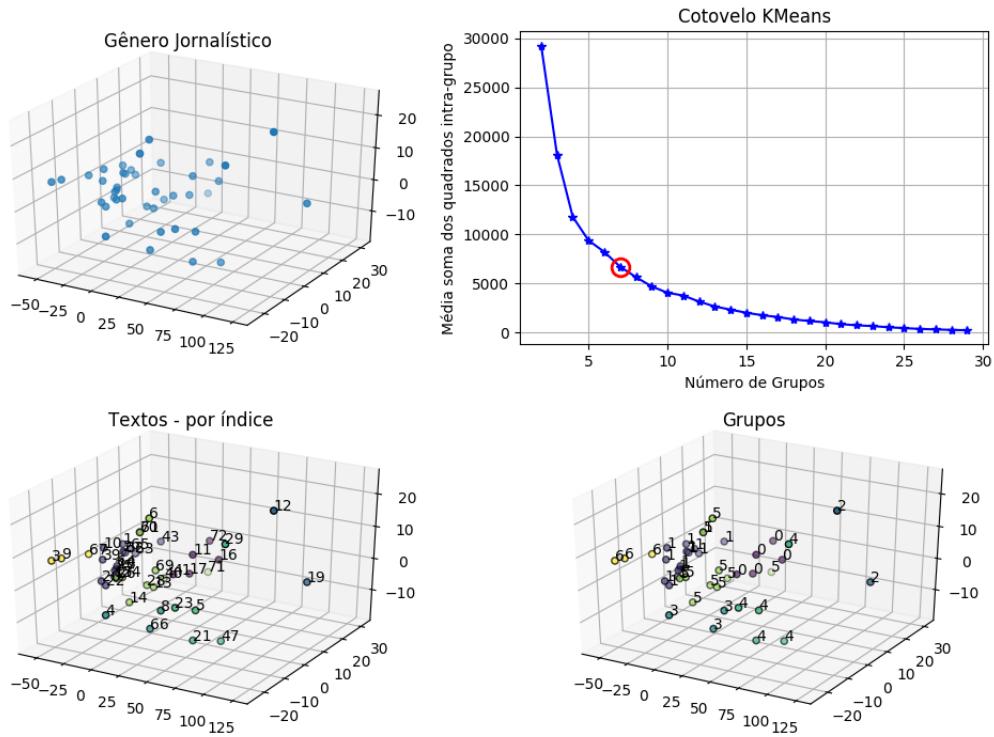
⁷<https://simpligo.sidle.al/nilcmetrix>

⁸A denominação vem do fato que se o gráfico relembrar um braço, então o “cotovelo” (ponto de inflexão da curva) é uma boa indicação de que o modelo subjacente se encaixa melhor naquele ponto.

Tabela 2. Lista de todas as métricas utilizadas.

Nome	Descrição
Complexidade Estrutural	
words_per_sentence	Média de palavras por sentença
sentences	Quantidade de sentenças no parágrafo
words	Quantidade de palavras no parágrafo
sentence_length_max	Quantidade máxima de palavras por sentença
sentence_length_min	Quantidade mínima de palavras por sentença
sentence_length_std	Desvio padrão da quantidade de palavras por sentença
yngve	Complexidade sintática de Yngve (árvores sintáticas fora do padrão de ramificação à direita)
frazier	Complexidade sintática de Frazier (baseada na profundidade das árvores sintáticas)
dep_distance	Distância na árvore de dependência
words_before_main_verb	Quantidade média de palavras antes dos verbos principais das orações principais das sentenças
clauses_per_sentence	Quantidade média de orações por sentença
sentences_with_zero_clause	Proporção de sentenças sem verbos em relação a todas as sentenças do parágrafo
sentences_with_one_clause	Proporção de sentenças com uma oração em relação a todas as sentenças do parágrafo
sentences_with_two_clauses	Proporção de sentenças com duas orações em relação a todas as sentenças do parágrafo
sentences_with_three_clauses	Proporção de sentenças com três orações em relação a todas as sentenças do parágrafo
sentences_with_four_clauses	Proporção de sentenças com quatro orações em relação a todas as sentenças do parágrafo
sentences_with_five_clauses	Proporção de sentenças com cinco orações em relação a todas as sentenças do parágrafo
sentences_with_six_clauses	Proporção de sentenças com seis orações em relação a todas as sentenças do parágrafo
sentences_with_7+_clauses	Proporção de sentenças com sete ou mais orações em relação a todas as sentenças do parágrafo
punctuation_diversity	Proporção de <i>types</i> de pontuações em relação à quantidade de <i>tokens</i> de pontuações no parágrafo
punctuation_ratio	Proporção de sinais de pontuação em relação à quantidade de palavras do parágrafo
non_svo_ratio	Proporção de orações que não estão no formato SVO (sujeito-verbo-objeto) em relação a todas as orações
Tipos de orações	
passive_ratio	Proporção de orações na voz passiva analítica em relação à quantidade de orações do parágrafo
relative_clauses	Proporção de orações relativas em relação à quantidade de orações do parágrafo
relative_pronouns_div_ratio	Proporção de <i>types</i> de pronomes relativos em relação à quantidade de <i>tokens</i> de pronomes relativos
subordinate_clauses	Proporção de orações subordinadas pela quantidade de orações do parágrafo
infinite_subordinate_clauses	Proporção de orações subordinadas reduzidas pela quantidade de orações do texto
coordinate_conj_per_clauses	Proporção de conjunções coordenativas em relação a todas as orações do texto
apposition_per_clause	Quantidade média de apostos por oração do texto
Correferência	
adjacent_refs	Média das proporções de candidatos a referentes na sentença anterior em relação aos pronomes pessoais do caso reto nas sentenças
anaphoric_refs	Média das proporções de candidatos a referentes nas cinco sentenças anteriores em relação aos pronomes anafóricos das sentenças
arg_owl	Quantidade média de referentes que se repetem nos pares de sentenças do texto
adj_arg_owl	Quantidade média de referentes que se repetem nos pares de sentenças adjacentes
stem_owl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças
adj_stem_owl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes
adj_cw_owl	Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes
coreference_pronoun_ratio	Média de candidatos a referente, na sentença anterior, por pronome anafórico do caso reto
Morfossintáticas	
verbs	Proporção de verbos em relação à quantidade de palavras do parágrafo
verbs_max	Proporção máxima de verbos por palavras em relação à quantidade de palavras das sentenças
verbs_min	Proporção mínima de verbos por palavras em relação à quantidade de palavras das sentenças
verbs_standard_deviation	Desvio padrão das proporções entre verbos e a quantidade de palavras das sentenças
verbal_time_moods_diversity	Quantidade de diferentes tempos-modos verbais que ocorrem no texto
adverbs	Proporção de advérbios em relação à quantidade de palavras do texto
adverbs_max	Proporção máxima de advérbios em relação à quantidade de palavras das sentenças
adverbs_min	Proporção mínima de advérbios em relação à quantidade de palavras das sentenças
adverbs_standard_deviation	Desvio padrão das proporções entre advérbios e a quantidade de palavras das sentenças
noun_ratio	Proporção de substantivos em relação à quantidade de palavras do parágrafo
nouns_max	Proporção máxima de substantivos em relação à quantidade de palavras das sentenças
nouns_min	Proporção mínima de substantivos em relação à quantidade de palavras das sentenças
nouns_standard_deviation	Desvio padrão das proporções entre substantivos e a quantidade de palavras das sentenças
pronoun_ratio	Proporção de pronomes em relação à quantidade de palavras do parágrafo
pronouns_max	Proporção máxima de pronomes em relação à quantidade de palavras das sentenças
pronouns_min	Proporção mínima de pronomes em relação à quantidade de palavras das sentenças
pronouns_standard_deviation	Desvio padrão das proporções entre pronomes e a quantidade de palavras das sentenças
adjective_ratio	Proporção de adjetivos em relação à quantidade de palavras do parágrafo
adjectives_standard_deviation	Desvio padrão das proporções entre adjetivos e a quantidade de palavras das sentenças
preposition_diversity	Proporção de <i>types</i> de preposições em relação à quantidade de <i>tokens</i> de preposições
syllables_per_content_word	Quantidade média de sílabas por palavra no parágrafo

Figura 2. Visualização dos parágrafos e grupos do gênero jornalístico.



5.2. Avaliação do Método

Neste trabalho, os agrupamentos foram gerados utilizando o algoritmo K-Means e AgglomerativeClustering, em seguida foram calculadas as medidas de silhueta (*Silhouette*) para os grupos e *V-Measure* [Rosenberg and Hirschberg 2007] para medir a concordância entre os dois algoritmos. Os resultados podem ser vistos na Tabela 3. A silhueta mede o quanto similar é um objeto em seu grupo, em comparação com os demais grupos, e varia de -1 a +1. No nosso cenário, o valor médio 0,38 pode ser considerado bom, tendo em vista que os parágrafos já possuem certa similaridade pela seleção prévia (parágrafos curtos). Já a *V-Measure* obtida reforça que os algoritmos concordam com a divisão dos objetos nos grupos em mais de 90%. A Homogeneidade (*Homogeneity*) avalia se cada grupo contém somente membros de uma única classe, a Completude (*Completeness*) avalia se todos os membros de uma classe estão no mesmo grupo, sendo a *V-Measure* a média harmônica entre elas duas.

Tabela 3. Resultados

Gênero	Número de Grupos	Itens por grupo Med (Min-Max)	K-Means Silhouette	Agglomerative Silhouette	Homogeneity	Completeness	V-Measure
Jornalístico	7	7 (2-14)	0,38	0,38	0,93	0,92	0,92
Literário	4	2 (1-4)	0,39	0,39	1,00	1,00	1,00
Divulgação científica	7	5 (2-15)	0,38	0,35	0,82	0,79	0,81
Média	6	11 (1,6-4,6)	0,38	0,37	0,92	0,90	0,91

A Tabela 1 mostra os alvos de seleção para montar o corpus de 50 parágrafos a partir do corpus inicial de 100 parágrafos (35 parágrafos dos gêneros jornalísticos e literários

e 15 do gênero de divulgação). O experimento realizado selecionou 7, 4 e 7 grupos (cf. Tabela 3). Assim, o trabalho final para montar o corpus de pesquisa pode ser realizado pela escolha manual, apoiada por algum critério importante para a pesquisa, como o tamanho dos parágrafos. No exemplo deste artigo, de 11 grupos serão selecionados 35 parágrafos e de 7 grupos de divulgação, os 15 parágrafos finais.

6. Considerações finais

A possibilidade de selecionar textos com características linguísticas específicas pode ser muito relevante em estudos de natureza experimental na área de psicolinguística. A contribuição desta pesquisa com um método de clusterização que atende esse propósito se mostrou bastante eficiente, pois o conjunto de métricas automáticas ajudou a agrupar parágrafos com características semelhantes, realizando uma anotação dirigida ao agrupamento. Com a lista dos parágrafos agrupados, a tarefa de selecionar manualmente a amostra final tornou-se bem mais simples e informada. É possível selecionar um número de itens de cada grupo, de forma aleatória, ou com alguma forma de ranqueamento (maiores ou menores parágrafos, por exemplo).

Acreditamos que a utilização dos recursos de PLN e Aprendizagem de Máquina não supervisionada podem ajudar bastante em tarefas trabalhosas como a anotação manual dos fenômenos da língua em um corpus. Como continuação deste trabalho, os autores pretendem disponibilizar uma ferramenta web com o método apresentado, para automatizar a análise e permitir que outros pesquisadores consigam replicar o experimento sem esforço de codificação.

7. Agradecimentos

À Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, processo número 2019/09807-0, pelo apoio financeiro.

Referências

- Aluísio, S. M., Cunha, A., Toledo, C., and Scarton, C. (2016). Computational tool for automated language production analysis aimed at dementia diagnosis. In *International Conference on Computational Processing of the Portuguese Language, Demonstration Session*.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguit.*, 22(2):249–254.
- Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2016). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.
- da Silva e Forster, R. A. M. (2013). *Aspectos do Processamento de Orações Relativas: Antecipação de Referentes e Integração de Informação Contextual*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).
- Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing, E-Book.
- Engbert, R., Longtin, A., and Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621 – 636.

- Faceli, K., Lorena, A. C., Gama, J., and de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. LTC - Livros Técnicos e Científicos, Rio de Janeiro.
- González-Garduño, A. V. and Søgaard, A. (2017). Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.
- Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. *Proceedings of the 12th European conference on eye movement*.
- Klerke, S., Castilho, S., Barrett, M., and Søgaard, A. (2015). Reading metrics for estimating task efficiency with mt output. In *Conference on Empirical Methods in Natural Language Processing*, pages 6–13. Association for Computational Linguistics.
- Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.
- Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16, pages 262–284.
- Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., and Klieg, R. (2018). Russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior Research Methods*, pages 1–18.
- Leitão, M. M., Ribeiro, A. J. C., and Maia, M. (2012). Penalidade do nome repetido e rastreamento ocular em português brasileiro. *Revista Lingüística*, v8 n2.
- Luke, S. G. and Christianson, K. (2017). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*.
- Maia, M., Lemle, M., and França, A. I. (2007). Efeito stroop e rastreamento ocular no processamento de palavras. *Ciências e Cognição* 2007, 12:02–17.
- Reichle, E. D., Pollatsek, A., and Rayner, K. (2006). E-z reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cogn. Syst. Res.*, 7(1):4–22.
- Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL)*, pages 410–420.
- Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.
- Teixeira, E. N., Fonseca, M. C. M., and Soares, M. E. (2014). Resolução do pronome nulo em português brasileiro: Evidência de movimentação ocular. *VEREDAS: Sintaxe das Línguas Brasileiras*, 18.

(Re)começando a discutir as locuções verbais

Elvis de Souza, Cláudia Freitas

Departamento de Letras
PUC-Rio – Brasil

elvis.desouza99@gmail.com, claudiafreitas@puc-rio.br

Abstract. *The need to annotate every word of every sentence in a corpus sheds light on grammatical issues not always highlighted in the linguistic description or which discussions do not come to definitive proposals. In this paper, we gather the postulations of [Câmara Júnior 1992], [Vilela and Koch 2001] and [Bagno 2012] on the issue of verbal phrases to assist us in grammar annotation of constructions such as [estar a V_{infinitive}], with an emphasis on the “a”, “de” and “para” particles that can figure in the center of the expressions. Despite some discrepancies between the authors’ observations, we eventually outline a treatment for the morphosyntactic annotation task.*

Resumo. A necessidade de anotar todas as palavras de todas as sentenças em um corpus joga luz sobre questões gramaticais nem sempre destacadas na descrição linguística ou cujas discussões não chegam a propostas definitivas. Neste trabalho, reunimos as postulações de [Câmara Júnior 1992], [Vilela and Koch 2001] e [Bagno 2012] sobre a questão das locuções verbais para nos auxiliar na anotação gramatical de construções como [começar a V_{infinitivo}], com ênfase nas partículas “a”, “de” e “para” que podem figurar no centro das expressões. A despeito de algumas divergências entre as observações dos autores, esboçamos, no final, um tratamento para a tarefa de anotação morfossintática.

1. Introdução

A anotação de corpora nos confronta com desafios novos a cada sentença. Assim ocorre no processo de revisão da anotação do corpus Bosque-UD [Rademaker et al. 2017], a versão em Universal Dependencies [Nivre et al. 2016] do corpus Bosque, que é parte integrante do projeto Floresta Sintá(c)tica [Afonso et al. 2002], com textos jornalísticos em português do Brasil e de Portugal.

O Universal Dependencies é um projeto de anotação de treebanks multilíngue, cujo objetivo é facilitar o desenvolvimento de parsers multilíngues. De um ponto de vista linguístico, a ideia é que o modelo possa ser compreendido e utilizado por não linguistas, e por isso as análises propostas se aproximam, em grande medida, daquelas das gramáticas tradicionais. Como em qualquer treebank, um corpus em UD deve conter, entre outros atributos morfossintáticos, uma classificação gramatical (substantivo, verbo, adjetivo etc.) e uma classificação sintática (objeto, sujeito, adjuntos etc.) para todas as palavras em todas as sentenças que compõem o corpus.

Ao longo do processo de revisão da anotação, um dos desafios com que nos deparamos foi o de anotar construções do tipo [estar a V_{infinitivo}]. Além da tarefa de decidir

a classe gramatical e a função sintática dos dois verbos, precisamos realizar a anotação também da partícula “a” entre a forma “estar” e o verbo no infinitivo que a segue. Comparativamente, precisamos decidir se tais ocorrências devem receber um tratamento igual ao de casos como os das assim nomeadas locuções verbais aspectuais – [acabar de V_{infinitivo}] – e as locuções verbais modais – [querer V_{infinitivo}]. No corpus, atualmente, as primeiras são tratadas como um caso de locução (*aux*) e, as segundas, como um caso de subordinação entre orações, em que a segunda é uma oração substantiva objetiva direta reduzida de infinitivo (*xcomp*) da segunda (figuras 1 e 2).

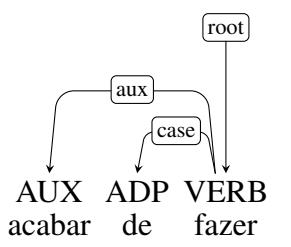


Figura 1. A anotação de *acabar de fazer* no Bosque-UD 2.4

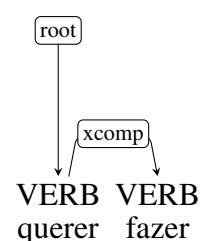


Figura 2. A anotação de *querer fazer* no Bosque-UD 2.4

Na gramática UD, a anotação sintática das partículas “de”, “a” e “para” entre dois verbos pode variar drasticamente caso consideremos que 1) elas fazem parte de uma locução verbal, indicando que os dois verbos estariam sendo unidos pela partícula e as três palavras formam uma unidade, ou 2) elas iniciam uma oração, sendo que o segundo verbo é, via de regra, complemento do primeiro verbo, que não é auxiliar. No caso das locuções verbais de tempo composto – [ter/haver V_{partícipio}] – há consenso nas diretrizes do projeto (e nas gramáticas) de que se trata de uma unidade verbal – e o verbo ter/haver deve ser anotado como de função sintática “auxiliar” e deve ser dependente do segundo verbo, indicando uma unidade verbal. Por outro lado, se construções como “gostar de cantar” devem ser consideradas uma locução ou uma subordinação entre orações não está claro. Especificamente, se a palavra “de” deve ser anotada como parte de uma locução verbal modal, ou como conjunção subordinativa (iniciando uma oração reduzida de infinitivo que complementa o primeiro verbo) é um ponto de discussão.

A fim de sustentar nossa decisão, lançamos mão dos dicionários e gramáticas específicos do português, de tal modo que o corpus seja anotado de maneira linguisticamente embasada, mas sem deixar de manter uma proximidade com abordagens das chamadas gramáticas tradicionais.

Tendo esse objetivo em mente, buscamos em [Câmara Júnior 1992], [Vilela and Koch 2001] e [Bagno 2012] o que se postulou sobre o assunto, dados que sistematizaremos na seção 2, apontando também algumas lacunas e divergências entre os autores. Por fim, na seção 3, pensaremos em algumas soluções para nossas questões a partir do que observamos nas gramáticas e tentaremos encarar a anotação de algumas sentenças complexas de nosso corpus com base nelas.

2. Revisão da literatura

As construções [estar a V_{infinitivo}] estão presentes apenas na parte portuguesa do corpus Bosque-UD, totalizando 80 sentenças, como em (1) e (2). Há também uma sentença do tipo [estar para V_{infinitivo}], na parte brasileira (3).

1. CP3-4 – «**Estamos a dotar** os computadores de um novo sentido» disse Steve d'Averio, director de marketing para a Europa da Logitech.
2. CP285-4 – António Guterres foi o primeiro convidado de uma série de debates com líderes políticos que o Inesc **está a promover**.
3. CF835-12 – Cursando economia na Faap, Kiko espera ansioso o seu telefone celular, que «**está para sair**».

Para dar direcionamento às questões, precisamos procurar referência em diferentes seções das gramáticas, sendo que alguns fenômenos são interpretados e nomeados diferentemente entre os autores. Entre as seções que contemplam o que procuramos, estão *preposição*, *verbo auxiliar*, *locução verbal*, *verbo modal* e *conjunção*.

2.1. Sobre as locuções verbais

Na *Gramática da Língua Portuguesa* [Vilela and Koch 2001], os autores enquadram as construções [estar a/para V_{infinitivo}] em duas categorias simultaneamente, sem distinção: verbos copulativos e verbos auxiliares de aspecto. Ambas as categorias estão dentro da seção *Verbos plenos e verbos auxiliares*, de tal modo que o primeiro verbo se configura como um verbo auxiliar, e o segundo, pleno. Segundo os autores, portanto, nas construções acima (1)-(3), a forma *estar* é a forma:

“(...) em que o peso gramatical é preponderante, ou porque o verbo se deslexicalizou e reforçou o seu peso gramatical (gramaticalizando-se) e necessita de um verbo pleno para poder funcionar como predicado ou porque o núcleo predutivo é constituído por um nome (*ter consideração por*), por um adjetivo (*ser inteligente*)” ([Vilela and Koch 2001], p. 72).

Por outro lado, o verbo no infinitivo nas construções (1)-(3) é verbo pleno, o que ocorre quando:

“(...) o conteúdo se dirige diretamente para a configuração da processualidade existente no mundo extralingüístico e que gramaticalmente pode funcionar como predicado da frase sem qualquer apoio ou suporte” ([Vilela and Koch 2001], p. 72).

Dentro da categoria de auxiliares de aspecto, esta que contém a construção em foco [estar a/para V_{infinitivo}], há também as construções do tipo [começar/continuar a V_{infinitivo}]. É possível concluir, portanto, que, de um ponto de vista formal, as construções são análogas e devem receber o mesmo tratamento morfossintático. Ainda dentro da seção de verbos auxiliares, Vilela e Koch inserem os verbos auxiliares de tempos compostos [ter/haver V_{particípio}] e auxiliares de modo [querer¹ V_{infinitivo}].

Embora as construções com o “estar” sejam mais marcadamente auxiliares, em outras construções, entretanto, julgar se um verbo está mais ou menos gramaticalizado, isto é, se funciona ou não como um verbo auxiliar em um dado contexto não é tarefa

¹São alguns dos poucos verbos inseridos na categoria *auxiliares de modo*, junto com [ter de/que, dever e poder V_{infinitivo}].

simples. Depende, por exemplo, de uma análise que compare o uso/função do verbo em uma sentença com o uso/função, do mesmo verbo, em outros contextos, que evidenciem o chamado sentido pleno desse verbo. A categorização como verbo pleno ou auxiliar pode, portanto, divergir para diferentes anotadores, assim como diverge para diferentes gramáticas.

Vejamos, como exemplo, a sentença (1) abaixo. Nela, o verbo “começar” está, indubitavelmente, exercendo sua função plena: é predicado verbal da sentença cujo sujeito é “A corrida sucessória”. No entanto, nas frases (2) e (3), embora as gramáticas nos digam que as formas de “começar” indicam apenas o aspecto do segundo verbo – aspecto inceptivo –, aceitamos como absolutamente possível uma leitura que assume o sentido pleno de “começar”, assim como em (1). Conseguimos, inclusive, parafrasear (2) e (3) como “começou a coordenação” e “começa a preparação”, respectivamente.

É possível argumentar que a diferença é sintática: em (1) o verbo é intransitivo e, em (2)-(3), possivelmente transitivo. A observação da frase (4), porém, desfaz a tese, pois trata-se de um “começar” transitivo com sentido pleno.

1. CF288-3 – A corrida sucessória **começa** esta semana com um quadro mais claro e definido do que o da semana passada.
2. CF28-1 – Pela segunda vez desde quando **começou** a coordenar as ações no Rio, há duas semanas, o Exército mudou o nome das operações.
3. CF118-2 – O escritório de Júlio Neves já **começa** a preparar novos estudos para o prolongamento deste corredor além do shopping Morumbi, em direção à ponte do Socorro.
4. CF39-2 – Diniz **começou** sua carreira automobilística em 1989, no Brasileiro de Fórmula Ford, campeonato em que obteve a sexta posição na classificação final.

Mesmo considerando que haveria um certo consenso em dizer que [começar a V_{infinitivo}] é uma locução verbal aspectual, Vilela e Koch não deixam claro se a noção de aspecto inclui as partículas “a”, “de” e “para” no escopo da locução verbal aspectual ou se elas não estão auxiliando na noção de aspecto junto ao verbo auxiliar. Em consequência, nós, na tarefa de anotação, precisamos decidir se tais partículas também podem ser consideradas de função auxiliar, e qual sua classe gramatical.

2.2. Sobre as preposições

Sem muito pensar no assunto, costumamos associar as formas das palavras “a”, “de” e “para” à de preposições, que, sendo (ilusoriamente) poucas, decoramos². Suponhamos que se possa dizer que as partículas “a”, “de” e “para”, que aparecem nas construções [comecei a V_{infinitivo}], [acabei de V_{infinitivo}] e [estar para V_{infinitivo}], são preposições. Em seu *Dicionário de Linguística e Gramática*, o autor [Câmara Júnior 1992] considera que preposições são:

“vocábulos que servem de morfema de relação para subordinar um substantivo como: adjunto a outro substantivo ou como complemento a um verbo. Esse processo de subordinação tem o nome de regência” ([Câmara Júnior 1992], p. 198).

²[Bagno 2012] nos alerta para a inadequação da tradição gramatical no tocante às preposições: decoramos, em média, 17, entre as quais poucas ainda são utilizadas atualmente, e deixamos de fora outras tantas que são mais usuais.

No entanto, ao postular que a partícula “a” em [começar a V_{infinitivo}] é preposição, fazemos diferente do que Mattoso definiu, pois no caso das locuções verbais não há um substantivo sendo relacionado, mas dois verbos; ou seja, o segundo verbo, quando muito, seria complemento do primeiro, e preposições não fazem relação entre verbos. Não é à toa que, no verbete “Aspecto”, Mattoso descreve as conjugações perifrásicas³ com “estar” sem nomear a partícula “a”: “o verbo auxiliar *estar*, conjugado com um gerúndio ou um infinitivo regido de *a*” ([Câmara Júnior 1992], p. 61).

Ao lidar com verbos modais, em sua *Gramática pedagógica do português brasileiro*, o autor [Bagno 2012] afirma que os verbos modais são auxiliares e os verbos que os seguem, seu complemento:

“a construção com os verbos modais se faz sempre com infinitivos na posição de verbo principal. Ao mesmo tempo, os verbos principais se constituem o complemento direto do verbo modalizador” ([Bagno 2012], p. 572).

Nesse ponto, a proposta vai ao encontro de Mattoso, pois, fazendo vista grossa e assumindo que preposições podem introduzir complementos que são verbos (orações subordinadas substantivas objetivas indiretas reduzidas de infinitivo), o segundo verbo, nessas construções, é complemento do primeiro.

Essa proposta, no entanto, traz algumas incongruências. Em primeiro lugar, do ponto de vista da anotação no ambiente UD, há uma contradição se levarmos ao cabo a observação de Bagno: não podemos considerar que o segundo verbo em [querer/poder/precisar V_{infinitivo}] é, ao mesmo tempo, complemento do primeiro verbo e verbo principal de uma locução verbal, pois, em UD, verbos auxiliares não podem ter complemento. Se encaramos que o segundo verbo é complemento do primeiro, ambos devem ser plenos.

Em segundo lugar, porque [Câmara Júnior 1992] argumenta, no verbete de conjugações perifrásicas, que:

“É má técnica de descrição gramatical considerar formas perifrásicas a combinação de dois verbos numa única oração em que ambos guardam a sua significação verbal e a significação total é uma das significações (**quero sair - vamos conversando** até a casa - já **tenho** uma carta **escrita**) e não houve a grammaticalização do primeiro verbo” ([Câmara Júnior 1992], p. 80).

Com a afirmação, além de partir do pressuposto de que seja fácil identificar quando um verbo guarda sua significação total – já vimos que não o é –, o autor utiliza o mesmo exemplo prototípico de “locuções verbais modais” – *quero sair* – para dizer que não concorda que se trate de uma locução verbal, na contramão tanto de [Vilela and Koch 2001] quanto de [Bagno 2012], que consideram tais construções como locuções verbais.

3. Desbravando o Bosque-UD

Para lidar com as partículas “a”, “de” e “para” no centro das locuções verbais, nenhuma consulta a gramática nos foi especialmente relevante. No entanto, uma exposição de

³“Conjugações perifrásicas”, como definidas em [Câmara Júnior 1992], têm definição idêntica à de locução verbal com que estamos lidando.

verbos na *Gramática Pedagógica do Português brasileiro* [Bagno 2012] lançou luz sobre um dado que nos parece esclarecedor (Tabela 1).

Tabela 1. Verbos auxiliares (6 primeiras entradas) em [Bagno 2012], p. 604

Verbo auxiliar	Exemplo
acabar	Ana acabou desistindo de viajar em julho.
acabar de	Ana acaba de desistir de viajar em julho.
acabar por	Ana acabou por viajar em julho.
andar	Ana anda pensando em viajar em julho.
cessar de	Ana ainda não cessou de sofrer com a separação.
começar	Ana começou falando dos pais.

A opção que Bagno faz por colocar as partículas “de” e “por” junto ao verbo na primeira coluna, mesmo que sem qualquer comentário sobre essa colocação, nos diz algo. A noção de auxiliaridade, de fato, comparece quando o verbo é acompanhado por tais partículas, evidenciando assim, por exemplo, que “acabar” seria diferente de “acabar de”; “vir” seria diferente de “vir a”, e, do mesmo modo, “começar a” seria diferente de “começar”.

A colocação de partículas (ou preposições) próximas ao verbo nos faz lembrar do conceito de *phrasal verbs* em inglês, quando uma preposição se junta a um verbo, criando uma entrada diferente tanto do verbo de origem quanto da preposição originária. Ainda que tenhamos dificuldade em chamar tais partículas de preposição, indicar que estamos diante de fenômenos semelhantes ao de *phrasal verbs* nos parece adequado. Para lidar com a anotação morfossintática das partículas “a”, “de”, “para” e “por”, portanto, assumiremos que elas são exigidas pelos verbos auxiliares, quando queremos torná-los auxiliares.

Continuando com a análise, no contexto UD, diríamos então que “começar a”, “vir a” e “acabar de” são expressões multi-palavras (MWEs) e, neste caso, a partícula associada se une ao verbo auxiliar pela relação de dependência *compound* (relação usada para os *phrasal verbs* do inglês). Em consequência, temos uma MWE do tipo “verbo auxiliar” e dependente do verbo principal, como anotado no formato UD nas figuras (3)-(7).



Figura 3. CP3-4 – «Estamos a dotar os computadores de um novo sentido» disse Steve d'Averio, director de marketing para a Europa da Logitech.



Figura 4. CF835-12 – Cursando economia na Faap, Kiko espera ansioso o seu telefone celular, que «está para sair».

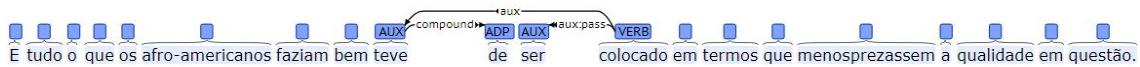


Figura 5. CF27-5 – E, assim, tudo o que os afro-americanos faziam bem teve de ser colocado em termos que menosprezassem a qualidade em questão.

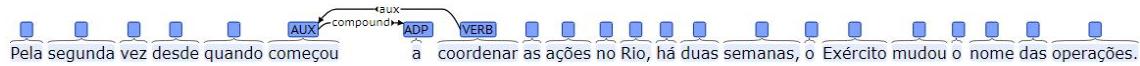


Figura 6. CF28-1 – Pela segunda vez desde quando começou a coordenar as ações no Rio, há duas semanas, o Exército mudou o nome das operações.



Figura 7. CF118-2 – O escritório de Júlio Neves já começa a preparar novos estudos para o prolongamento deste corredor além do shopping Morumbi, em direção à ponte do Socorro.

Como resultado e de forma a contribuir para a discussão, listamos as expressões multi-palavras que indicam aspecto⁴ no corpus Bosque-UD de duas maneiras: na tabela 2, por ordem alfabética, e na tabela 3, por frequência. Este novo tipo de *compound* verbal (o *compound* “phrasal verb”) aparece em 570 ocorrências no Bosque, distribuídas em 31 combinações distintas.

4. Considerações finais

Para nos auxiliar na tarefa de anotar construções como [estar a/para V_{infinitivo}] e [começar a V_{infinitivo}] analisamos duas gramáticas ([Vilela and Koch 2001] [Bagno 2012]) e um dicionário de linguística e gramática ([Câmara Júnior 1992]) da Língua Portuguesa. Elucidativos em alguns pontos, a análise das partículas “a”, “de” e “para” nas locuções verbais não foi explicitamente abordada por nenhuma das obras. Forçados a tomar uma decisão devido à tarefa de anotação, encontramos semelhanças com os chamados *phrasal verbs*, o que nos satisfaz do ponto de vista da análise e, simultaneamente, encontra alinhamento com a decisão do inglês (e talvez de outros treebanks de UD), critério relevante no contexto de um projeto de anotação multilíngue.

O próximo passo é implementar a decisão e investigar se ela traz impactos na consistência interna da anotação, o que pode ser verificado, de maneira indireta, por uma diminuição das confusões entre as relações de auxiliaridade (*aux*) e subordinação (*ccomp/xcomp*), por exemplo. Em caso positivo, teremos bons argumentos não apenas para incluir esta proposta de classificação na discussão sobre locuções verbais, mas também para defender a produtividade de uma descrição linguística que se articula com o PLN.

Agradecimentos

Elvis de Souza é bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) no projeto *Construção de datasets para o PLN de língua portuguesa*. Número do processo da bolsa: 128693/2019-3.

⁴Lembramos aqui que não interferimos no que foi considerado locução verbal, acatando a decisão original realizada pelo parser PALAVRAS [Bick 2000].

MWE	Frequência
acabar de	11
acabar por	30
andar a	3
chegar a	22
começar a	58
começar por	6
continuar a	57
continuar por	1
deixar de	30
dever a	1
estar a	124
estar para	1
estar por	1
ficar a	4
ficar de	1
haver a	1
haver de	2
haver que	1
ir a	3
ir de	1
parar de	3
passar a	43
poder a	3
ser de	3
tender a	1
ter a	9
ter de	62
ter que	3
tornar a	1
vir a	42
voltar a	42

Tabela 2. Lista das 31 expressões multi-palavras que indicam aspecto no Bosque-UD por ordem alfabética.

MWE	Frequência
estar a	124
ter de	62
começar a	58
continuar a	57
passar a	43
vir a	42
voltar a	42
acabar por	30
deixar de	30
chegar a	22
acabar de	11
ter a	9
começar por	6
ficar a	4
andar a	3
ir a	3
parar de	3
poder a	3
ser de	3
ter que	3
haver de	2
continuar por	1
dever a	1
estar para	1
estar por	1
ficar de	1
haver a	1
haver que	1
ir de	1
tender a	1
tornar a	1

Tabela 3. Lista das 31 expressões multi-palavras que indicam aspecto no Bosque-UD por ordem de frequência.

Referências

- Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá (c) tica: a treebank for portuguese. In *quot; In Manuel González Rodrigues; Carmen Paz Suarez Araujo (ed) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)(Las Palmas de Gran Canaria Espanha 29-31 de Maio de 2002) Paris: ELRA.* ELRA.
- Bagno, M. (2012). *Gramática pedagógica do português brasileiro.* Parábola Ed.
- Bick, E. (2000). The parsing system palavras. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.*
- Câmara Júnior, J. M. (1992). Dicionário de linguística e gramática: referente à língua portuguesa.
- Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.
- Vilela, M. and Koch, I. (2001). Gramática da língua portuguesa: Gramática da palavra, gramática da frase, gramática do texto/discurso. *Coimbra: Almedina.*

Quantificando (e qualificando) o sujeito oculto em português

Cláudia Freitas, Elvis de Souza, Luisa Rocha

Departamento de Letras
PUC-Rio – Brasil

claudiafreitas@puc-rio.br, elvis.desouza99@gmail.com, l.rocha7@globo.com

Abstract. *In information extraction we seek to find out who does what, when, where, and why. In Portuguese, it is possible to construct sentences with the omission of its subject. This omission – easy for humans but difficult for machines – causes information extraction to be compromised since we cannot fill the gap of who is responsible for the action. In this paper, we present a quantification of the hidden subject in Portuguese in corpora of different textual genres: journalistic, encyclopedic and literary. To do so, we make use of morphosyntactically annotated corpora and use a tool that allows us to interrogate corpora annotated with syntactic dependencies.*

Resumo. *Na extração de informações buscamos descobrir quem faz o quê, quando, onde, e por que. Diferentemente do inglês, na língua portuguesa é possível construir sentenças inteiras com a omissão do sujeito. Essa omissão – de fácil recuperação para humanos, mas difícil para máquinas – faz com que a extração de informações fique comprometida, uma vez que não temos como preencher a lacuna de quem é o responsável pela ação, já que boa parte da função sintática do sujeito corresponde à função semântica de agente. Neste trabalho, apresentamos uma quantificação do sujeito oculto em português em corpora de diferentes gêneros textuais: jornalístico, enciclopédico e literário. Para tanto, fazemos uso de corpora morfossintaticamente anotados e utilizamos uma ferramenta que nos permite interrogar corpora anotados com dependências sintáticas.*

1. Apresentação e motivação

Toda descrição é feita a partir de um ponto de vista. Ainda que, na descrição das línguas, a descrição seja frequentemente assumida como uma área/tarefa em si, lembramos que ela não é, e nem poderia ser, neutra ou desvinculada de uma motivação. Neste trabalho, assumimos uma descrição motivada por uma aplicação: a extração automática de informação em textos, uma das tarefas do PLN.

De maneira geral e muito simplificadamente, na extração de informações buscamos descobrir *quem faz o quê* (e *quando, onde, e por que*, dentre outros). Diferentemente do inglês (que concentra boa parte dos trabalhos de PLN), no entanto, a língua portuguesa permite a omissão do sujeito. Um dos tipos de omissão acontece em contextos em que o sujeito é facilmente recuperável – quer pelas pistas flexionais do verbo, quer pelo nosso conhecimento de mundo¹. Nesses casos, temos o chamado “sujeito oculto” segundo a tradição gramatical.

¹No entanto, como bem notou um dos revisores, embora a desinência verbal possa identificar a “pessoa”

A omissão do sujeito – facilmente recuperável pelos humanos, mas difícil para as máquinas – faz com que a extração de informações (e tarefas relacionadas, como a extração de citação (*Quotation Extraction*) e a anotação de papéis semânticos) fique comprometida, uma vez que não temos como preencher a lacuna de quem é o responsável pela ação, já que boa parte da função sintática do sujeito corresponde à função semântica de agente.

Mas qual o tamanho do problema para a língua portuguesa? E como quantificá-lo? Qual a correlação entre o sujeito oculto e gêneros textuais? Em trabalho anterior [Martins and Freitas 2019], realizamos de maneira semi-automática a quantificação de sujeitos ocultos em 6 verbetes do Dicionário Histórico Biográfico Brasileiro (DHBB), uma encyclopédia sobre a história política brasileira, publicada pelo CPDOC/FGV, que já vem sendo objeto de extração de informação [Higuchi et al. 2019]. Especificamente, foram escolhidos 6 verbetes biográficos que, juntos, totalizavam 25 mil palavras. Essa pequena amostra foi lida integralmente para a identificação dos casos de sujeito oculto. Em seguida, o texto foi processado automaticamente pela ferramenta UD-Pipe [Straka and Straková 2017] para a contabilização total dos verbos. Neste pequeno exercício, obtivemos uma distribuição desigual do sujeito oculto por verbete: o verbete com mais sujeitos ocultos apresentava o fenômeno em 45% das frases, e aquele com menos sujeitos ocultos, 18%. Considerando o tamanho do DHBB (mais de 8 milhões de palavras), e a relevância da explicitação do sujeitos para a extração de informação, decidimos ampliar o estudo, levando a cabo uma análise de todo o DHBB, e comparando os resultados com textos de outros gêneros textuais. Um desafio associado à tarefa é a existência de uma ferramenta que possibilite a contagem, já que no exercício de [Martins and Freitas 2019], todo o trabalho foi feito manualmente.

Neste trabalho, apresentamos uma quantificação do sujeito oculto em português em corpora de diferentes gêneros textuais: jornalístico, encyclopédico e literário. Para tanto, fazemos uso de corpora morfossintaticamente anotados e utilizamos uma ferramenta que nos permite interrogar corpora anotados com dependências sintáticas.

2. O que conta como sujeito oculto?

A língua portuguesa possui alguns fenômenos que envolvem a omissão de sujeito: o sujeito oculto propriamente, o chamado sujeito indeterminado e, ainda, as orações sem sujeito. Considerando nossa motivação principal, tarefas de PLN, não fizemos distinção, em nossa contagem, entre sujeito oculto (1) e sujeito indeterminado (2) - ou seja, nossa busca (e quantificação) pelas frases com “sujeito oculto” considera ambos os casos, já que, em ambos, não há sujeito explícito, ainda que exista um sujeito. A gramática [Cunha and Cintra 2008], por exemplo, considera o primeiro “sujeito oculto (determinado)” e o segundo “sujeito indeterminado”, sendo a diferença entre eles a possibilidade de determinação do sujeito pela desinência do verbo. Do mesmo modo, não levamos em conta as chamadas orações sem sujeito, como frases com verbos impersonais (3) e fenômenos da natureza.

a que corresponde o sujeito, nem sempre é possível recuperar o sujeito sem resolver a correferência do pronome pessoal. Isto porque as segundas pessoas são conjugadas como as terceiras pessoas (por exemplo, para preencher o sujeito da oração *Conseguiram um grande avanço* temos três possibilidades: *vocês*, *eles* ou *elas*.)

1. CP22-3: “Eu tentei, o senhor Vance tentou, se for respeitado, urrah!”, **comentou**.
2. CP31-3: Sempre que surge um problema, **chamam-na**
3. CP23-8: – **Há**, no ar, uma certa ideia de invasão.

Também não consideramos o sujeito oculto em orações subordinadas ou coordenadas à oração principal. Isto porque, nesses casos, o sujeito deve poder ser retomado no âmbito da frase. Ou seja, os sujeitos ocultos contabilizados foram apenas aqueles das orações principais.

Por fim, sabemos também que, em português, é possível que o *-se* exerça a função de um sujeito indeterminado. Neste caso, temos a seguinte situação: a presença formal de um elemento que conta como sujeito mas que, na prática, funciona como um índice de indeterminação. No entanto, e diferentemente dos demais casos de indeterminação, com o *-se* não é possível identificar quem é o sujeito; não é possível determiná-lo. Como nosso interesse está em apenas distinguir o sujeito oculto – porque são aqueles em que seria possível recuperar o sujeito – dos demais casos, não nos preocupamos em dar ao *-se* sujeito, neste momento, um tratamento especial, e também excluímos esses casos de nossa contagem.

Não somos os primeiros a se interessar pela a omissão do sujeito de um ponto de vista quantitativo. Em [Sardinha et al. 2014], por exemplo, a omissão do sujeito – codificada como *subjdrop* – é um dos critérios elencados para a caracterização de gêneros textuais em português. No entanto, não sabemos quais fenômenos são abarcados sob o referido rótulo. Os autores indicam que os critérios tiraram proveito da anotação feita pelo PALAVRAS [Bick 2000], e sabemos que o PALAVRAS reconhece (e distingue) verbos de orações sem sujeito explícito (sujeito oculto) e verbos de orações sem sujeito formal (oração sem sujeito), informação disponível desde 2008 nos corpora da Floresta Sintá(c)tica [Freitas et al. 2008]. Talvez, para a caracterização de gêneros textuais, seja irrelevante a diferença. No PLN, não é, dado que, nos sujeitos ocultos, recurso estilístico, podemos (e queremos) recuperar o sujeito sintático da oração.

3. Método e resultados

A pesquisa foi realizada nos corpora Bosque-UD (versão 2.4), com 9.366 frases; DHBB, com 323.301 frases; e em um subconjunto das obras de Machado de Assis (especificamente todos os contos, crônicas e romances), que totalizam 323.301 das frases do corpus OBras. Todo o material foi anotado pelo UDPipe [Straka and Straková 2017] e está no formato Universal Dependencies (UD) [Nivre et al. 2016]². Apenas o Bosque-UD teve sua anotação gramatical revista [Rademaker et al. 2017], por ser o material de treino do parser UDPipe.

Para realizar as pesquisas, utilizamos a ferramenta *Interrogatório*, um dos ambientes da ET, uma Estação de Trabalho para busca, revisão, edição e avaliação de corpora anotados [de Souza and Freitas 2019] – trata-se de uma ferramenta que surgiu motivada exatamente pela tarefa de contar sujeitos ocultos, já que não temos notícia de uma ferramenta de fácil acesso capaz de realizar contagens em material anotado com dependências sintáticas, formato subjacente à abordagem UD.

²<http://universaldependencies.org>

O principal desafio na contabilização dos sujeitos ocultos está no fato de que precisamos contar algo que não está presente. Na sintaxe de dependências, a primeira etapa foi encontrar todos os verbos de orações principais que não têm um sujeito que dele dependa. Tomando o Bosque-UD como exemplo, a pesquisa retornou 2774 frases. No entanto, essa expressão de busca retornou também outros tipos de frases, como construções com o verbo *haver* impessoal (exemplo 3). Criamos então, na ferramenta, um filtro para eliminar essas frases. Em seguida, fizemos mais um filtro, para eliminar as frases em que o verbo indicava um fenômeno da natureza.

Depois de aplicados os filtros, ficamos com 1480 sentenças, ou seja, cerca de 16% das frases do Bosque-UD. No DHBB, utilizando os mesmos critérios e expressões de busca, verificamos que 39.5% do corpus apresenta sujeitos ocultos. Por fim, no material literário (ou, ao menos, na escrita de Machado de Assis), 28.42% das frases continha sujeito oculto (tabela 1). Os resultados indicam que, independentemente do tipo de texto, os números são altos, justificando um tratamento especial para o fenômeno no PLN em português.

Tabela 1. Distribuição de sujeitos ocultos por corpus

Corpus	Frases com sujeito oculto
Bosque-UD (v.2.4)	15.8%
DHBB	39.5%
Machado de Assis	28.42%

Um dado interessante é a diferença na distribuição do fenômeno. O DHBB é, de longe, o corpus com mais sujeitos ocultos. A constatação é facilmente explicada quando sabemos que o material possui dois tipos de verbetes: biográficos e temáticos, sendo os verbetes do primeiro tipo a maioria. E, justamente por se tratar de um artigo biográfico, a omissão do sujeito – na imensa maioria das vezes, correspondente à pessoa verbetada – funciona como um recurso estilístico capaz de trazer fluidez ao texto, evitando repetições desnecessárias. Em seguida, temos o texto literário e o Bosque, composto por textos jornalísticos. O material de Machado de Assis possui o dobro de sujeitos ocultos do Bosque. Por outro lado, o Bosque-UD teve sua anotação revista, e o corpus com as obras de Machado, não. Além disso, vale lembrar que tanto este corpus, como o DHBB, foram anotados por um modelo que foi treinado no Bosque-UD. Deste modo, é possível que a contagem sofra efeitos de uma anotação sintática malfeita.

A fim de verificar o quanto os resultados da tabela 1 são confiáveis, analisamos manualmente uma amostra de 150 frases identificadas como sujeito oculto, 50 de cada corpus/gênero. Destas, 134 (90%) foram avaliadas como corretas, o que nos dá confiança quanto aos números apresentados, sobretudo no que se refere ao DHBB, onde todas as frases identificadas estavam corretas. A tabela 2 apresenta a distribuição dos resultados.

O corpus com textos literários (e diacrônico) é o que mais apresenta erros. A análise dos casos errados (apenas 16 erros) indicou que o principal motivo do erro (9 casos) decorre do processamento automático, e acontece quando temos um sujeito posposto ao verbo (1), sobretudo se estamos diante de um sujeito oracional (2). Em seguida, temos 3 erros que decorrem da nossa forma de pesquisa: buscamos pela ausência de sujeito na oração principal, mas não é raro que orações adverbiais antecedam a oração principal,

Tabela 2. Resultados da análise manual de 150 frases, por

Corpus	Acertos
Bosque-UD (v.2.4)	44 (88%)
DHBB	50 (100%)
Machado de Assis	40 (80%)
Total de acertos	134 (90%)

trazendo com elas o sujeito (3). Isto é algo que precisamos melhorar para números mais precisos, mas não acreditamos que a frequência dessas construções seja capaz de interferir de maneira significativa na contagem final – foram apenas 3 casos desses em 150 frases. Por fim, tivemos também 4 ocorrências, no Bosque-UD, da construção com sujeito indeterminado (construções com “tratar-se de”), e também precisaremos eliminar esses casos da busca.

1. Mas por outro lado, sem a apresentação de Miss Dollar, *seria o autor* obrigado a longas digressões, que encheriam o papel sem adiantar a ação
2. Era conveniente ao romance *que o leitor ficasse muito tempo sem saber quem era Miss Dollar*.
3. Se eu dirigisse uma federação, *apresentaria* balanços mensais e liberaria minhas contas bancárias

4. Considerações finais

Apresentamos aqui um primeiro estudo sobre a quantificação do sujeito oculto em português, levando em conta também o gênero textual. Os resultados indicam que o gênero é um aspecto relevante no que se refere à frequência do sujeito oculto. No caso de uma enciclopédia biográfica, por exemplo, e quando pensamos em cadeias de relações como *quem faz o quê* (e *quando* e *onde*), em cerca de 40% dos casos não temos como resolver a dimensão *quem*. Por outro lado, em textos jornalísticos, a quantidade de frases sem sujeito cai para cerca de 15%, o que, se não é muito, também não é insignificante.

Outro ponto a ser destacado é a necessidade de ferramentas capazes de auxiliar linguistas na sua tarefa de manipulação de grandes corpora anotados. Em nosso caso, a necessidade de contar de forma simples o sujeito oculto acabou levando ao desenvolvimento de um ambiente complexo para trabalhar com corpus, de uma maneira linguisticamente motivada.

Por fim, reforçamos os ganhos do lado linguístico e do lado computacional ao se pensar uma descrição do português motivadas pelos desafios empíricos do PLN.

Referências

- Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis.
- Cunha, C. and Cintra, L. (2008). *Nova gramática do português contemporâneo*. Lexikon.
- de Souza, E. and Freitas, C. (2019). Et: uma estação de trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente. In *VI Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic 2019)*.

- Freitas, C., Rocha, P., and Bick, E. (2008). Um mundo novo na floresta sintá(c)tica – o treebank do português. *Calidoscópio*, 6(3):142–148.
- Higuchi, S., Santos, D., Freitas, C., and Rademaker, A. (2019). Distant reading brazilian politics. In Navarretta, C., Agirrezabal, M., and Maegaard, B., editors, *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, volume 2364, Copenhagen, Denmark. <http://ceur-ws.org/Vol-2364/>.
- Martins, F. and Freitas, C. (2019). Sujeitos ocultos em verbetes biográficos: Contornando dificuldades da extração automática de informações. In *XI Congresso Internacional da ABRALIN*.
- Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.
- Sardinha, T. B., Kauffmann, C., and Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in brazilian portuguese. *Corpora*, 9(2):239–271.
- Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Avaliação do uso da Diversidade Contextual e da Frequênci para a Tarefa de Identificação de Palavras Complexas em Simplificação Lexical

Nathan Siegle Hartmann¹, Sandra Maria Aluísio¹

¹Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematics and Computer Sciences
University of São Paulo

{nathansh, sandra}@icmc.usp.br

Abstract. *Lexical Simplification (LS) has the function of exchanging words or expressions by synonyms that can be understood by a greater number of people. The focus of this article is on the first stage of LS systems, called Complex Word Identification (CWI), that is, identifying which words in a sentence are considered complex by a particular target audience. The objective of this article was to evaluate two CWI methods of the threshold-based approach (word frequency and contextual diversity) in the Leg2Kids subtitle corpus, contrasting this approach with the one assisted by machine learning, using the same frequency and contextual diversity resources in Leg2Kids.*

Resumo. *A Simplificação Lexical (SL) tem a função de trocar palavras ou expressões por sinônimos que podem ser entendidos por um maior número de pessoas. O foco deste artigo é na primeira etapa dos sistemas de SL, chamada de Identificação de Palavras complexas (IPC), isto é, identificar quais palavras de uma sentença são consideradas complexas por um determinado público-alvo. O objetivo deste artigo foi avaliar dois métodos da tarefa de IPC da abordagem baseada em threshold (frequência de palavras e diversidade contextual) no корпус de legendas Leg2Kids, contrastando esta abordagem com a assistida por aprendizado de máquina, usando os mesmos recursos de frequência e diversidade contextual no Leg2Kids.*

1. Introdução

A Simplificação Lexical (SL) tem a função de trocar palavras ou expressões por sinônimos que podem ser entendidos por um maior número de pessoas. Um sistema automático para SL realiza os seguintes passos em *pipeline*: (i) análise da complexidade lexical, que seleciona as palavras ou expressões que são consideradas complexas para um leitor e/ou tarefa; (ii) busca por substitutos, em geral, sinônimos com o mesmo sentido usado no contexto; e (iii) ranqueamento dos sinônimos do passo (ii) de acordo com o quanto simples eles são para o leitor e/ou tarefa [Specia et al. 2012]. Após a escolha do sinônimo adequado, há a troca da palavra em foco pelo sinônimo, que pode pedir ajustes na escrita das palavras da oração, como a adequação de gênero e/ou número. Na Figura 1, exemplificamos o processo de SL para uma sentença em português. Primeiramente, identifica-se a palavra complexa que deve ser simplificada, no caso, *precauções*. O segundo passo é listar os possíveis candidatos à simplificação, de forma a não comprometer o significado

da sentença. A próxima etapa é ordenar os candidatos pela simplicidade e adequação ao contexto. Finalmente, escolhemos o substituto mais adequado, que leva em consideração o público alvo (muitas vezes o substituto não é simples o bastante). O último passo é realizar a substituição da palavra e adequá-la à sentença, verificando o gênero, número e grau das palavras.

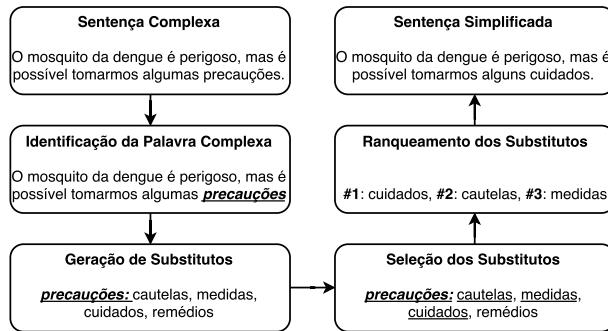


Figura 1. Exemplo de Simplificação Lexical.

O foco deste artigo é na primeira etapa dos sistemas de SL – identificar quais palavras de uma sentença são consideradas complexas por um determinado público-alvo [Shardlow 2013]. Essa tarefa é conhecida como Identificação de Palavras Complexas (IPC), do inglês *Complex Word Identification (CWI)*, e tem atraído a atenção da comunidade de pesquisa nos últimos anos [Paetzold and Specia 2016], [Zampieri et al. 2017], [Yimam et al. 2018]. IPC é um passo importante no *pipeline* de um sistema de SL. Por um lado, um modelo conservativo de IPC pode falhar na detecção de muitas palavras complexas, deixando-as sem adaptação e limitando, assim, a utilidade do sistema de SL. Por outro lado, um modelo de IPC agressivo pode identificar palavras simples como complexas, fazendo simplificações desnecessárias, aumentando assim o risco de incorrer em erros na etapa de substituição [Lee and Yeung 2018].

Segundo [Paetzold 2016], há cinco categorias de estratégias para IPC: (1) Simplificação de todas as palavras (criando um modelo agressivo de IPC); (2) Baseada em *threshold* (delimita um limiar rígido e simples para a classificação considerando, por exemplo, o tamanho das palavras); (3) Baseada em léxico; (4) IPC implícita (considera que todas as palavras de uma sentença são alvo de simplificação); e (5) IPC assistida por Aprendizagem de Máquina (AM). A categoria baseada em AM é a mais atual e com melhores chances de sucesso, dado um *dataset* para treinamento da tarefa.

Para o português, no melhor do nosso conhecimento, o único trabalho que avaliou a tarefa de IPC foi [Watanabe et al. 2009], seguindo a abordagem baseada em léxico no projeto PorSimples, cujo público alvo eram adultos com baixa escolaridade. Além disso, apenas recentemente [Hartmann et al. 2018] apresentou o SIMPLEX-PB¹, o primeiro córpus publicamente disponível de SL para o Português Brasileiro (PB), voltado para crianças do 3º ao 9º anos do Ensino Fundamental (descrito na Seção 4.2), o que nos dá a oportunidade de avaliar métodos de todas as abordagens para o público alvo infantil.

O objetivo deste artigo é avaliar dois métodos da abordagem baseada em *threshold* (frequência de palavras e diversidade contextual (DC)) (descritos na Seção 4.3) em um

¹github.com/nathanshartmann/simplex

grande córpus de legendas de filmes e séries para crianças compilado neste trabalho, o Leg2Kids² (descrito na Seção 3). O primeiro método porque frequência é a escolha mais popular desta abordagem [Paetzold 2016] e o segundo método, que é definido como o número de documentos diferentes em que uma palavra aparece dentro de um córpus³, porque tem se mostrado como um melhor substituto para a frequência em várias tarefas relacionadas ao reconhecimento e aprendizado de novas palavras em experimentos de leitura de sentenças para crianças [Rosa et al. 2017]. Além disso, contrastamos os dois métodos acima com a abordagem assistida por Aprendizagem de Máquina, que usa também os mesmos recursos de frequência e a DC do Leg2Kids.

2. Trabalhos Relacionados em Identificação de Palavras Complexas

Em 2016, a avaliação conjunta sobre CWI na SemEval [Paetzold and Specia 2016] tratou da IPC para a língua inglesa, sendo o melhor resultado o do time que combinou vários subsistemas das abordagens baseada em *threshold*, léxico e AM com métodos baseados em votação, atingindo uma precisão de 0,14 e cobertura de 0,77. Naquele trabalho, a frequência das palavras foi a *feature* que mais agregou na predição da tarefa. Os subsistemas que usaram AM foram treinados com *features* morfológicas, léxicas e semânticas.

Em 2017, [Zampieri et al. 2017] revisitaram os resultados do SemEval CWI de 2016, tentando entender a razão da baixa performance de muitos sistemas no *dataset* da avaliação conjunta, além de determinar o teto superior da tarefa de IPC no *dataset* da competição (0,6 F1), usando a saída dos sistemas participantes. A avaliação sobre o *dataset* revelou 3 fatos, confirmando a relação da anotação de não-nativos com o desempenho baixo dos sistemas: (i) 50% das palavras mais complexas (anotadas por mais anotadores) não receberam o mesmo rótulo no conjunto de treinamento e teste; (ii) palavras que foram anotadas mais frequentemente como complexas no conjunto de treinamento tenderam a ser mais fáceis para os sistemas identificarem; e (iii) palavras que foram atribuídas como complexas são, em média, mais longas do que as não complexas.

Já a SemEval de 2018 trouxe um *dataset* multilíngue para a tarefa de IPC. No geral, as abordagens baseadas em engenharia de *features* (baseadas geralmente em tamanho da palavra e frequência) desempenharam melhor do que abordagens neurais baseadas em *word embeddings*, mas nesta edição mais sistemas empregaram *deep learning* e esses modelos estão cada vez melhores, indicando uma tendência na resolução da tarefa.

[Lee and Yeung 2018] trouxeram uma visão de como a tarefa de IPC pode ajudar a personalizar sistemas de SL, dado que a maioria deles é treinada para encontrar uma substituição ótima para uma dada palavra complexa ou mesmo uma lista de substituições para todos os usuários, não considerando as diferenças em níveis de proficiência dos usuários alvos. Os autores concluem que, mesmo um modelo simples de IPC, baseado em listas de vocabulário para séries escolares, pode ajudar a reduzir o número de simplificações desnecessárias.

²<http://nilc.icmc.usp.br/leg2kids>

³No caso de córpus de legendas, DC é definida como o número de filmes/séries em que uma palavra aparece [Adelman et al. 2006].

3. O córpus Leg2Kids

Para compilar um córpus que represente o léxico em português mais ouvido por crianças, entramos em contato com a equipe do Open Subtitles⁴, o maior repositório de legendas com um acervo de aproximadamente 5 milhões de legendas para o Português brasileiro. A Open Subtitles nos disponibilizou um córpus de 36.413 legendas de filmes e séries dos gêneros Família e Animação, pois acreditam que esses melhor descrevem o material que as crianças têm acesso, dado que não há um metadado específico de conteúdo e seu público-alvo.

Realizamos um pré-processamento nas legendas do córpus, removendo as marcações de tempo existentes em cada trecho da legenda (essas marcações definem o intervalo de tempo em que um trecho será exibido na tela). Removemos, também, marcações dos editores da legenda, como endereços de páginas web, agradecimentos, patrocínio, entre outros. O córpus foi, então, sentenciado e tokenizado pela ferramenta NLTK⁵.

Leg2Kids contém um total de 153.791.083 *tokens* e 452.312 *types*, atingindo um *type-token ratio* (TTR) de 0,294%. No comparativo com outros córpuses do mesmo gênero, esse valor de TTR implica em uma maior riqueza lexical do que o SUBTLEX-PT-BR [Tang 2012] (0,22% TTR) mas menor riqueza que o Escolex [Soares et al. 2014] (1,5% TTR). Concluimos que o TTR do Leg2Kids é baixo, dado que 30% das palavras do córpus ocorrem uma única vez e, ao analisarmos as 90% palavras mais frequentes, elas não ocorrem mais que 58 vezes - um contraste ao compararmos com a palavra de conteúdo mais frequente do córpus (*estar*), que ocorre pouco mais de 1 milhão de vezes.

4. Avaliação da Diversidade Contextual e da Frequência para IPC

Nesta seção, apresentamos os recursos de dicionários do MEC na Seção 4.1, pois foram usados no dataset SIMPLEX-PB, descrito na Seção 4.2. Na seção 4.3 apresentamos as configurações do experimento realizado a na seção 4.4 apresentamos os resultados obtidos.

4.1. Dicionários do Programa Nacional do Livro Didático (PNLD) do MEC

Para desenvolver um modelo que mensure complexidade lexical, fizemos uso dos dicionários selecionados pelo PNLD⁶, programa do MEC, categorizados pelo nível de complexidade lexical esperado em cada etapa escolar [Hartmann et al. 2018]. Esses dicionários foram utilizados na identificação de palavras difíceis para crianças do Ensino Fundamental, na compilação do córpus SIMPLEX-PB [Hartmann et al. 2018] de SL, descrito na Seção 4.2: O dicionário de nível 1, formado pelo dicionário Caldas Aulete com a Turma do Cocoricó, contempla o 1º ciclo do Ensino Fundamental 1 (1º ao 3º anos) e possui 1.371 entradas; o dicionário de nível 2 é composto pelo Dicionário Escolar da Língua Portuguesa, Dicionário Ilustrado de Português, e Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Pica-Pau Amarelo, e contempla o 2º ciclo do Ensino Fundamental 1 (4º e 5º anos), possuindo 8.171 entradas; o dicionário de

⁴<https://www.opensubtitles.org>

⁵<https://www.nltk.org/>

⁶<http://portal.mec.gov.br/pnld>

nível 3, composto aqui pelo Minidicionário Contemporâneo da Língua Portuguesa, contempla o Ensino Fundamental 2 (6º ao 9º anos), e possui 29.970 entradas. Podemos dizer que o *dataset* SIMPLEX-PB já apresenta um viés de personalização pleiteado por [Lee and Yeung 2018], como apresentado na Seção 2.

4.2. SIMPLEX-PB

O SIMPLEX-PB é composto por 1.582 sentenças com palavras contidas no dicionário nível 3 (cf. Seção 4.1), que contempla o léxico a ser aprendido por crianças do 6º ao 9º anos, não contidas nos dicionários anteriores. Com relação a crianças do 1º ao 5º anos do Ensino Fundamental, todas as palavras identificadas pelo córpus SIMPLEX-PB são complexas e alvos de simplificação. Na Figura 2, podemos ver um trecho do SIMPLEX-PB, que é estruturado de forma tabular e conta com sentenças (coluna *sentence*), palavras complexas (coluna *target_word*) e sinônimos (coluna *synonyms*). Para o presente trabalho, nosso interesse é exclusivamente as palavras complexadas.

instance	sentence	synonyms	target_word
INSTÂNCIA 0	Por enquanto, o Circo da Física se apresenta a...	[estender, alargar, amplificar, aumentar, ampl...]	expandir
INSTÂNCIA 1	O pesquisador não vira os exemplares dessa esp...	[pegar, colher, pegaram, recolher, capturaram,...]	colearam
INSTÂNCIA 528	(Universidade de São Paulo), registrar o qu...	[demonstração, preleção, lecionação, ensino, i...]	palestras
INSTÂNCIA 647	O tempo de *gestação* do peixe-boi é de um ano...	[prenhez, gravidez]	gestação
INSTÂNCIA 61	Aliado à observação de organismos vivos , isso...	[desenvolver, elaborar, criar, concluir, escre...]	formular
INSTÂNCIA 62	Há muito tempo , em 1801 , o francês François...	[intitular, denominar, nomear]	batizou
INSTÂNCIA 63	Para minha felicidade , os dois toparam *compa...	[partilhar, divulgar, dividir, repartir, parti...]	compartilhar
INSTÂNCIA 64	Havia também um desenho animado que ajudou a *	[divulgar, difundir]	popularizar

Figura 2. Trecho retirado do SIMPLEX-PB.

4.3. Configurações do experimento

Avaliamos aqui duas das abordagens descritas em [Paetzold 2016]: (i) baseada em *threshold* e (ii) baseada em AM, com base no córpus de legendas de filmes e séries infantis Leg2Kids. Desenvolvemos também dois modelos *baselines*: (i) classifica todas as palavras como simples; e (ii) classifica todas as palavras como complexas.

A abordagem baseada em *threshold* considerou a média da frequência e a média da diversidade lexical no Leg2Kids como limiares na classificação da complexidade lexical. Por exemplo, se uma dada palavra possui frequência maior que a média das frequências no Leg2Kids, essa palavra é considerada complexa, caso contrário ela é considerada simples.

A abordagem baseada em AM fez uso da Regressão Logística como método de classificação, por ser um método clássico e recomendado quando a variável dependente é de natureza binária. Para IPC, o uso de AM é mais interessante do que a abordagem simples que se baseia em *threshold* porque podemos usar o SIMPLEX-PB como córpus de treinamento, fazendo uso das anotações de IPC, e também extraír *features* do Leg2Kids para conseguir uma melhor generalização nos resultados. Essa generalização permite, idealmente, uma maior cobertura no universo das palavras que podemos identificar como complexas e, consequentemente, simplificar, não se restringindo somente àquelas listadas nos dicionários do MEC.

Consideramos como complexas, todas as palavras marcadas como complexas no córpus SIMPLEX-PB. Utilizamos duas estratégias na demarcação de palavras simples: (i) selecionamos, aleatoriamente, palavras contidas nos textos do SIMPLEX-PB diferentes daquelas marcadas como complexas; (ii) selecionamos palavras dos dicionários de nível 1 e 2, que contêm vocabulário considerado difícil para crianças entre o 1º e 5º anos do Ensino Fundamental. O uso dessas duas abordagens na seleção de palavras simples nos possibilita a avaliação da abordagem baseada em léxico e também avaliar a adequação em relação ao propósito ao qual os dicionários foram selecionados pelo MEC já que, quanto maior o nível do dicionário, mais complexas são as palavras contidas nele.

Nossa abordagem de AM foi calibrada com: (i) somente uso da frequência das palavras; (ii) somente uso da diversidade contextual das palavras; e (iii) ambas. Uma avaliação nesse cenário nos permite verificar qual das duas *features* tem melhor resultado na tarefa de IPC e, também, se o uso combinado delas é mais eficaz do que o uso separado.

Utilizamos *10-fold cross-validation* na separação dos dados entre treinamento e teste, estratégia que dá bons subsídios na garantia da generalização do modelo, que não só é interessante mas também necessária para que a solução possa ser amplamente utilizada. A métrica de avaliação foi a F1, que consiste na média harmônica entre a precisão e a cobertura, avaliando tanto a acertividade quanto a amplitude de atuação do modelo.

4.4. Resultados do experimento

Na Tabela 1, reportamos os resultados do experimento de comparação das abordagens de *threshold* e AM em IPC.

Os métodos baseados em *threshold* obtiveram desempenho inferior aos *baselines* e não mostraram ser uma boa escolha para a tarefa. Esse resultado está alinhado com o atestado para o inglês [Paetzold 2016].

Os métodos baseados em AM, por sua vez, bateram com ampla margem os *baselines*. Ambas as *features* frequência e diversidade contextual mostraram ser bons *proxies* na classificação da complexidade lexical, mas a primeira se mostrou mais efetiva - resultado alinhado com os encontrados por [Paetzold and Specia 2016], que diz que a forma mais efetiva de determinar a complexidade de uma palavra é consultando a sua frequência em um córpus de qualidade. Interessante observar que o uso combinado das duas *features* não agregou performance ao modelo. Podemos entender que ambas as informações modelam a ocorrência de palavras em córpus: enquanto a frequência mensura, de forma bruta, o volume de ocorrências de uma palavra em córpus, a diversidade contextual calcula a frequência com que uma palavra ocorre em documentos distintos. Podemos inclusive fazer um paralelo com a *feature* clássica de AM, o TF-IDF, em que o TF (*term frequency*) é análogo à frequência das palavras e o IDF (*inverse document frequency*) é análogo à diversidade contextual.

Por fim, comparamos o desempenho na IPC entre a abordagem que seleciona palavras simples por meio de dicionários (DIC) e a abordagem que seleciona palavras aleatórias da língua para serem palavras simples (ALE). O desempenho da abordagem ALE superou com boa margem a DIC (ver Tabela 1). Um estudo para entendimento do ocorrido se faz necessário, já que a abordagem DIC usa um léxico com graduação de complexidade de palavras e a abordagem ALE usa palavras genéricas da língua.

Abordagem	Features	Método	F1
Baseline 1	Todas as palavras simples		0,33
Baseline 2	Todas as palavras complexas		0,33
Palavras da língua (aleatório) (ALE)	Frequência	Régressão Logística	0,88
	Diversidade Contextual	Régressão Logística	0,81
	Frequência e Diversidade Contextual	Régressão Logística	0,88
	Frequência maior que média no Leg2Kids	Threshold	0,10
	DC maior que média no Leg2Kids	Threshold	0,12
Dicionários nível 1 e 2 (DIC)	Frequência	Régressão Logística	0,79
	Diversidade Contextual	Régressão Logística	0,72
	Frequência e Diversidade Contextual	Régressão Logística	0,79
	Frequência maior que média no Leg2Kids	Threshold	0,23
	DC maior que média no Leg2Kids	Threshold	0,19

Tabela 1. Resultados na avaliação de IPC para o PB.

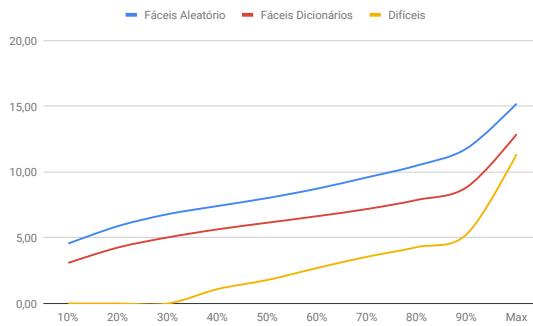


Figura 3. Distribuição da feature Frequência nos experimentos realizados.

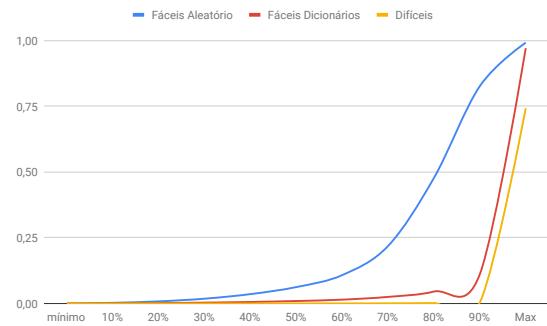


Figura 4. Distribuição da feature Diversidade Contextual nos experimentos realizados.

Como apresentado na Figura 3 e Figura 4, respectivamente, palavras genéricas da língua ocorrem muito mais frequentemente no córpus de legendas e com diversidade contextual muito maior entre as legendas do córpus do que as palavras de dicionários infantis do MEC (tanto as dos dicionários de nível 1 e 2, fáceis, quanto as do dicionário de nível 3, difíceis). Métodos de AM usam essa diferença na distribuição das *features* para traçar o limiar de classificação da tarefa que, no nosso trabalho, é distinguir dois grupos lexicais. Assim, o *gap* de desempenho é justificável porque as diferenças entre as distribuições das *features* frequência e diversidade contextual são substancialmente mais distintivas entre as palavras genéricas da língua e as palavras, aqui, definidas como complexas, do que as palavras simples vindas de dicionários e as palavras complexas.

Assim, o desempenho da abordagem ALE, na realidade, nos mostra que foi possível diferenciar palavras referente a crianças do 6º ao 9º anos do Ensino Fundamental de palavras gerais da língua, o que é interessante, pois mostra que conseguimos identificar o nicho de palavras infantis entre o universo de palavras da Língua Portuguesa e esse é, inclusive, o objetivo deste trabalho. Em relação a comparação dos modelos ALE e DIC, não podemos inferir nada já que o léxico do modelo ALE privilegia o resultado apresentado no experimento e mascara a real boa performance do modelo DIC.

5. Conclusões e Trabalhos Futuros

O objetivo deste artigo foi avaliar dois métodos da abordagem baseada em *threshold* (frequência de palavras e diversidade contextual) no córpus de legendas Leg2Kids, contrastando esta abordagem com a assistida por Aprendizagem de Máquina, usando os mesmos recursos de frequência e diversidade contextual no Leg2Kids.

As contribuições desse trabalho são: (1) a disponibilização pública do Leg2Kids; (2) a verificação da frequência como *feature* mais distintiva para a tarefa binária de IPC, via abordagem de AM; (3) a distinção das palavras dos dicionários infantis e, consequentemente, ranquear palavras pela sua complexidade - atendendo à tarefa de IPC; e (4) a divulgação desta tarefa para o português, alinhando-a com os avanços para o inglês, por exemplo.

São vários os trabalhos futuros que antevemos para essa pesquisa: avaliar outros métodos de AM para a tarefa além da Regressão Logística, por exemplo, Árvores de decisão, Random Forest, Bagging, Boosting, SVM; incrementar o número de *features* da abordagem baseada em AM, usando tamanho das palavras, número de sílabas, número de sentidos e sinônimos em tesouros; além de avaliar modelos avançados de *deep learning*, que são uma tendência da área para a tarefa.

Referências

- Adelman, J. S., Brown, G. D., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.
- Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. (2018). Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.
- Lee, J. and Yeung, C. Y. (2018). Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Paetzold, G. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.
- Paetzold, G. H. (2016). *Lexical Simplification for Non-Native English Speakers*. PhD thesis, University of Sheffield.
- Rosa, E., Tapia, J. L., and Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS One*, 12(6).
- Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.
- Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, A., de Almeida, J. J., Pinheiro, A. P., and Comesana, M. (2014). Escolex: A grade-level lexical database from european portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1):240–253.

- Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM-2012)*, pages 347–355. Association for Computational Linguistics.
- Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, 24:208–214.
- Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Aluísio, S. M. (2009). Facilita: Reading assistance for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication*, pages 29–36. ACM.
- Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.
- Zampieri, M., Malmasi, S., Paetzold, G., and Specia, L. (2017). Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Teste de Memória de Trabalho de Leitura: Versão Computadorizada Padronizada do *Reading Span Test* para o Português Brasileiro

Laiane F. N. Vasconcelos¹, Priscilla de A. Almeida¹, Gustavo L. Estivalet¹, José Ferrari-Neto¹

¹Universidade Federal da Paraíba (UFPB), Programa de Pós-Graduação em Linguística (PROLING), João Pessoa - PB, Brasil

*laianefigueiredo@gmail.com, prisca.albuquerque@gmail.com,
gustavoestivalet@hotmail.com, joseferrarin@ibest.com.br*

Abstract. *The working memory is related to the study of language, and its span capacity plays a crucial role in reading comprehension. Thus, the creation of a computerized and standardized version of the Reading Span Test (RST) in Brazilian Portuguese (BP) contributes for the investigation of the functioning of the central executive control during reading in speakers of this language. This article presents the methodology used for the development of this new version of the RST. We remark that its standardized methodology allows the comparison to other international studies and the computation of different scores of working memory.*

Resumo. *A memória de trabalho apresenta relação com o estudo da linguagem, e a capacidade de memória de trabalho desempenha um papel crucial na compreensão leitora. Assim, a criação de uma versão computadorizada e padronizada do Reading Span Test (RST) para o português brasileiro (PB) contribui para investigação do funcionamento do controle executivo central durante a leitura em indivíduos falantes desta língua. Este artigo apresenta a metodologia utilizada para o desenvolvimento desta nova versão do RST. Destaca-se que sua padronização metodológica permite a comparação com outros estudos internacionais e a computação de diferentes pontuações de capacidade de memória de trabalho.*

1. Introdução

A memória de trabalho (MT) desperta interesse em pesquisas relacionadas à Psicolinguística Experimental, pois, em estudos que investigam sua relação com o processamento da linguagem, ela exerce um papel decisivo, desde a aprendizagem de novas palavras até a produção e a compreensão da linguagem. Conforme Just e Carpenter (1978), a capacidade de MT desempenha um papel crucial na compreensão leitora. Contudo, ainda existem muitas discussões acerca do funcionamento dessa capacidade e de como ela influencia os processos de leitura.

Daneman e Carpenter (1980) desenvolveram o teste de capacidade de MT de leitura (*Reading Span Test* - RST) para medir a capacidade de MT verbal, utilizado até

os dias atuais. Outras versões e adaptações desse teste foram criadas e passaram também a ser usadas como instrumentos de investigação das funções executivas centrais. Segundo van den Noort *et al.* (2008), uma das vantagens da criação de versões padronizadas é possibilitar a comparação direta entre os diferentes grupos de pesquisa e entre diferentes idiomas.

Sendo assim, a partir da versão original e das adaptações subsequentes do RST, o presente trabalho tem como objetivo apresentar e descrever a metodologia utilizada na criação da Versão Computadorizada e Padronizada do RST para o português brasileiro (PB) - VCP-RST-PB). Os estudos apontam que a MT é um sistema responsável por manter temporariamente ativadas informações importantes para execução de tarefas complexas (DANEMAN; HANNON, 2001), portanto, ela se torna relevante em estudos de Psicolinguística, Psicopedagogia e Educação no PB, que visam verificar sua influência durante o desempenho da leitura.

2. Memória de Trabalho

Antes de se iniciar a apresentação do experimento desenvolvido, será realizada uma breve explanação sobre a MT e sua pertinência em testes psicolinguísticos. Conforme Uehara e Landeira-Fernandez (2010), ela é essencial para o desempenho de atividades cognitivas do dia-a-dia, além de ser importante para o desempenho em atividades linguísticas, como o diálogo e a leitura.

Conforme Baddeley (*apud* Mota, 2015), a MT pode ser definida como um sistema que tem como função o armazenamento e o processamento temporário de informação para realização de atividades cognitivas complexas. Observa-se que esse sistema apresenta capacidade limitada e funciona como uma espécie de interface entre o processamento e o armazenamento temporário de informações, associando-se à mecanismos de atenção e de memória de longo prazo (UEHARA; LANDEIRA-FERNANDEZ, 2010). O modelo de MT proposto Baddeley e Hitch (1974) é constituído de subsistemas que atuam de forma interativa, sendo eles a alça fonológica, a alça visuo-espacial e executivo central. O executivo central controla os demais subsistemas através da manipulação e da conservação de informações.

Durante a leitura, as informações do texto devem estar na MT para serem ativadas quando necessário. Porém, essas informações podem ser perdidas, uma vez que a capacidade da MT é limitada (MILLER, 1956). Logo, testes como o RST buscam verificar o limite da capacidade individual dos sujeitos durante a leitura de frases. Observa-se que a capacidade de MT fornece informações pertinentes em relação a retenção e falhas de memória no processamento da informação linguística (RODRIGUES, 2001). Para Daneman e Carpenter (1980), limitações na MT podem se relacionar diretamente a prejuízos no processamento da linguagem. Por exemplo, se o sujeito executa de maneira ineficiente processos específicos de compreensão em leitura, ele passa a ter menos recursos disponíveis para armazenar informações já processadas.

3. Testes de Capacidade de Memória de Trabalho

O RST (DANEMAN; CARPENTER, 1980) é um teste composto por cartões organizados em três conjuntos, cada um com duas, três, quatro, cinco e seis frases. Estas são apresentadas progressivamente nos conjuntos (de 2 a 6 frases). As frases possuem

entre 13 e 16 palavras, são independentes umas das outras e devem ser lidas em voz alta pelo participante. A palavra final (palavra-alvo) de cada frase deve ser memorizada e relembrada quando solicitado. A tarefa segue até o nível de retenção que o sujeito consegue memorizar ou até o final dos 3 conjuntos (caso tenha memorizado todas as palavras), determinando a capacidade de MT. Esse teste influenciou estudos posteriores e recebeu diversas adaptações metodológicas (e.g., JUST; CARPENTER, 1992; WATERS *et al.*, 1987; WATERS; CAPLAN, 1996; WALTER, 2007).

Van den Noort *et al.* (2008) desenvolveram uma versão computadorizada e padronizada do RST em quatro línguas (alemão, holandês, inglês e norueguês) com o objetivo de se comparar a padronização dos materiais entre as línguas. Essa versão do RST apresenta algumas adaptações em relação ao controle dos materiais e à aplicação do teste, tais como: 1) controle do tamanho das sentenças por meio do número de a. palavras, b. sílabas e c. letras; 2) controle da frequência das palavras-alvo no final das frases; 3) controle de concretude das palavras-alvo; 4) controle de plausibilidade das sentenças utilizadas; 5) apresentação aleatória dos conjuntos de 2 a 6 frases; e 6) pontuação calculada a partir do total de palavras recordadas no experimento. Na seção seguinte será apresentada a metodologia de desenvolvimento, construção, aplicação e análise do experimento aqui desenvolvido, a VCPRST-PB.

4. Metodologia

A VCPRST-PB foi desenvolvida nos programas Paradigm Experiments¹ (Perception Research Systems Incorporated, EUA) e DMDX² (DMASTR, Tucson, EUA) (FORSTER; FORSTER, 2003). A versão-Paradigm pode ser aplicada gratuitamente através da utilização do programa Paradigm Player³, contudo, para edição do experimento, uma licença válida é necessária. Já a versão-DMDX possui código aberto, podendo ser aplicada e editada gratuitamente.

4.1. Teste de Concretude

A versão do RST proposta por van den Noort *et al.* (2008) aponta a importância do teste de concretude para as palavras-alvo. Palavras concretas têm mais chances de serem recuperadas pela MT devido ao processamento imagético diferente das palavras abstratas (JANCZURA *et al.*, 2007). Os autores apontam que palavras mais concretas tendem a gerar imagens mentais que podem ser usadas como recursos na execução de tarefas que envolvam a memória e a linguagem.

Com o objetivo de se selecionarem as palavras-alvo para VCPRST-PB, foram selecionados 120 substantivos concretos a partir do corpus Léxico do Português Brasileiro (LexPorBR) (ESTIVALET; MEUNIER, 2015) obedecendo os seguintes critérios: a. frequência entre 100 e 500 (por milhão de palavras), b. entre 2 e 4 sílabas, c. entre 5 e 8 letras, e, d. entre 0 e 7 vizinhos ortográficos. Essas palavras foram testadas em um experimento de concretude com escala Likert de 5 pontos, aplicado através do Google Forms em 80 participantes (maiores de 18 anos e universitários). Um teste

¹ <http://www.paradigmexperiments.com/>

² [http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm/](http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm)

³ <http://www.paradigmexperiments.com/ParadigmPlayer/ParadigmPlayer.html>

estatístico de chi-quadrado de independência (*chi-square test*) com distribuição binomial foi aplicado para se verificar a relação entre as respostas 5 e demais respostas da escala Likert do teste de concretude. Os resultados foram estatisticamente significativos χ^2 (476, N = 80) = 883,13, $p < 0,001$, sugerindo que todas as 120 palavras do teste de concretude receberam mais notas 5 do que as demais notas. Assim, foram selecionadas as 100 palavras com menor p-valor para utilização na VCPRST-PB.

4.2. Teste de Plausibilidade

Para cada uma das 100 palavras-alvo selecionadas, procuraram-se no Corpus do Português⁴ duas frases que contivessem as referidas palavras-alvo como palavras finais, totalizando 200 sentenças. Em seguida, essas frases foram adaptadas com o objetivo de apresentarem naturalidade de ocorrência no PB, bem como se adequarem aos critérios de controle em relação ao número de: a. palavras (12 a 16), b. sílabas (24 a 28) e c. letras (55 a 65), visando homogeneidade das sentenças. A partir dessas frases, realizou-se um teste de julgamento de plausibilidade (MORAES *et al.*, 2016) das mesmas em uma escala Likert de 5 pontos através do Google Forms. Devido ao grande número de frases, o teste foi dividido em quatro blocos, contendo 50 frases cada. Cada bloco de frases foi realizado por 30 participantes (maiores de 18 anos e universitários), totalizando 120 participantes.

Para cada frase, calculou-se um escore (soma do número de participantes X escala Likert de resposta) e em cada um dos pares que continham a mesma palavra-alvo selecionou-se a frase com maior escore, ou seja, com maior plausibilidade. Além disso, novamente um teste estatístico de chi-quadrado de independência com distribuição binomial foi aplicado para se verificar a diferença entre as respostas 5 e demais respostas das frases de cada par do teste de plausibilidade. Os resultados foram estatisticamente significativos χ^2 (396, N = 30) = 507,07, $p < 0,001$, sugerindo que em cada par de frases, uma delas possuía maior plausibilidade que a outra e corroborando a seleção das frases com maior escore. Em seguida, a partir das 100 frases mais plausíveis selecionadas, realizou-se um teste estatístico teste-z (*z-score test*) para se verificar a diferença entre as respostas das frases selecionadas. Os resultados não foram significativos ($z = 0,819$, $p = 0,41$), indicando que não houve diferença de plausibilidade entre as 100 frases selecionadas, sendo todas elas perfeitamente plausíveis para utilização na VCPRST-PB.

4.3. Versão Computadorizada Padronizada do *Reading Span Test* para o PB

4.3.1. Participantes

A VCPRST-PB foi aplicada em 23 participantes voluntários, sendo que 12 destes realizaram o teste na modalidade oral e 11 na modalidade escrita. Todos os participantes eram maiores de 18 anos (média de idade 21 anos), falantes nativos do PB e cursavam ou já concluíram o ensino superior.

⁴ <https://www.corpusdoportugues.org/>

4.3.2. Materiais

As 100 frases finais foram distribuídas em 5 séries de 20 frases. Em cada série, foram construídos conjuntos de 2, 3, 4, 5 ou 6 frases. As frases em cada série foram controladas conforme o número de a. palavras, b. sílabas, c. letras das frases, assim como ao controle de a. sílabas, b. letras, c. vizinhos ortográficos e d. frequência das palavras-alvo, conforme apresentado na Tabela 1.

Tabela 1: Médias e desvios-padrão entre parênteses das variáveis de controle para a elaboração das frases da VCPRST-PB

	<i>Palavras frase</i>	<i>Sílabas frase</i>	<i>Letras frase</i>	<i>Sílabas palavra</i>	<i>Letras palavra</i>	<i>Viz. orto. palavra</i>	<i>Frequência palavra</i>
<i>Série 1</i>	13,1 (0,9)	26,5(1,1)	59,5 (2,3)	2,7 (0,8)	6,3 (1,1)	2,1 (2,0)	274 (138)
<i>Série 2</i>	13,3 (1,2)	26,4(1,3)	59,2 (3,0)	2,7 (0,7)	6,3 (0,8)	2,5 (1,9)	260 (114)
<i>Série 3</i>	13,0 (1,0)	26,2(1,1)	58,4 (2,9)	2,6 (0,7)	6,5 (1,1)	2,2 (1,6)	239 (101)
<i>Série 4</i>	12,9 (1,0)	26,6(1,2)	59,1 (3,3)	2,8 (0,8)	6,1 (1,2)	2,2 (2,3)	230 (98)
<i>Série 5</i>	12,9 (0,8)	26,2(1,2)	59,6 (3,3)	2,5 (0,6)	6,3 (1,0)	2,6 (2,2)	244 (92)
Total	13,0 (1,0)	26,4(1,2)	59,1 (3,0)	2,7 (0,7)	6,3 (1,0)	2,3 (2,0)	249 (109)

Visando reduzirem-se possíveis interferências e estratégias de memorização das palavras-alvo, as séries foram organizadas em ordem pseudo-randomizada, evitando-se i. a sequência de palavras iniciadas com a mesma letra e ii. a sequência de frases longas e curtas no mesmo conjunto. Segue abaixo o exemplo de um conjunto de duas frases seguidas das palavras que devem ser recordadas, conforme Quadro 1.

Quadro 1: Exemplo de um conjunto de 2 frases, seguidas das palavras-alvo a serem recordadas na VCPRST-PB

“Use muita água e sabão para evitar manchas e depois seque com uma toalha.” TOALHA
“O jornal português divulgou que o atleta tem uma suspeita de fratura no nariz.” NARIZ

4.3.3. Procedimentos

Cada frase é apresentada individualmente no centro da tela do computador durante oito segundos (8.000 ms) ou até que o participante aperte a tecla “espaço”. Os conjuntos são apresentados aos participantes de forma aleatória em cada uma das 5 séries do experimento. O participante inicia o experimento com uma tela de instruções informando que ele deverá 1) ler em voz alta as frases apresentadas o mais rapidamente possível, 2) memorizar a última palavra de cada frase e 3) ao final de cada conjunto de frases, quando solicitado, recordar e dizer em voz alta (ou escrever, na versão escrita) as palavras memorizadas. Em seguida, o participante realiza um treino com três conjuntos de frases (2, 3 e 4 frases) com a finalidade de se habituar à tarefa. Finalmente, o participante realiza o teste completo contendo as 100 frases experimentais.

5. Resultados e Discussão

5.1. O experimento: VCPRST-PB

Em relação à seleção dos estímulos do experimento apresentado aqui, destaca-se que todos os materiais foram selecionados a partir de corpora representativos e controlados a partir de testes de aceitabilidade (Teste de Concretude para as palavras-alvo e Teste de Plausibilidade para a aceitação das frases experimentais), minimizando-se a interferência dos pesquisadores na seleção dos materiais. Conforme Moraes *et al.* (2016), o teste de plausibilidade possibilita a criação de frases mais aceitáveis no mundo real, mostrando-se um bom instrumento para as escolhas dos estímulos dos experimentos psicolinguísticos, garantindo maior controle dos materiais nas condições experimentais e nas variáveis independentes.

Com o objetivo de se acrescentar maior rigor na seleção dos materiais que no experimento proposto por van den Noort *et al.* (2008), além da frequência e tamanho das palavras-alvo, controlou-se também o número de vizinhos ortográficos, variável importante nos processos de leitura e acesso lexical (ESTIVALET; MEUNIER, 2015). As médias e os desvios-padrão das frequências das palavras-alvo por séries foram melhor controlados que no experimento de van den Noort *et al.* (2008), reforçando a pertinência dos materiais utilizados e a confiança nos resultados da VCPRST-PB.

A fim de verificar-se o equilíbrio e o controle das sentenças nas diferentes séries, o tempo de leitura (TL) de cada sentença foi registrado (VAN DEN NOORT *et al.*, 2008). Realizou-se um teste estatístico de Análise de Variância (ANOVA) com as variáveis: média do TL das séries como variável dependente; séries (5) e modalidade (oral ou escrita) como variáveis independentes. Os resultados apontaram um efeito principal não-significativo entre as séries $F(4, 538) = 1,87$, $p = 0,11$, mas um efeito principal significativo de modalidade $F(1, 538) = 26,33$, $p < 0,001$. A falta de efeito significativo entre as séries sugere que os materiais utilizados na VCPRST-PB foram satisfatoriamente controlados, não havendo diferenças significativas do TL em função dos materiais utilizados e da distribuição dos materiais entre as 5 séries do experimento. As médias dos TL da VCPRST-PB foram maiores que van den Noort *et al.* (2008), este resultado possivelmente pode ser explicado pelo fato que, de uma forma geral, as palavras do PB possuem em média mais letras que as palavras das línguas testadas pelos autores, e na VCPRST-PB, manteve-se o critério em relação ao número de palavras das frases, aumentando-se o número médio de letras em relação àquele estudo.

Em relação às diferenças de TL entre as duas modalidades, os participantes leram as sentenças na modalidade escrita significativamente de maneira mais rápida que na modalidade oral. Esta diferença se deve, provavelmente, à necessidade dos participantes lerem as frases rapidamente com o objetivo de manterem, de forma simultânea, a repetição das palavras memorizadas na alça fonológica para produção oral, diferente do processo de retenção da forma ortográfica na alça visuo-espacial para produção escrita (BADDELEY, 1986).

5.2. Capacidade de memória de trabalho: VCPRST-PB

Em relação à pontuação do teste de capacidade de MT, a VCPRST-PB permite calcular três pontuações com características diferentes. A primeira pontuação nomeada “Span

MT” calcula o tradicional *span* máximo de capacidade de MT estabelecido por Daneman e Carpenter (1980). A segunda, “Memória Total” calcula de forma flexível o total de palavras-alvo recordadas em todo experimento, conforme van den Noort *et al.* (2008). Já a terceira, “Span Conjunto”, proposta neste trabalho, calcula a média do total de palavras recordadas nos conjuntos em que todas as palavras foram relembradas. Essa última pontuação proposta tem o objetivo de oferecer uma medida menos conservadora que o “Span MT” e mais robusta do que “Memória Total”, apresentando um coeficiente híbrido entre estes extremos.

No que diz respeito à análise das pontuações da capacidade de MT, os dados foram analisados através de três ANOVAs com as diferentes pontuações como variável dependente e séries (5) e modalidade (oral e visual) como variáveis independentes. A pontuação “Span MT” (DANEMAN; CARPENTER, 1980) não apresentou diferenças significativas de série $F(4, 98) = 0,31, p = 0,86$, nem de modalidade $F(1, 98) = 0,74, p = 0,38$, indicando que a seleção dos materiais foi bem controlada e não provocou diferenças de capacidade de MT, assim como a modalidade de resposta parece não influenciar essa capacidade. A pontuação “Memória Total” (VAN DEN NOORT, 2008) também não apresentou diferenças significativas de série $F(4, 98) = 0,61, p = 0,65$, nem de modalidade $F(1, 98) = 0,27, p = 0,60$, corroborando que os materiais foram equilibradamente distribuídos nas séries e não provocam diferenças de capacidade de MT, não ocorrendo influência também da modalidade de resposta. A pontuação “Span Conjunto”, proposta aqui, também não apresentou diferenças significativas de série $F(4, 98) = 0,18, p = 0,94$, nem de modalidade $F(1, 98) = 0,03, p = 0,84$, apontando mais uma vez a eficácia do controle dos materiais distribuídos entre as séries, assim como a indiferença entre a modalidade de resposta na capacidade de MT. Os resultados das médias e desvios-padrão dos TL e diferentes pontuações de capacidade de MT por séries e modalidades são apresentados na Tabela 2.

Tabela 2: Médias e desvios-padrão entre parênteses das 3 pontuações de capacidade de MT e dos TL das frases por séries e modalidade.

	<i>Span MT</i>	<i>Memória Total</i>	<i>Span Conj.</i>	<i>Média do TL (ms)</i>
<i>Escrito</i>				
<i>Série 1</i>	3,0 (1,0)	11,5 (2,6)	2,5 (0,9)	5625 (1346)
<i>Série 2</i>	3,1 (0,9)	12,8 (3,3)	2,7 (0,8)	5799 (1253)
<i>Série 3</i>	2,7 (1,0)	12,6 (2,8)	2,6 (0,8)	5662 (1336)
<i>Série 4</i>	2,8 (1,0)	12,1 (3,1)	2,6 (0,9)	5608 (1365)
<i>Série 5</i>	2,8 (1,1)	12,0 (3,4)	2,6 (0,8)	5702 (1169)
<i>Total</i>	2,8 (1,0)	12,2 (3,1)	2,6 (0,8)	5681 (1249)
<i>Oral</i>				
<i>Série 1</i>	2,8 (0,6)	11,0 (2,9)	2,4 (0,6)	5822 (1461)
<i>Série 2</i>	2,6 (0,6)	11,9 (2,2)	2,4 (0,6)	6478 (1051)
<i>Série 3</i>	2,7 (0,8)	12,2 (2,1)	2,4 (0,7)	6128 (1278)
<i>Série 4</i>	3,0 (1,0)	12,2 (2,3)	2,6 (0,8)	6385 (1023)
<i>Série 5</i>	2,5 (0,6)	12,3 (2,7)	2,4 (0,6)	6222 (1213)
<i>Total</i>	2,7 (0,7)	11,9 (2,4)	2,5 (0,6)	6210 (1166)

Destaca-se que o “Span da MT” (DANEMAN; CARPENTER, 1980), medida clássica e mais conservadora, considera de forma progressiva o limite da capacidade de MT, porém levando em conta somente a capacidade máxima dessa capacidade. Já, a

“Memória Total” (VAN DEN NOORT *et al.*, 2008) é uma medida muito flexível e geral que não considera o limite de capacidade da MT, uma vez que leva em conta todas palavras relembradas, independentemente da capacidade máxima por conjunto de frases.

Por sua vez, a pontuação “Span Conjunto”, proposta neste trabalho, mede a capacidade de MT ao longo das 5 séries, pontuando de forma ponderada o número de palavras totais recordadas apenas nos conjuntos de frases que o participante acertou todas as palavras do conjunto. Observa-se que as médias dessa pontuação foram menores que na medida tradicional porque a “Span Conjunto” considera o número de palavras recordadas em todos conjuntos que todas palavras foram relembradas. Portanto, tendo em vista que, em geral, a maior parte dos participantes acerta muitos conjuntos com poucas frases (2-3) e poucos conjuntos com muitas frases (4-6), a média possui uma tendência para baixo, diferente da medida “Span MT” que considera somente o maior conjunto que todas palavras foram relembradas. Sendo assim, as características da pontuação “Span Conjunto” indicam que ela pode ser utilizada como medida principal, flexível e objetiva, da capacidade de MT.

6. Conclusões

A relação entre a MT e o processamento linguístico permite a investigação do processamento da linguagem por inúmeros caminhos de hipóteses. O presente trabalho se comprometeu com o desenvolvimento da VCPRST-PB, tornando-se um modelo de teste de capacidade de MT em PB comparável a outras línguas. Como resultado final, obteve-se um experimento com material altamente controlado, uma vez que se realizou o controle nas frases quanto ao número de palavras, sílabas e letras, e, nas palavras-alvo quanto à frequência, número de letras, sílabas e vizinhos ortográficos. Todas as palavras e as frases foram retiradas e controladas através da pesquisa de corpora, obtendo-se um teste de capacidade de MT padronizado e computadorizado, que por sua vez, pode ser comparado com outros estudos internacionais (VAN DEN NOORT *et al.*, 2008). Espera-se que a versão do teste criada e apresentada aqui possa contribuir para investigação da MT e estudos que desejam testar e/ou controlar a capacidade de MT. Todos experimentos do Paradigm e do DMDX, escritos e orais, da VCPRST-PB podem ser baixados em: <http://www.lexicodoportugues.com/experimentos/vcprst-pb>.

Referências

- Baddeley, A. D.; Hitch, G. (1974). Working Memory. In: Bower, G.A. (Ed). *Recent advances in learning and motivation*. New York: Academic Press.
- Baddeley, A.D. (1986). *Working memory*. New York: Oxford University Press.
- Daneman, M.; Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Daneman, M.; Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General*, 30(2), 208-223.
- Estivalet, G. L.; Meunier, F. (2015). The Brazilian Portuguese Lexicon: An instrument for psycholinguistic research. *PLoS ONE*, 10(12), e0144016.

- Forster, K.; Forster, J. (2003). DMDX: A Windows Display Program with Millisecond Accuracy. *Behavior Research Methods: Instruments and Computers*, 35(1), 116-124.
- Janczura, G. A.; Castilho, G. M.; Rocha, N. O.; van Erven, T. J. C. (2007). Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa*, 23(2), 195-204.
- Just, M. A.; Carpenter, P. A. (1978). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87 (4), 329-354.
- Just, A. M.; Carpenter, P. A. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review*, 99(1), 122-149.
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.
- Moraes, B. M.; Leite, J. E. R.; Soares, A. P.; Oliveira, H. M. (2016). A importância do teste de plausibilidade na validação de frases em experimentos psicolinguísticos. *Revista Prolíngua*, 11(1), 17-26.
- Mota, M. B. (2015). Sistemas de memória e processamento da linguagem: um breve panorama. *Revista Lingüística*, 11 (1), 205- 215.
- Uehara, E.; Landeira-Fernandez, J. (2010). Um panorama sobre o desenvolvimento da memória de trabalho e seus prejuízos no aprendizado escolar. *Ciências & Cognição*, 15(2), 31-41.
- Van den Noort, M.; Bosh, P.; Haverkot, M.; Hugdahl, K. (2008). A Standard Computerized Version of the Reading Span Test in Different Languages. *European Journal of Psychological Assessment*, 24(1), 35-42.
- Walters, G. S.; Capland, D.; Hildebrandt, N. (1987). Working memory and written sentence comprehension. In: M. Coltheart (Ed.). *Attention and Performance XII, The psychology of reading*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc. 531- 555.
- Walters, G. S.; Capland, D. (1996). The Measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology*, 49A, 51-75.
- Walter, C. (2007). First-to second-language reading comprehension: not transfer, but access. *International Journal of Applied Linguistics*, 17(1), 14-37.
- Rodrigues, C. (2001). Contribuições da memória de trabalho para o processamento da linguagem: Evidências experimentais e clínicas. *Working Papers em Linguística UFSC*, 5, 124-144.

Acesso Lexical de Formas Irregulares Flexionadas em Número em Português Brasileiro

Jefferson Alves da Rocha¹, José Ferrari-Neto¹

¹Programa de Pós-graduação em Linguística (PROLING)

Universidade Federal da Paraíba (UFPB)

Campus I - Cidade Universitária - Castelo Branco - João Pessoa, PB - Brasil

jefferson.rocha16@gmail.com, joseferrarin@ibest.com.br

Abstract. This paper investigates the access and representation of inflected nominal forms in number in Brazilian Portuguese (BP). Initially, three different hypotheses were observed from the psycholinguistic point of view, that is, different ways of conceiving the mode of storage and retrieval of a given irregular word in the mental lexicon. Subsequently, these hypotheses were related to the present study. Here, an experimental masked priming paradigm was used. The stimuli present in this priming experiment were selected from the type of morphological irregularity of the lexical items. The results showed that the irregular forms are accessed through their base form, which leads to a decomposition model of access and lexical representation

Keywords: Mental lexicon. Number bending. Priming.

Resumo. Este trabalho procurou investigar o acesso e a representação de formas nominais flexionadas em número em Português Brasileiro (PB). De início, observaram-se três diferentes hipóteses do ponto de vista psicolinguístico, isto é, formas distintas de se conceber o modo de armazenamento e de recuperação de uma dada palavra irregular no léxico mental. Posteriormente, relacionaram-se essas hipóteses ao presente estudo. O estudo em questão consistiu em um priming encoberto. Os estímulos presentes no experimento de priming foram selecionados a partir do tipo de irregularidade dos itens morfológicos. Os resultados apontaram que as formas irregulares são acessadas por meio da sua forma de base, o que se direciona para um modelo decomposicional de acesso e representação lexical.

Palavras-chave: Léxico mental. Flexão de número. Priming.

1. Introdução

O estudo do léxico mental, compreendido como o repositório dos itens lexicais de uma língua, é concorrente, em linhas gerais, à investigação sobre como esses itens são nele estocados, assumindo-se uma noção que assemelha o léxico mental a um sistema de memória de longo prazo para palavras e elementos constituintes. A maneira como se dá essa estocagem acaba por determinar, em grande medida, a forma como eles são recuperados, em operações de geração de sentenças em uma dada língua. Acesso e recuperação são, portanto, os dois grandes pontos de investigação no âmbito de uma teoria do léxico mental.

Porém, algumas questões auxiliares também acabam por se destacar, quando se investigam as duas questões iniciais. Uma dela diz respeito à natureza da informação a ser armazenada – que tipo de item constitui uma unidade armazenável? Pode-se pensar que essas unidades são morfemas – unidades mínimas significativas – o que leva a necessidade de se postularem regras que combinam essas unidades no momento em que são recuperadas do léxico. Ou pode-se propor que são palavras, restando, então, a caracterização do que vem a ser de fato uma palavra – formas flexionadas de um mesmo item contam como uma unidade armazenável? Se sim, o mesmo vale para as formas com flexão irregular?

Problemas como esse têm sido investigados na pesquisa psicolinguística sobre o acesso e a representação lexical. O presente trabalho tem por objetivo apresentar dados empíricos que, longe de encerrar a questão, visam principalmente contribuir para a discussão sobre as teorias que vêm sendo propostas a respeito. Ele se organiza da seguinte forma: na seção 2, tratamos de questões relacionadas às teorias sobre o léxico mental; na seção 3, trazemos a descrição do estudo realizado no alemão, que inspirou o estudo realizado em PB; na seção 4 deste trabalho, será tratado do referido estudo em PB. Por fim, trazemos as considerações finais.

2. Teorias sobre o Léxico Mental

De acordo com estudos psicolinguísticos, o processamento de palavras morfológicamente complexas se desenvolve de duas maneiras distintas. O modo de armazenamento de uma dada palavra fornece algumas indicações sobre como se pode conceber o acesso lexical. Um dos fatores presentes nos modelos de processamento de palavras flexionadas é a decomposição entre raiz e afixos (GIRAUDO & GRAINGER, 2000; MARINKOVIC, 2004; STOCKALL & MARANTZ, 2006), por exemplo, em oposição ao acesso de formas por completo no léxico.

Outros estudos, diferentemente dos apontados anteriormente, consideram que todas as formas complexas são compostas em um momento inicial de reconhecimento de palavras. (TAFT, 1979, 1981, 1988; TAFT & FORSTER, 1975; TAFT et al., 1986.). Podem-se mostrar, ainda, estudos que consideram um modelo com duas rotas de acesso para o reconhecimento de palavras. O acesso pode se basear em formas completas de palavras ou por meio da decomposição dos morfemas. Dentre os modelos de dupla rota de acesso, pode-se mencionar, inicialmente, o modelo de Morfologia Endereçada Aumentado - *Augmented Addressed Morphology model* - (BURANI & CARAMAZZA, 1987; CARAMAZZA et al., 1988). O modelo diz que a rota de acesso pela palavra por inteiro tem precedência. O acesso pela rota de análise dos morfemas só seria ativado se o acesso por inteiro falhasse. Já o modelo de dupla rota assume a ativação em paralelo das duas formas de acesso. (FRAUENFELDER & SCHREUDER, 1992; SCHREUDER & BAAYEN, 1995; BAAYEN et al., 1997). Outros estudos se direcionam para o modelo de Mecanismo Duplo - *Dual Mechanism model* - (PINKER & PRINCE, 1994; CLAHSEN et al., 1997; CLAHSEN, 1999; SONNENSTUHL et al., 1999.). Esse modelo consiste em analisar a distinção entre flexão padrão e irregular, procurando determinar como a forma das palavras complexas é representada e processada. Em relação à forma das palavras padrão, pode-se dizer que são representadas por base e afixo, processadas ao se inserir elementos de uma dada categoria sintática, independentemente das propriedades fonológicas e semânticas. Por outro lado, palavras

irregulares são armazenadas por inteiro como formas completas, ressalvadas questões de frequência.

Em português brasileiro (PB), pode-se investigar essa questão analisando-se o acesso e a representação de formas flexionadas em número. Isso porque essa língua apresenta uma morfologia de plural que distingue formas regulares e formas irregulares quanto à expressão do número gramatical. Em PB, a flexão regular se dá pelo acréscimo do morfema -s ao final das palavras com tema em -a, -e ou -o. Já a irregular se manifesta em nomes com terminações -r, -s, -z ou -m, e -ão, o que faz com que o morfema de plural sofra alomorfia. Estudos sobre o acesso e a representação dessas formas no léxico mental têm apontado que efeitos de frequência influenciam o tempo de reconhecimento de palavras em PB. (ROCHA, 2016) Assim, o estudo do processamento de formas flexionadas em número em PB se configura como uma forma de se prover evidências empíricas sobre a organização dos itens lexicais no léxico mental, seu acesso e sua representação.

3. Experimento no alemão

Num experimento realizado por Sonnenstuhl e Huth (2002), investigaram-se os plurais terminados em -n no alemão, por meio de um experimento de *priming*. A hipótese assumida neste estudo foi a de que prováveis informações lexicais são diretamente influenciadas no reconhecimento do alvo, já que a entrada correspondente já teria sido ativada pelo *prime*. Dessa forma, pôde-se investigar se o -n de plurais em alemão é representado de forma completa no léxico, não sofrendo decomposição quando de seu acesso, ou se sua representação se dá por meio de seus morfemas constituintes, o que levaria a um acesso por decomposição. Foram testados quatro tipos de plurais. Para cada tipo de plural foram selecionados trinta nomes como itens de teste. O alvo em todas as condições foi o singular. Cada alvo foi apresentado em uma condição idêntica (*blume/blume*), uma condição plural-singular (*blumen/ blume*) e uma condição controle (*wand/ blume*). Cada item foi controlado por frequência e por tamanho de sílabas em cada tipo de plural. De acordo com os pressupostos experimentais, os nomes femininos que possuem o n de plural tendem a ser mais frequentes em relação aos nomes não femininos com o afixo n de plural. Mediram-se os efeitos do *priming* morfológico, comparando o tempo de resposta da mesma tarefa na condição plural e na condição idêntica, e, separadamente, também a condição controle para cada tipo de plural.

Os resultados apontaram dados significativos em relação por tipo de *prime*, por sujeito e por tipo de plural. Obteve-se efeito principal para tipo de plural e que tal efeito foi distinto nos tempos de reação das formas controle. Houve também efeito de *priming*, já que o tempo de reconhecimento nas condições idênticas e plurais foi significativamente mais curto em relação às condições controle. De acordo com os autores, esses efeitos de facilitação indicaram que as entradas lexicais do alvo foram ativadas pelo *prime*. Esses resultados confirmaram que os plurais irregulares não são decompostos entre a base e o afixo. Desse modo, podem ser representados como formas completas no léxico mental.

4. Experimento em PB: *Priming* Encoberto

Procurou-se investigar a flexão de número por meio de quatro tipos distintos de formação de plural existentes nessa língua. Para testar possíveis diferenças entre os plurais irregulares em PB, utilizou-se uma técnica de *priming* encoberto, montada no programa *Paradigm*. Essa técnica experimental consiste em um processo de ativação lexical, já que apresenta informações sobre as entradas lexicais das formas flexionadas em número. Com isso, pode-se analisar se a segunda palavra ativaria a representação da entrada lexical da primeira palavra. Isso possibilitaria a investigação sobre a representação e o acesso dessas formas, especificamente, em relação ao *priming*. Informações lexicais podem ser influenciadas no reconhecimento do alvo, já que a entrada correspondente já foi ativada pelo *prime*. Apresentaram-se pares de palavras na tela do computador, a primeira palavra foi o *prime* e a segunda foi o alvo.

4.1 Método

4.1.1 Estímulos

Assim como no experimento realizado no alemão, aqui também utilizamos quatro tipos de plurais irregulares. Para cada tipo de plural foram selecionados doze pares de substantivos. O alvo em todas as condições era o singular. Cada alvo foi apresentado em uma condição idêntica, uma condição plural e uma condição controle. Em uma dada palavra como *cidadão*, por exemplo, cujo plural é *cidadãos*, foi inserida dentre os estímulos experimentais na condição idêntica (*prime*: cidadão - alvo: cidadão), na condição plural (*prime*: cidadãos - alvo: cidadão) e na condição controle (*prime*: abdômen - alvo: cidadão), por exemplo. A seguir, podem-se observar exemplos dos estímulos distribuídos pelos quatro tipos de plurais e pelas condições experimentais.

Tabela 1. Descrição e exemplos de estímulos experimentais.

TIPO DE PLURAL	CONDIÇÃO	PRIME	ALVO
I - Palavras terminadas em <i>ão</i> no singular. O plural se realiza por <i>ões</i> e <i>ãos</i> .	IDÊNTICA	<i>coração</i>	<i>coração</i>
	PLURAL	<i>corações</i>	<i>coração</i>
	CONTROLE	<i>manequim</i>	<i>coração</i>
II - Palavras terminadas em <i>r</i> ou <i>z</i> no singular. O plural é realizado por meio do acréscimo de <i>es</i> .	IDÊNTICA	<i>ditador</i>	<i>ditador</i>
	PLURAL	<i>ditadores</i>	<i>ditador</i>
	CONTROLE	<i>casaco</i>	<i>ditador</i>
III - Palavras terminadas em <i>al</i> , <i>el</i> e <i>ol</i> no singular. No plural, há a supressão do <i>l</i> e acréscimo de <i>is</i> .	IDÊNTICA	<i>avental</i>	<i>avental</i>
	PLURAL	<i>aventais</i>	<i>avental</i>
	CONTROLE	<i>camarão</i>	<i>avental</i>
IV - Palavras terminadas em <i>m</i> no singular. No plural, há a supressão do <i>m</i> e acréscimo de <i>ns</i> .	IDÊNTICA	<i>armazém</i>	<i>armazém</i>
	PLURAL	<i>armazéns</i>	<i>armazém</i>
	CONTROLE	<i>auditor</i>	<i>armazém</i>

Inicialmente, realizou-se uma divisão por quatro tipos de plurais irregulares em PB. De acordo com a tabela anterior, o primeiro grupo foi composto por palavras terminadas em *ão* no singular e o plural se realizou por *ões* e *ãos*. O segundo grupo foi composto por palavras terminadas em *r* ou *z* no singular e o plural foi realizado por meio do acréscimo de *es*. O terceiro grupo foi composto por palavras terminadas em *al*, *el* e *ol* no singular e, no plural, houve a supressão do *l* e acréscimo de *is*. O quarto grupo

foi composto por palavras terminadas em *m* no singular e, no plural, houve a supressão do *m* e acréscimo de *ns*.

Posteriormente, verificou-se a frequência dessas palavras no *Corpus do Português* para inserção delas no experimento. Nesse *corpus*, investigamos as frequências das formas do singular e do plural correspondente. Para dissuadir os participantes do experimento, acrescentaram-se setenta e dois pares de nomes, que funcionaram como itens distratores. Acrescentou-se, ainda, à lista de palavras do experimento, cento e vinte pseudopalavras. Para a construção das pseudopalavras, realizou-se a troca de sílabas de palavras experimentais e de distratoras.

4.1.2 Participantes

Para este experimento, participaram noventa estudantes da graduação e da pós-graduação da UFPB (Universidade Federal da Paraíba). Todos foram testados individualmente na sala do Laboratório de Processamento Linguístico (LAPROL), situado na UFPB.

4.1.3 Variáveis

As variáveis independentes foram: *tipo de plural irregular* e *tipo de relação entre prime-alvo (idêntica, plural ou controle)*. As variáveis dependentes foram: *o tempo de reação* e *o número de acertos*.

4.1.4 Procedimento

De início, explicou-se para o participante no que consistia o experimento proposto por meio da tarefa de *priming*. Apresentou-se a instrução do experimento, na tela do computador, antes do inicio do teste experimental. A instrução dizia que o participante veria uma sequência de letras seguida de outra sequência de letras no centro da tela do computador. Essas sequências seriam antecedidas por uma sequência de fixação de # (*hashtag*). A instrução dizia ainda que a tarefa do participante era identificar se a segunda sequência de letras era ou não uma palavra do português. O procedimento experimental permitia ao participante apertar a tecla verde do teclado em caso afirmativo ou a tecla vermelha em caso negativo para a pergunta realizada. Dizia-se, ainda, que para o inicio do experimento, o participante deveria clicar na barra de espaço.

Os estímulos foram expostos em letras brancas, fonte *New Times Roman*, tamanho doze, em uma tela de fundo preto. O *prime* ficou exposto por 50ms e o *alvo* ficou exposto até o tempo de reação do participante. Por fim, foram apresentados os agradecimentos pela participação do sujeito, finalizando o experimento. Cada participante levou em média quinze minutos para responder o teste.

4.1.5. Resultados

Realizamos duas análises, assim como foi feito no estudo dos plurais irregulares no alemão. A primeira análise consistiu na análise dos tipos de plurais e a segunda consistiu na análise do efeito de *priming*.

Os tempos médios de reação foram os seguintes: tipo A - 850ms; tipo B - 878ms; tipo C - 907ms; tipo D - 833ms. Os tempos médios de reação dos tipos de plurais se mostraram diferentes um dos outros. Pode-se observar que o tipo D teve o

menor tempo de reação, seguido pelo tipo A e pelo tipo B, já o tipo C obteve maior tempo de reação.

Desse modo, as diferenças nos tempos de reação dos tipos de plurais podem ser explicadas pela existência de diferentes níveis de complexidade na formação do plural irregular em PB. O tipo D, por exemplo, seria composto por palavras como *jardim*. Nesse caso, esse tipo de formação do plural com supressão do *m* e acréscimo de *ns* se aproximaria mais de um tipo de plural regular. O tipo A seria mais complexo que o D e menos complexo do que os outros dois tipos. Em pares de palavras como *cidadão/cidadões* só haveria o acréscimo do morfema *s*, indicativo de plural também em formas regulares. Já em pares de palavras como *camarão/camarões* haveria uma mudança em relação à vogal nasal do vocábulo, culminando na alteração de *ão* para *ões*. Essa mudança não aumentaria a complexidade morfológica, nesse caso. Em relação ao tipo B, os vocábulos na formação do plural consistem no acréscimo do morfema *es* em *caçador/caçadores* ou *capataz/capatazes*. O nível de complexidade mais elevado foi observado no tipo C. Nesse caso, a formação do plural em *l* se tornou mais custosa do ponto de vista do acesso lexical. Os pares de palavras como *animal/animais*, *coquetel/coquetéis* e *cachecol/cachecóis* formam o plural com a supressão do *l* e acréscimo de *is*. A possibilidade de diferentes formas no singular formarem palavras no plural com a mesma terminação morfológica aumentaria o tempo de reação e tornando o tipo de plural com nível de complexidade mais elevado.

Tabela 2. ANOVA - Tipos de plural.

Tabela da ANOVA					
	GL.	SomadeQuadrados	QuadradoMédio	Estat..F	P.valor
Fator	3	3041873,739	1013957,913	5,796259272	0,000598095
Resíduos	3894	681189699,8	174933,1535		

De acordo com a tabela dois, pode-se observar que o P-valor foi significativo, além disso, assevera-se que esse resultado, portanto, mostra que há diferenças entre os tipos de plurais irregulares analisados no teste. Essas diferenças podem ser percebidas nas tabelas a seguir.

Tabela 3. Médias - tipos de plural.

Agrupamentos		
Fator	Médias	Grupos
Tipo C	907,2571	a
Tipo B	878,0627	ab
Tipo A	850,6502	b
Tipo D	833,5839	b

A tabela três apresenta os tempos de reação dos tipos de plurais irregulares no estudo em questão. Aqui, podem-se observar as diferenças existentes ao agrupar os tipos de plurais. Desse modo, de acordo com a tabela, o tipo A e o tipo D foram iguais, representados por *b*. O tipo C foi diferente dos outros três tipos, representado por *a*. O tipo B foi igual ao tipo C e igual aos tipos A e D, representado por *ab*. Pode-se observar na tabela a seguir, a comparação realizada entre os tipos de plurais.

Tabela 4. Comparações entre tipos de plurais.

Comparações Múltiplas				
Níveis	Centro	Limite Inferior	Limite Superior	P-valor
Tipo B - Tipo A	27,41257305	-21,22582121	76,05096731	0,469073968
Tipo C - Tipo A	56,6069206	7,870077215	105,343764	0,015121956
Tipo D - Tipo A	-17,06629365	-65,95348401	31,8208967	0,806288825
Tipo C - Tipo B	29,19434755	-19,31840784	77,70710295	0,409645546
Tipo D - Tipo B	-44,4788667	-93,14266139	4,18492799	0,087379209
Tipo D - Tipo C	-73,67321425	-122,4354068	-24,91102172	0,000606338

De acordo com a tabela quatro, as comparações que mostraram efeitos significativos foram a segunda e a última. Assim, no nível Tipo C - Tipo A, o P-valor foi significativo. Esse nível tratou de comparar as condições com palavras terminadas em *ão*, que formam o plural em *aõs* ou *ões*, e as condições com palavras terminadas em *l*, que formam o plural com *is*, como já foi tratado anteriormente. Outro resultado significativo foi obtido no nível Tipo D - Tipo C. Esse nível tratou de comparar as condições com palavras terminadas em *m*, que formam o plural em *ns*, e as condições com palavras terminadas em *l*, como foi descrito no parágrafo anterior.

A análise discutida a seguir será sobre o tipo de relação entre o *prime* e o alvo no experimento. Conforme os resultados obtidos, os tempos médios de reação das condições foram as seguintes: *idêntico* - 848ms, *plural* - 833ms e *controle* - 920ms. A condição *controle* foi a mais lenta, tendo em vista que duas palavras sem nenhuma relação morfo-semântica apresentariam esse resultado. A condição *plural* foi a mais rápida. A explicação para esse efeito está no fato de que a estrutura morfológica de uma dada palavra no singular ativaría outra palavra processada no plural anteriormente. A condição *idêntica* obteve tempo de reação um pouco mais lento do que a condição *plural*. Desse modo, pode-se pensar que uma dada forma no singular, mostrada como *prime*, possui efeito de facilitação sobre a mesma forma também no singular, mostrada como alvo.

A seguir, mostra-se a tabela com o nível de significância do P-valor a respeito da condição *prime*-alvo.

Tabela 5. ANOVA - prime-alvo.

Tabela da ANOVA					
	GL.	Soma de Quadrados	Quadrado Médio	Estat..F	P.valor
Fator	2	5612086,787	2806043,393	16,10554844	1,08196E-07
Resíduos	3895	678619486,7	174228,3663		

Conforme a tabela cinco, o P-valor foi significativo. Isso mostrou que o tipo de relação entre o *prime* e o alvo é levado em consideração no processamento lexical dos itens analisados aqui. Pode-se observar, ainda, de acordo com a tabela cinco, que houve uma diferença na variável investigada.

Tabela 6. Médias - relação prime-alvo.

Agrupamento		
Fator	Médias	Grupos
Controle	920,4294	a
Idêntico	848,3069	b
Plural	833,5812	b

A tabela seis mostra os tempos médios de reação das condições *prime-alvo*, além das diferenças ou semelhanças existentes ao compará-los. A condição *idêntica* e a condição *plural* foram iguais entre si, representadas por *b*. Já a condição *controle* foi diferente das outras duas condições, representadas por *a*. A seguir, mostram-se as comparações entre as condições *prime-alvo*.

Tabela 7. Comparações entre relação prime-alvo.

Comparações Múltiplas				
Níveis	Centro	Limite Inferior	Limite Superior	P-valor
Idêntico-Controle	-72,12242638	-110,4201124	-33,82474032	3,08394E-05
Plural-Controle	-86,84813449	-125,3162855	-48,37998344	3,9886E-07
Plural-Idêntico	-14,72570811	-53,15004251	23,69862628	0,641255126

Na tabela sete, pode-se constatar que houve efeito significativo em *Idêntico-Controle* e em *Plural-Controle*. Ao comparar tais condições, o P-valor foi significativo, corroborando com o efeito observado na análise.

5. Considerações finais

Os resultados obtidos, no presente estudo, em PB, estão na mesma direção daqueles obtidos no estudo realizado no alemão. A condição *plural* obteve tempo mais curto do que aquelas da condição *idêntica* e da condição *controle*. Já as palavras da condição *plural* tiveram tempo mais curto do que aquelas da condição *controle*. Isso já era uma previsão do estudo em questão, já que formas sem nenhuma relação morfológica tenderiam a ter um custo maior no processamento. Desse modo, houve efeito de facilitação entre *prime* e *alvo* nas condições idênticas e plurais.

Ao analisar os dados obtidos, além das questões relacionadas ao acesso e à representação das formas irregulares flexionadas em número em PB, pode-se constatar que os afixos são ativados pela base. Isso permite dizer que as formas analisadas aqui vão na direção de um modelo decomposicional de acesso. Pode-se assumir que a representação de uma forma no plural, como *corações*, por exemplo, facilita o acesso da forma correspondente no singular. A base da referida forma ativa variações flexionais daquele item. Dessa forma, o acesso lexical se evidencia mais rapidamente a partir da estrutura armazenada no léxico mental.

Em relação aos tipos de plural, pode-se pensar em diferenças com maior ou menor complexidade dos morfemas. Os dados obtidos puderam verificar diferenças significativas ao comparar os diferentes tipos de plurais irregulares analisados aqui. Essas diferenças sugerem salientar que determinadas terminações de plurais foram processadas mais rapidamente, devido a menor complexidade estrutural. Isso poderia assemelhar essas terminações a formas regulares. Já as terminações de plurais que

foram processadas menos rapidamente apresentam complexidade morfológica maior que aquelas mencionadas anteriormente. Com isso, podem-se explicar as diferenças obtidas ao comparar a formação dos tipos de plural.

Referências

- Baayen, R. H. and Dukstra, T; Schreuder, R. (1997) "Singulars and plurals in dutch: Evidence for a parallel dual route model", Nova York, Journal of Memory and Language, v. 37, p. 94-117.
- Burani, C. and A. Caramazza. (1987) "Representation and Processing of Derived Words", Language and Cognitive Processes, v. 2, 217-227.
- Câmara Jr., J. M. (2013) "Estrutura da língua portuguesa", 45 ed., Petropólis - RJ, Editora Vozes.
- Caramazza, A., Laudani, A. and Romani, C. (1988) "Lexical access and inflection morphology", Cognition, v. 28, 2 ed., p. 297-332.
- Clahsen, H; Eisenbeiss, S; Hadler, M and Sonnenstuhl, I. (1999) "The mental representation of inflected words: An experimental study of adjectives and verbs in german", Language, p. 510-540.
- Davies, M. (2017) "Corpus do português", <http://www.corpusdoportugues.org>, November.
- Frauenfelder, U. H. and Schreuder, R. (1992) "Constraining psycholinguistic models of morphological processing and representation: The role of productivity". In: Yearbook of morphology (1991), Dordrecht : Springer Netherlands.
- Giraudo, H. and Grainger, J. (2000) "Effects of prime word frequency and cumulative root frequency in masked morphological priming", Language and Cognitive Processes, p. 421-444.
- Marinkovic, K. (2004) "Spatiotemporal dynamics of word processing in the human cortex", Neuroscientist, 10, p. 142-152.
- Pinker, S. and Prince, Alan. (1994) "Regular and irregular morphology and the psychological status of rules of grammar", Language, p. 230-251.
- Rocha, J. A. (2016) "Acesso e representação das formas nominais flexionadas em número em português brasileiro: Um estudo sobre o léxico mental" (dissertação de mestrado), UFPB - Universidade Federal da Paraíba, João Pessoa - PB.
- Schreuder, R. and Baayen, R. H. (1995) "Modeling morphological processing", In: Morphological aspects of language processing, Editado por L. B. Feldman, p. 131-154.
- Sonnenstuhl, I and Huth, A. (2002) "Processing and representation of german -n plurals: a dual mechanism approach", Brain and language, p. 276-290.
- Stockall, L. and Marantz, A. (2006) "A single route, full decomposition model of morphological complexity", MEG evidence, The Mental Lexicon, p. 85-123.
- Taft, M. and Forster, K. I. (1975) "Lexical storage and retrieval of prefixed words", In: Journal of Verbal Learning and Verbal, Behavior, p. 638-647.

Conjugador verbal do português brasileiro e análise morfonológica dos radicais

Gustavo Lopez Estivalet¹

¹Laboratório de Processamento Linguístico (LAPROL) - Universidade Federal da Paraíba (UFPB), João Pessoa-PB, Brasil.

gustavoestivalet@hotmail.com

Abstract. *The main objective of this work was to develop a verbal conjugator of Portuguese with the generalities of regular verbs and specificities of irregular verbs. The secondary objectives were: i. outline the function of the developed instrument; ii. describe the general characteristics of verbal conjugation in Portuguese, and iii. analyze the morphophonological irregularities of verbal stems in Portuguese. The verbal conjugator was programmed in R language and allows the conjugation of regular and irregular verbs, as well as pseudoverbs and neologisms of Portuguese in three forms: simple, reflexive, and pronominal. This instrument is available for free on <https://lexicodoportugues.shinyapps.io/Conjugator/>.*

Resumo. *O objetivo principal deste trabalho foi desenvolver um conjugador verbal do português comportando as propriedades gerais dos verbos regulares e específicas dos verbos irregulares. Os objetivos secundários foram: i. delinejar o funcionamento do instrumento desenvolvido; ii. descrever as características gerais da conjugação verbal do português e iii. analisar as irregularidades morfonológicas dos radicais verbais do português. O conjugador verbal foi desenvolvido em linguagem de programação R e permite a conjugação de verbos regulares e irregulares, assim como pseudoverbos e neologismos do português em três formas: simples, reflexiva e pronominal. Este instrumento está disponível gratuitamente em <https://lexicodoportugues.shinyapps.io/Conjugator/>.*

1. Introdução

Nos últimos anos, uma intensa discussão em torno da arquitetura do funcionamento do processamento morfológico tem se dado, especialmente em relação às classes flexionais. Por um lado, a arquitetura de “item e processamento” prevê a representação de unidades mínimas no léxico mental e a computação morfológica para formação de formas flexionadas; por outro lado, a arquitetura de “palavra e paradigma” prevê a representação de formas flexionadas como palavras inteiras a partir das relações com os paradigmas flexionais estabelecidos e produtivos (Blevins, 2006). Neste sentido, grande parte da discussão tem se dado em torno da representação e do processamento das formas irregulares apresentando modificações morfonológicas (e.g., impera[dor]/impera[triz], poder/p[O]de) e alomorfia (e.g., [ag]ir/[aj]o, [sab]er/[soub]e, ir/[v]ou); em suma, as formas irregulares são armazenadas no léxico mental como unidades inteiras ou possuem computações específicas para sua flexão?

As línguas latinas possuem morfologia rica e a conjugação verbal é uma categoria prototípica de produtividade flexional com grande número de formas em função de modo, tempo, pessoa e número (Villalva, 2007). Nesse sentido, a conjugação verbal de verbos específicos é frequentemente consultada para a verificação de formas irregulares assim como para apropriação de novas palavras, tais como estrangeirismos e neologismos. Para tanto, conjugadores verbais disponibilizados *online* na internet se tornaram uma ferramenta pertinente e frequentemente utilizada para verificação de formas verbais flexionadas, além disso, um algoritmo de conjugação e flexão verbal do português pode contribuir enormemente para o desenvolvimento e implementação do processamento de línguas naturais.

Sendo assim, o objetivo primário do presente trabalho foi desenvolver um conjugador verbal do português brasileiro (PB) para fins de pesquisa, consulta e aplicações computacionais. Os objetivos secundários foram: i. implementar um algoritmo de conjugação verbal do PB baseado na teoria linguística, ii. analisar as irregularidades morfológicas dos radicais verbais do PB e iii. oferecer uma ferramenta *online* e *offline* gratuita e aberta com aplicação em palavras e pseudopalavras do PB. Portanto, este trabalho justifica-se a partir da necessidade de descrição das características específicas da conjugação verbal do português e sua utilização no desenvolvimento de ferramentas de processamento de línguas naturais, assim como a exploração destes recursos em pesquisas em psicolinguística e linguística computacional. Ainda, esta ferramenta pode ser explorada como instrumento de consulta sobre a conjugação verbal do PB assim como recurso didático de ensino do PB (Estivalet & Meunier, 2015).

As perguntas que guiaram esta pesquisa foram: i. Quais as regras de sufixação para flexão verbal do português? ii. Quais são as subclasses de verbos irregulares do português? iii. Como as variáveis grafotáticas e de n-grama influenciam a flexão verbal de formas irregulares do português? A hipótese nula é que a primeira classe de conjugação de verbos com infinitivos terminados em [-ar] é completamente regular e produtiva, enquanto as segunda e terceira classes de conjugação de verbos com infinitivos terminados em [-er] e [-ir], respectivamente, é irregular e improdutiva (Veríssimo & Clahsen, 2009). A hipótese de trabalho da presente pesquisa é que todas as classes de conjugação do português apresentam subclasses com diferentes graus de irregularidade e produtividade (Kilani-Schoch & Dressler, 2005) baseado nas estruturas grafotáticas e fonotáticas, assim como nas frequências de n-gramas do português (Albright, 2002).

Para realização da presente investigação, primeiramente foram pesquisados os mecanismos regulares e irregulares da flexão verbal do português e em seguida foi implementado um algoritmo computacional para flexão verbal do PB baseado na teoria linguística (Embick & Halle, 2005). Deu-se especial atenção para análise morfológica das irregularidades dos radicais com objetivo de definirem-se subclasses verbais com diferentes graus de irregularidades e produtividade, que por sua vez, possam também contemplar a irregularidades de neologismos, estrangeirismo e pseudopalavras do PB.

2. Conjugação verbal do português brasileiro

Em geral, os dicionários dos PB não apresentam as formas verbais flexionadas, mas somente as entradas lexicais nas formas infinitivas, já as gramáticas tradicionalmente apresentam as conjugações dos tempos verbais de verbos prototípicos. Atualmente, faz-se o uso de conjugadores verbais na internet quando se deseja consultar a forma flexionada específica de um verbo¹. Ainda, algumas ferramentas didáticas já foram desenvolvidas sobre a conjugação verbal, contudo continuam indisponíveis para o grande público, cobram taxas para sua utilização e/ou não apresentam a teoria utilizada para desenvolvimento (Vasilévski & Araújo, 2011). Sendo assim, a ferramenta e o algoritmo apresentados no presente trabalho têm como objetivo contornar estas limitações da descrição, implementação e utilização da conjugação verbal do PB.

2.1. Tempos verbais

A conjugação verbal do PB apresenta três modos, 12 tempos verbais, três pessoas do discurso e dois números de pessoas. O I. modo indicativo contém seis tempos verbais: 1. presente, 2. pretérito perfeito, 3. pretérito imperfeito, 4. pretérito-mais-que-perfeito, 5. futuro do presente e 6. futuro do pretérito; o II. modo subjuntivo contém três tempos verbais: 7. presente, 8. pretérito imperfeito e 9. futuro; o III. modo imperativo contém os tempos verbais: 10. afirmativo e 11. negativo. Além disso, a conjugação verbal do PB apresenta três formas nominais: i. infinitivo, ii. particípio passado e iii. gerúndio; o infinitivo contém o a. infinitivo impessoal e o b. infinitivo pessoal. As pessoas do discurso são 1^a, 2^a e 3^a nos números singular e plural, tipicamente representados pelos pronomes pessoais do caso reto: eu, tu, ele|ela, nós, vós, eles|elas, respectivamente. Além disso, destaca-se no PB a substituição do pronome de 2^a pessoa do singular “tu” pelo pronome “você” e do pronome de 1^a pessoa do plural “nós” pela expressão “a gente” com conjugação da 3^a pessoa do singular, assim como a substituição do pronome de 2^a pessoa do plural “vós” pelo pronome “vocês” com conjugação da 3^a pessoa do plural (Bassani & Luguinho, 2011).

O futuro do pretérito do indicativo também é conhecido como modo condicional. Além dos tempos simples, o português possui também as formas compostas construídas a partir dos verbos “ter/haver” conjugado nos diferentes tempos verbais seguido do particípio passado do verbo principal, que por sua vez podem ser regulares (e.g., am[ado], com[ido], dorm[ido]), irregulares (e.g., d[ito], ab[erto]) ou abundantes (e.g., entreg[ado]/entreg[ue], salv[ado]/sal[vo]).

2.2. Conjugações

O português apresenta tipicamente três classes definidas em relação à vogal temática, verbos com a forma infinitiva terminada em: [a]r, [e|o]r ou [i]r. A primeira classe em [-ar] é a classe mais regular e produtiva, possuindo maior número de verbos (mais de 4000 verbos). A segunda classe [-er|-or] apresenta irregularidade, número reduzido de verbos (cerca de 240 verbos) e verbos auxiliares e modais de grande utilização (e.g., ser, ter, haver, dever, fazer, querer, poder). A terceira classe em [-ir] apresenta verbos mais irregulares e um número mais reduzido (cerca de 160) (Villalva, 2007).

¹ Por exemplo: <https://www.conjugacao.com.br/>, <http://www.conjugador.com.br/>.

As formas verbais se apresentam em três conjugações: simples, reflexiva e pronominal. A conjugação simples apresenta apenas as formas flexionadas (e.g., eles lavaram); a conjugação reflexiva apresenta as formas flexionadas ligadas por hífen a clíticos expressados pelos pronomes reflexivos representados pelos pronomes pessoais do caso oblíquo átonos: me, te, se, nos, vos, se (e.g., eles lavaram-se); a conjugação pronominal apresenta as formas flexionadas, que por sua vez pode apresentar mudanças específicas, ligadas por hífen a clíticos expressados pelos pronomes objeto direto/indireto representados pelos pronomes pessoais do caso oblíquo átonos: me, te, o/lo/no/a/la/na/lhe, nos, vos, os/los/nos/as/las/nas/lhes (e.g., eles lavaram-no, lavá-las).

2.3. Variáveis morfonológicas

Para o desenvolvimento do presente trabalho, observaram-se três variáveis em relação à forma dos radicais verbais: i. regularidade (regular, grau de irregularidade, supressão) (Embick & Halle, 2005), ii. morfonologia (marca ortográfica da vogal temática, mudança de consoante ortográfica) e iii. distância ortográfica de Levenshtein (DOL) (número de inserções, exclusões e substituição entre formas) (Levenshtein, 1966).

Portanto, a partir dessas variáveis, propôs-se a classificação das formas irregulares em subclasses a partir do grau de irregularidade e da consistência grafotática entre os diferentes lexemas de um mesmo lema. Primeiramente os verbos foram classificados como i. regular, ii. irregular ou iii. supletivo; em seguida os verbos irregulares foram analisados a partir das formas alomórficas; enfim, avaliaram-se as irregularidades grafotáticas específicas das formas flexionadas (Albright, 2002).

3. Metodologia

3.1. Análise

Com o objetivo de se implementar um algoritmo computacional baseado diretamente na teoria linguística de “item e processamento”, mais especificamente de acordo ao modelo da morfologia distribuída, a flexão verbal do PB foi analisada a partir de dois tipos de morfemas: morfemas flexionais contendo traços morfossintáticos (sufixos) e morfemas lexicais de conteúdo semântico (radicais). Sendo assim, realizou-se uma análise computacional linguística i. das regras de aplicação dos sufixos flexionais (Bassani & Luguinho, 2011) e ii. da morfonologia e da alomorfia dos radicais lexicais verbais das subclasses dos verbos irregulares (Embick & Halle, 2005).

Em relação às operações de sufixação para flexão verbal, partiu-se da 1^a classe de verbos terminados em [-ar] produtiva e procedeu-se com a formulação de regras de ajuste para as 2^a [-er] e 3^a [-ir] classes (e.g., IND.PRET.IMP.3S: [va]_{PRET.IMP} → [ia]_{PRET.IMP}/_[e|i]_{TH}, ama[va] → com[ia]|dorm[ia]). Em relação aos radicais verbais, realizou-se primeiramente uma análise morfonológica ortográfica das formas flexionadas (lexemas) que possuíam divergências em relação às formas infinitivas (lema) no caractere imediatamente anterior a vogal temática (e.g., [caç]ar~[cac]ei, [ag]ir~[aj]o, [passe]ar~[passei]o); em seguida, analisaram-se os radicais alomórficos através da comparação da consistência grafotática dos demais caracteres finais dos radicais verbais (e.g., [sab]er~[se]i~[soub]e~[saib]a, [v]ir~[venh]o~[vinh]a). Todas as análises foram realizadas a partir do dicionário UNITEM/DELAF (Muniz, 2004).

3.2. Algoritmo

O algoritmo do conjugador verbal do português foi implementado no programa R (R Core Team, 2014), sendo exploradas principalmente as funções de base `paste` e `gsub`. A primeira permite a concatenação de dois ou mais conjuntos de caracteres e foi utilizada para formação de formas verbais flexionadas através da operação de junção entre o radical (formado pela raiz e vogal temática) e os sufixos flexionais. A segunda permite a substituição de caracteres específicos e foi utilizada para as operações de ajuste ortográfico (morfofonológico). As expressões regulares (REGEX) foram especialmente exploradas para manipulação de caracteres e para o desenvolvimento de funções específicas (Fitzgerald, 2012), explorando-se especialmente os conjuntos de caracteres, limitadores de palavra e operadores de repetição (e.g., `^.*[aei]r$` = infinitivo, formas finalizadas pelas vogais temáticas seguidas de morfema infinitivo).

Para análise dos radicais alomórficos, explorou-se a DOL entre as formas dos radicais dos lexemas e lemas. A DOL é definida a partir do cálculo do número de operações de inserções, exclusões e substituições necessárias para comparação de duas formas (e.g., *sab/saib*: 1 inserção, *sab/soub*: 1 substituição + 1 inserção = 2, *sab/se*: 1 exclusão + 1 substituição = 2) (Levenshtein, 1966).

3.3. Processamento

Os seguintes procedimentos para formação de formas verbais flexionadas foram implementados: i. para os verbos regulares, procede-se de forma linear a junção do radical com os sufixos flexionais, seguida das regras de ajuste; ii. para os verbos irregulares, uma terceira etapa realiza a análise da irregularidade e produtividade do radical e ajusta-os de acordo aos radicais alomórficos correspondentes (Albright, 2002). As operações alomórficas das subclasse dos verbos irregulares são realizadas a partir da análise da estrutura grafotática dos radicais.

Para as conjugações reflexiva e pronominal, são acrescentados os pronomes clíticos na posição de ênclise correspondentes a seleção do usuário: tipo de conjugação (simples, reflexiva, pronominal); conjugação pronominal (mas/sing, mas/plur, fem/sing, fem/plur). Enfim, são realizadas operações de ajustes (e.g., *amar/amá-lo*). Além disso, o conjugado possui a opção de apresentação dos rótulos das formas flexionadas (e.g., *lavas_C1.IND.PRE.2S*, classe: 1^a conjugação [-ar], modo: indicativo, tempo verbal: presente, pessoa: 2^a, número: singular).

4. Resultados e discussão

4.1. Conjugador verbal do português brasileiro

Como resultado, o algoritmo aberto e livre implementado no programa R (R Core Team, 2014) e o aplicativo disponibilizado gratuitamente na internet permitem a conjugação completa de todos os verbos do português, além disso, esse instrumento permite a conjugação verbal de neologismos e estrangeirismo, assim como de pseudoverbos que obedecem as regras grafotáticas do PB (Keuleers & Brysbaert, 2010). Observa-se que estas características são exatamente as principais limitações de outros conjugadores verbais do português (Vasilévski & Araújo, 2011).

Na Figura 1 abaixo, são apresentados exemplos do funcionamento do aplicativo online: a esquerda a conjugação simples em [-ar] com exibição dos rótulos, no centro a conjugação reflexiva em [-er] e a direita a conjugação pronominal feminina/plural em [-ir]. Na parte superior, a caixa de texto permite a entrada de qualquer verbo na forma infinitiva (i.e., REGEX: `^.*[aeoi]r$`). Abaixo, seleciona-se o tipo de conjugação, no caso de conjugação pronominal, pode-se selecionar o gênero e número da conjugação. Enfim, a opção TAG apresenta os rótulos de cada forma flexionada. Portanto, as formas geradas pelo algoritmo podem facilmente ser utilizadas por outros programas de processamento de línguas naturais explorando os rótulos contendo os traços morfossintáticos das formas flexionadas (Estivalet & Margotti, 2014).

The figure consists of three side-by-side screenshots of a web application for Portuguese verb conjugation. Each screenshot shows a search bar at the top with a verb root ('lavar', 'comer', or 'abrir'). Below the search bar are sections for 'Conjugation type' (radio buttons for 'normal', 'refl', and 'pron'), 'Gender/Number' (radio buttons for 'ms', 'mp', 'fs', and 'fp'), and a 'Tag' checkbox. The main area displays a table of conjugated forms with columns for 'pres', 'perpast', 'imppast', 'pluspast', and 'presut'. The first screenshot shows results for 'lavar' (simple past), the second for 'comer' (reflexive past), and the third for 'abrir' (pronominal past). Each row includes the base form followed by its conjugated forms and their respective morphological tags.

Figura 1. Exemplos de utilização do aplicativo *online*.

4.2. Análise do corpus verbal

A análise do dicionário UNITEX/DELAF (Muniz, 2004) apresentou o total de lemas e lexemas e o total de formas simples de lemas e lexemas conforme a Tabela 1 abaixo. No primeiro par de colunas, apresenta-se o número total de lemas (infinitivos) e lexemas (formas flexionadas) encontrados no dicionário. No segundo par de colunas, apresenta-se o número de lemas e lexemas somente das conjugações simples, isto é, sem as formas reflexivas e pronominais. O terceiro par de colunas apresenta o número de lemas e lexemas das formas irregulares, ou seja, formas flexionadas que apresentaram $DOL > 0$. O quarto par de colunas apresenta o número de formas suppletivas, considerada com substituição da primeira letra e/ou $DOL > 2$. Destaca-se que a DOL entre cada par lexema/lema foi calculada a partir das formas ortográficas após a decomposição do morfema de infinitivo [-r] e do morfema de vogal temática [-a|e|o|i-] do lema em relação ao lexema com mesmo número de letras (e.g., fal[es]/fal[ar] = 0, cac[es]/caç[ar] = 1, durm[a]/dorm[ir] = 1, soub[e]/sab[er] = 2).

Sendo assim, pode-se calcular o coeficiente de família morfológica flexional, definido pela relação entre o número de lexemas em função do número de lemas (i.e., CFM = lexemas/lemas) (Baayen, 2008). Observa-se um CFM bastante alto para todas formas encontradas no dicionário ($10.430.109/14.278 = 730$), esse coeficiente alto se deve ao grande número de formas flexionadas (lexemas) por forma infinitiva (lema), pois o corpus completo apresenta todas formas flexionadas de todos tipos de conjugação (i.e., simples, reflexiva e pronominais). Diferentemente, o CFM das formas simples está mais próximo do valor esperado ($925.646/14.278 = 64$), pois: 6 pessoas/número X 12 tempos verbais = 72, considerando-se as formas homógrafas (Villalva, 2007). Em seguida, o CFM das formas irregulares ($25.299/2.109 = 12$) indica que apenas 12 formas

flexionadas de cada infinitivo são irregulares, ou seja, em torno de 18,75% das formas totais da conjugação completa. Enfim, nos verbos supletivos, o CFM (301/10 = 30) indica que cerca de 47% das formas são supletivas, ou seja, possui modificação da primeira letra ou modificações em mais de 2 letras do radical (Albright, 2002).

Tabela 1. Número de lemas (Lm) e lexemas (Lx) do dicionário UNITEX/DELAf.

Classe	Lemas	Lexemas	Lm Sim.	Lx Sim.	Lm Irr.	Lx Irr.	Lm Supl.	Lx Supl.
1^a [-ar]	12.718	9.394.572	12.718	824.594	1.352	16.175	1	43
2^a [-er]²	839	605.587	839	53.958	471	6.316	6	193
3^a [-ir]	721	429.950	721	47.094	281	2.808	3	65
Total	14.278	10.430.109	14.278	925.646	2.104	25.299	10	301

Destaca-se que de todas as formas analisadas, 14,73% dos lemas apresentam 2,73% de lexemas irregulares; dentre os irregulares, apenas 0,47% dos lemas apresentam 1,19% de lexemas supletivos. Portanto, mais de 85% de todas as formas verbais flexionadas do PB são produzidas a partir de lemas regulares através de regras simples de junção entre o radical regular e os sufixos flexionais (Rocha, 1999). Todavia, destaca-se que muitos verbos frequentes do PB estão entre os menos de 15% de verbos irregulares, sendo assim, cabe realizar-se uma análise das irregularidades dos radicais a partir de uma perspectiva gradual, para enfim estabelecerem-se subclasses.

4.3. Subclasses verbais

A análise das subclasses irregulares que apresentam DOL = 1 com modificações na última vogal ou consoante de ligação final do radical apontou 18 subclasses no PB. Observa-se o que a 1^a conjugação regular em [-ar] é a classe que apresenta menos subclasses (3), mas produtivas com verbos terminados em “-car” com modificações entre “c/qu” e terminados em “-çar” com modificações entre “ç/c” (e.g., educar/eduque, caçar/cace). Esse mesmo padrão produtivo pode ser encontrado de forma invertida na 2^a conjugação em [-er] “-cer” “c/ç” (e.g., escurecer/escureço). Outras subclasses semiprodutivas são as terminações em “-ger” com modificações entre “g/j” da 2^a conjugação e inversamente em “-gir” com modificações entre “g/j” da 3^a conjugação em [-ir] (e.g., proteger/protejo, emergir/emergo), conforme apresentada na Tabela 2. Portanto, essas modificações na última letra do radical permitem estabelecer as seguintes regras de alomorfia ortográfica simples para os casos apontados: c/_ar→qu/_e|i, ç/_ar→c/_e|i, c/_er|ir→ç/_a|o, g/_er|ir→j/_a|o, gu/_er|ir→g/_a|o (Estivalet & Margotti, 2014).

Em seguida, as subclasses que apresentaram produtividade (mais de 10 lemas) com DOL = 1 e modificações na penúltima letra do radical (alomorfia de vogal) são: -azer (45), -elir (11), -erir (33) e -izar (14) (e.g., fazer/fizer, impelir/impilo, ferir/firo, enraizar/enraízo). Somente a classe “-izar” pertence a 1^a conjugação e, por sua vez, apresenta simplesmente uma modificação de acento “i/í” na última vogal do radical.

² Desses, as seguintes formas eram em [-or]: Lemas = 39, Lexemas = 31.746, Lexemas Simples = 2.574, Lemas Supletivos = 1, Lexemas Supletivos = 63.

Tabela 2. Frequências das subclasses verbais que apresentam irregularidades na última letra do lema (infinitivo) em seus lexemas (formas flexionadas).

	[-ar]	[-er]	[-ir]	Total
Vogal	-iar (10)	-oer (8), -uer (8)	-uir (28)	54
Consoante	-car (954), -çar (366)	-ber (5), -cer (381), -der (4), -ger (21), -rer (6), -ver (2), -zer (79)	-cir (2), -dir (13), -gir (87), -rir (3), -vir (4)	1927
Total	1330	514	137	1981

Finalmente, a análise das subclasses restante com DOL > 1 apresentou 57 subclasses (1^a conjugação [-ar] = 5, 2^a conjugação [-er] = 30, 3^a conjugação [-ir] = 22), com padrões produtivos (mais que 5 lemas) em: -aber (8), -aver (10), -azer (7) e -guir (14). Portanto, a partir do funcionamento dessas subclasses verbais, as regras de alomorfia do radical podem ser estabelecidas para conjugação de verbos irregulares (Embick & Halle, 2005), assim como para adequação de conjugação de neologismos, estrangeirismos e pseudoverbos do PB (Keuleers & Brysbaert, 2010).

Enfim, os verbos irregulares podem ser classificados em i. supletivos, ii. ajuste da última vogal ou consoante de ligação do radical, iii. alomorfia da última vogal do radical e/ou iv. alomorfia do radical. A análise desenvolvida e algoritmo implementado no presente instrumento apresenta relações diretas com a teoria linguística e estruturas hierárquicas das classes e subclasses verbais (Kilani-Schoch & Dressler, 2005). As subclasses verbais produtivas do PB analisadas aqui são apresentadas na Figura 2.

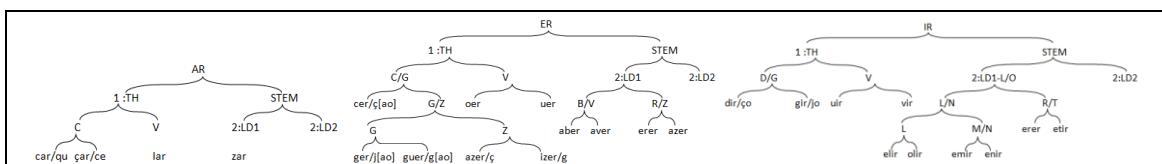


Figura 2. Subclasses verbais produtivas do PB.

5. Considerações finais

Como resultado, o instrumento desenvolvido permite a conjugação completa de todos os verbos do português brasileiro. Além disso, ele permite a conjugação de pseudoverbos que obedecem às estruturas e regras grafotáticas do português. O instrumento possui como opções as conjugações: simples, reflexiva e pronominal (masc/fem, sing/plur); assim como a apresentação dos rótulos das formas flexionadas. O conjugador verbal do PB está disponível gratuitamente: <https://lexicodoportugues.shinyapps.io/Conjugator/>; o algoritmo em: http://www.lexicodoportugues.com/downloads/lexporbr_conjugator.R/.

A análise das irregularidades morfológicas e alomórficas dos radicais verbais do PB permitiu uma melhor compreensão dos fenômenos relacionados a hierarquia e processos alomórficos e ortográficos dos radicais verbais do PB. Espera-se que a análise linguística do sistema verbal, assim como o instrumento aqui desenvolvido, sejam úteis para trabalhos e investigações futuras em torno da descrição e funcionamento do sistema verbal do PB.

Referências

- Albright, A. (2002). Islands of Reliability for Regular Morphology: Evidence from Italian. *Language*, 78(4), 684–709.
- Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.
- Bassani, I. de S., & Lunguinho, M. V. (2011). Revisitando a flexão verbal do português à luz da Morfologia Distribuída: um estudo do presente, pretérito imperfeito e pretérito perfeito do indicativo. *ReVEL*, 199–227.
- Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 42(03), 531.
- Embick, D., & Halle, M. (2005). On the status of stems in morphological theory. In Geerts, T. & Jacobs, H. (Eds.), *Proceedings of Going Romance 2003*. Amsterdam: John Benjamins, 59–88.
- Estivalet, G. L., & Margotti, F. W. (2014). Diálogos entre a flexão verbal do Português e do Francês (Dialogues entre la flexion verbale du Portugais et du Français). *Estudos da Lingua(gem)*, 12(2), 31–49.
- Estivalet, G. L., & Meunier, F. (2015). The Brazilian Portuguese Lexicon: An Instrument for Psycholinguistic Research. *PLOS ONE*, 10(12), e0144016.
- Fitzgerald, M. (2012). *Introdução às expressões regulares*. São Paulo, Brasil: Novatec Editora.
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627–633.
- Kilani-Schoch, M., & Dressler, W. U. (2005). *Morphologie naturelle et flexion du verbe français* (Vol. 488). Tübingen: Gunter Narr Verlag Tübingen.
- Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and reversals. *Soviet Physics*, 10(8), 707–710.
- Muniz, M. C. M. (2004). *A construção de recursos linguístico-computacionais para o português do Brasil: o projeto de Unitex-PB*. Universidade de São Paulo.
- R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna/Austria.
- Rocha, L. C. de A. (1999). *Estruturas Morfológicas do Português*. Belo Horizonte, MG: Editora UFMG.
- Vasilévski, V., & Araújo, M. J. (2011). Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português. *LinguaMATICA*, 3(2), 107–118.
- Veríssimo, J., & Clahsen, H. (2009). Morphological priming by itself: A study of Portuguese conjugations. *Cognition*, 112(1), 187–194.
- Villalva, A. (2007). *Morfologia do Português*. Lisboa, PT: Universidade Aberta.

Córpus 4P: um córpus anotado de opiniões em português sobre produtos eletrônicos para fins de summarização contrastiva de opinião

Raphael Rocha da Silva¹ e Thiago Alexandre Salgueiro Pardo¹

¹Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos – SP – Brasil

Resumo. Este artigo descreve a construção do córpus 4P, um córpus de opiniões em português brasileiro sobre telefones celulares e câmeras digitais. O córpus conta com a anotação manual de aspectos e polaridades das opiniões. As 642 sentenças do córpus foram coletadas de 542 comentários opinativos publicados por compradores no site Buscapé e se referem a quatro produtos diferentes. Esse córpus deve subsidiar pesquisas na área de Análise de Sentimentos, mais especificamente, em Sumarização Contrastiva de Opinião.

Abstract. This paper describes the construction of the 4P corpus, which is a corpus of opinions in Brazilian Portuguese about cell phones and digital cameras, with manually annotated aspects and polarities. The 642 sentences that compose the corpus were collected from 542 opinionated comments posted by users on Buscapé website and refer to four different products. This corpus is intended to subsidize research in the area of Sentiment Analysis, specially in Contrastive Opinion Summarization.

1. Introdução

A summarização contrastiva de opinião é uma tarefa de Análise de Sentimentos que tem como objetivo a geração de resumos que comparam entidades com base em textos opinativos [Liu 2012]. Por exemplo, pode-se desejar comparar dois produtos eletrônicos (como dois modelos de celular) por meio dos comentários que as pessoas publicam sobre eles na Web. Veja, por exemplo, as iniciativas na literatura especializada [Liu et al. 2005, Lerman e McDonald 2009, Kim e Zhai 2009, Jin et al. 2016].

A criação do córpus relatado neste texto – o **córpus 4P** – se deu para possibilitar a investigação de métodos de summarização contrastiva para o português. Para isso, é necessário um conjunto de textos de tamanho suficiente, com anotação confiável e que ofereça casos de teste diversificados para se avaliar os métodos. Cada um dos métodos publicados na literatura usa seu próprio conjunto de dados para testes, e esses córpus ou não foram disponibilizados publicamente ou não têm características adequadas para confrontar os métodos. O córpus apresentado aqui preenche essa lacuna.

Na linha do que fizeram outros trabalhos [Liu et al. 2005, Lerman e McDonald 2009, Kim e Zhai 2009, Jin et al. 2016], optou-se por coletar textos opinativos sobre 4 produtos eletrônicos (daí a origem do nome do córpus). Essa escolha deve-se principalmente à facilidade de coletar esse tipo de texto (dada sua abundância em algumas páginas da web) e também por se acreditar que a identificação

de aspectos opinativos nesse domínio é uma tarefa mais bem definida do que em outros casos (por exemplo, avaliações de serviços, resenhas de livros e discussões políticas), como sugerido em [Vargas e Pardo 2018].

Aspectos costumam ser o foco de muitas tarefas de Análise de Sentimentos [Liu 2012]. Por exemplo, se uma pessoa opina sobre um telefone móvel, ela pode falar sobre a tela, o peso, o tamanho, etc. Essas características e partes da entidade avaliada são o que se convencionou chamar de **aspectos** (ou aspectos opinativos, para diferenciar de outros possíveis usos). Uma **entidade** é o objeto alvo da opinião; pode ser um produto, serviço, organização, indivíduo ou evento, entre outros [Liu 2012].

No córpus anotado neste trabalho, cada sentença foi manualmente rotulada quanto aos **aspectos** avaliados e quanto à **polaridade** da opinião, que indica se a opinião é positiva ou negativa. Além dessas anotações, foram identificadas sentenças que não contêm opiniões (por exemplo, ‘comprei esse celular semana passada’) e sentenças que não pertencem ao escopo de avaliação do produto (por exemplo, ‘demorou muito para chegar’).

O córpus foi composto a partir de 542 comentários extraídos do site Buscapé, que é um site brasileiro que oferece serviço gratuito de busca de produtos e comparação de preços em lojas virtuais, permitindo que seus leitores publiquem avaliações sobre os produtos. Os comentários coletados contêm 642 sentenças, sendo que foram manualmente identificadas nelas 1.658 opiniões¹. Formalmente, considera-se, neste trabalho, que uma opinião é um par formado por um aspecto e sua polaridade. Foram identificadas no córpus 1.416 opiniões positivas e 228 negativas², além de 244 outras passagens que não contêm opinião, que foram separadas em várias categorias para melhor caracterizar o córpus.

Visando subsidiar o teste de métodos de sumarização em contextos variados de uso, diferentes arranjos do córpus foram produzidos. Além da configuração inicial, foram construídos subconjuntos do córpus para permitir avaliar os métodos em situações com variedade de dados. Foram criados 12 arranjos dos dados a partir da seleção de sentenças sobre as entidades avaliadas. A seleção foi feita de modo que os arranjos tivessem características diversificadas entre si, como quanto à quantidade de aspectos citados, proporção entre opiniões positivas e negativas, etc.

A seguir, uma breve revisão literária é apresentada. O córpus construído neste trabalho é apresentado na Secção 3. Algumas considerações finais são feitas na Secção 4.

2. Trabalhos relacionados

A tarefa da sumarização contrastiva de opinião aparece na literatura geralmente voltada para a comparação entre opiniões de compradores sobre produtos [Jin et al. 2016, Lerman e McDonald 2009, Kim e Zhai 2009, Liu et al. 2005], embora também se encontrem trabalhos que a executam para outros tipos de textos, como assuntos controversos [Guo et al. 2015]. Alguns trabalhos estudam a tarefa para textos não opinativos e usam outros tipos de textos em seus conjuntos de dados, como artigos jornalísticos sobre assuntos polêmicos [Park et al. 2011, Wang et al. 2012].

Para a tarefa de comparar produtos a partir de textos opinativos, [Liu et al. 2005]

¹Algumas sentenças contêm mais de uma opinião, como ‘A câmera é boa mas o aparelho trava muito’, onde se identificam uma opinião positiva sobre a câmera e uma opinião negativa sobre o aparelho.

²As outras 14 não são claramente positivas ou negativas.

fizeram um conjunto de dados manualmente etiquetado com comentários em inglês sobre 15 produtos eletrônicos. As opiniões foram coletadas do site Epinions. Os autores não dão outras informações sobre o conjunto de dados. [Lerman e McDonald 2009] coletaram comentários em inglês sobre 56 produtos eletrônicos de 15 tipos (câmeras, computadores, sistemas de GPS, tocadores de MP3, etc.) de várias fontes (CNet, Epinions, PriceGrabber, etc). Cada produto tem um mínimo de 4 comentários e a média de comentários por produto é 70. O córpus não foi disponibilizado.

Para estudar a geração de sumários contrastando opiniões sobre um mesmo produto, [Kim e Zhai 2009] usaram um conjunto de dados com comentários etiquetados sobre 12 produtos coletados do site da Amazon. Para testar a generalidade do método, também foi usado um conjunto extra formado por 50 comentários favoráveis e 50 contrários ao uso de aspartame.

O conjunto criado por [Condori e Pardo 2017] com textos em português mostrou-se adequado para a summarização de opinião estudada em seu trabalho, que tem como objetivo selecionar as opiniões mais relevantes dos textos sobre uma entidade para compor um sumário. Entretanto, para a summarização contrastiva de opinião, buscam-se conjuntos maiores para que haja maior possibilidade de confrontar elementos concorrentes.

A seguir, relata-se a criação do córpus para fins de summarização contrastiva.

3. O córpus 4P

Após escolhidas as entidades e coletadas as opiniões sobre elas, a primeira etapa da criação do córpus foi definir como ele deveria ser anotado. Foram feitas leituras dos textos coletados e anotações prévias para entender melhor o problema e definir as regras de anotação. Então, o conjunto passou por uma etapa de anotação automática para que os anotadores precisassem apenas conferir a anotação e não ter que inserir todas as informações eles mesmos. Após a revisão manual, o córpus passou por processos de extensão, limpeza e simplificação, para posterior disponibilização. Esses passos são descritos a seguir.

3.1. Regras para anotação

A próxima seção relata o processo prático da anotação, e esta descreve as regras usadas para identificar opiniões. As regras foram definidas empiricamente por meio da inspeção manual dos textos coletados, tendo como base ideias publicadas em [Liu 2012] e [Vargas e Pardo 2018].

3.1.1. Identificação de aspectos

Um aspecto é o assunto principal de que trata uma opinião. Os aspectos foram identificados seguindo uma listagem de 16 aspectos para celulares³ e 18 aspectos para câmeras⁴.

³Acessório, armazenamento, bateria, câmera, desempenho, design, peso, preço, produto, resistência, sistema, som, tamanho, tela, usabilidade e outro.

⁴Acessório, armazenamento, áudio, bateria, design, foco, foto, funcionalidade, imagem, peso, preço, produto, resistência, tamanho, tela, usabilidade, vídeo e zoom.

Além desses, deveria ser atribuído um aspecto especial caso a opinião se referisse ao produto de maneira geral (e não a um aspecto específico), como na sentença ‘*esse celular é supimpa*’; opiniões assim são ditas **genéricas**.

3.1.2. Identificação de polaridades

Cada aspecto identificado deveria ser associado a uma polaridade que refletisse o sentimento expresso em relação a ele. A identificação de polaridade foi feita de acordo com as possibilidades listadas abaixo (os exemplos são reais e extraídos do próprio córpus):

- **positivo:** algo bom, desejável – ‘*Celular muito rápido*’.
- **positivo fraco:** algo condicionalmente bom ou parcialmente bom – ‘*É um valor caro a se pagar, mas vale a pena*’.
- **positivo forte:** algo excepcionalmente bom – ‘*Perfeito, sem qualquer reclamação*’.
- **negativo:** algo ruim, indesejável – ‘*O preço é alto demais*’.
- **negativo fraco:** algo condicionalmente ruim ou parcialmente ruim – ‘*Creio que para usuários mais exigentes não compensaria*’.
- **negativo forte:** algo excepcionalmente ruim – ‘*Foi o pior aparelho que já comprei*’.
- **mediano:** algo no meio da escala entre desejável e indesejável – ‘*Aparelho razoável*’.
- **relativo:** trecho subjetivo onde não há um conceito claro de valor entre desejável ou indesejável – ‘*Design discreto*’.
- **dual:** trecho que indica algo simultaneamente bom e ruim sobre um mesmo aspecto – ‘*Aceita SD mas não expande a memória interna*’.
- **irresoluto:** indica indecisão ou falta de opinião – ‘*Ainda estou avaliando*’.
- **conselho:** informação que ajuda a usar melhor o produto – ‘*Recomendo que seja comprada uma capa de proteção*’.
- **experiência:** relata a experiência de uso do produto de forma que não remete a uma opinião – ‘*Uso para falar e web*’.

3.1.3. Trechos não avaliativos

Algumas sentenças coletadas não são úteis para avaliar o produto em questão. Essas sentenças foram separadas em três classes:

- **fora de escopo:** trechos que falam não sobre o produto, mas sobre alguma entidade relacionada a ele ou à experiência de compra, como fabricante, vendedor, transportadora, etc. – ‘*Chegou bem rápido*’.
- **contextualização:** trechos que contêm informação adicional que pode agregar valor ao comentário da pessoa, mas não ajuda a avaliar o produto se lido isoladamente – ‘*Foi presente*’.
- **irrelevante:** trechos que não se relacionam ao produto em questão e sequer agregam valor a comentários sobre o produto – ‘*Não sei porque estou respondendo esta pesquisa*’.

Também foram identificados trechos **duplicados** (quando a mesma pessoa publica duas vezes o mesmo comentário ou repete algo no mesmo comentário) e trechos que têm problemas de texto, como trechos **ininteligíveis** ('*xvcxcvc*') e **quebrados** ('*pouca, recém comprada*').

3.1.4. Segmentação de texto

Os colaboradores foram instruídos a identificar todas as opiniões contidas no texto. Sentenças que contivessem mais de uma opinião deveriam ser divididas em trechos (oracões, usualmente) de forma que cada trecho contivesse apenas uma opinião. As informações sobre a divisão de sentenças foram registradas para que também seja possível usar o córpus com as sentenças completas.

Os colaboradores foram orientados a reescrever trechos para que cada trecho fosse sentido se lido isoladamente. Por exemplo, a sentença '*o produto é bom mas caro pra dedéu*' poderia ser quebrado nos trechos '*o produto é bom*' e '*o produto é caro pra dedéu*'. Isso seria útil para uma eventual tarefa onde convém considerar trechos contendo uma única opinião. Além disso, essa divisão traz mais segurança à anotação, pois permite identificar exatamente onde cada opinião foi encontrada.

Se achasse necessário, os anotadores também poderiam unir duas sentenças em casos que essa não formasse uma opinião que não se formaria com as sentenças isoladas, como no trecho '*O que eu mais busco num celular é a qualidade e possibilidade de edição nas configurações da câmera. E nesse quesito o S7 não deixa a desejar.*'

3.2. Trabalho de anotação

Após coletados os dados, os textos foram automaticamente divididos em sentenças e a classificação de opiniões começou com uma etapa de identificação automática com um método⁵ simples que identifica aspectos por meio de palavras-chave (como estudaram [Vargas e Pardo 2018]) e polaridades por meio de meta-informações sobre a avaliação⁶. Depois da fase automática, duas pessoas trabalharam revisando a anotação. Toda a anotação foi feita em ferramentas de edição de texto.

Os anotadores receberam arquivos de texto puro, um para cada produto, estruturados como no Quadro 1. Dentro de cada arquivo, os comentários aparecem ordenados pela data de publicação. Cada sentença se inicia com um identificador do tipo (006.015) que indica o comentário de onde a sentença foi extraída e a posição da sentença no arquivo. Depois, existem dois pares de colchetes: o primeiro é preenchido com a polaridade e o segundo contém os aspectos identificados na etapa automática. Depois, há um separador (formado por dois dois-pontos) seguido da sentença.

Com os arquivos de texto como no Quadro 1, os anotadores deveriam corrigir as opiniões que foram identificadas automaticamente e dividir a sentença quando necessário. Foram definidas combinações de caracteres para especificar as polaridades. Se

⁵A ferramenta usada está disponível em github.com/raphsilva/naive-opinion-miner.

⁶Quando uma pessoa publica uma avaliação sobre um produto no Buscapé, ela deve também indicar se recomenda ou não o produto. Considera-se, no método, que se uma pessoa não recomenda o produto, todas as sentenças que ela escreveu são negativas (análogo para positivas).

Quadro 1. Formato de dados recebido pelos anotadores.

(006.015) [+] [PRODUTO TAMANHO DESEMPENHO] :: Sempre quis um aparelho da linha S e o S7 tem o tamanho perfeito para não chamar muita atenção e desempenho fantástico.

(007.017) [+] [PRODUTO] :: Qualidade.

(007.018) [+] [PRODUTO] :: Produto deixa a desejar, bordas metálicas riscam com facilidade, botão home então nem se fala.

(007.019) [+] [TEL A] :: Complicado de encontrar películas compatíveis com a tela toda.

Quadro 2. Revisão manual feita no exemplo do Quadro 1.

(006.015) [+.] [PRODUTO] :: Sempre quis um aparelho da linha S.
(006.015) [+] [TAMANHO] :: O S7 tem o tamanho perfeito para não chamar muita atenção.
(006.015) [+] [DESEMPENHO] :: Desempenho fantástico.

b(07.017) [.] [PRODUTO] :: Qualidade.

(007.018) [-] [PRODUTO] :: Produto deixa a desejar.

(007.018) [-] [DESIGN] :: Bordas metálicas riscam com facilidade.

(007.018) [-] [OUTRO] :: Botão home ruim.

(007.019) [-] [ACESSÓRIO] :: Complicado de encontrar películas compatíveis com a tela toda.

uma sentença não contivesse informação útil para avaliar o produto, os anotadores deveriam identificar isso no começo da linha com caracteres predefinidos: por exemplo, uma letra ‘b’ no começo de uma linha indica que o texto contido ali está quebrado. O Quadro 2 mostra um exemplo de trabalho dos anotadores.

3.3. Extensão do conjunto

Os dados coletados são opiniões sobre quatro produtos extraídas do site Buscapé. São dois tipos de produtos: celulares e câmeras. Os comentários sobre os celulares formam o subconjunto rotulado **D1**, e os sobre as câmeras formam o conjunto **D2**. Para diversificar os testes, foram criados artificialmente outros casos de teste a partir de D1 e D2.

O conjunto **D3** é um subconjunto de D1 do qual foram excluídas algumas sentenças de modo a equilibrar a quantidade de opiniões positivas e negativas para cada entidade. Ambos os conjuntos D1 e D2 têm muito mais opiniões positivas do que negativas (ver Tabela 1 a seguir). Esse novo conjunto pode ser usado para simular um cenário em que exista forte controvérsia entre as opiniões sobre as entidades.

O conjunto **D4** é um subconjunto de D1 de onde foram excluídas aleatoriamente algumas sentenças para que uma das entidades ficasse com uma quantidade de texto muito menor do que a outra. Com ele, pode-se avaliar casos em que uma das entidades tem mais destaque.

O conjunto **D5** contém os comentários mais recentes de D2, o conjunto **D6** contém os comentários mais antigos de D2 e o conjunto **D8** contém comentários aleatórios de D1. Eles permitem simular situações com conjuntos pequenos.

O conjunto **D7** é um subconjunto de D2 que contém apenas quatro aspectos. Foram escolhidos quatro aspectos (os quatro mais frequentes de D2) e descartadas todas as sentenças que não os citam. Com ele, simula-se situação em que uma entidade tem poucos aspectos.

3.4. Limpeza e simplificação do córpus

Para uso prático em trabalhos de Análise de Sentimentos, foi projetada uma versão limpa do córpus. Além de removidas as sentenças que não são úteis para avaliar o produto a que deveriam se referir (sentenças que foram marcadas pelos anotadores como fora de escopo, irrelevantes, ininteligíveis, etc), foram removidas sentenças genéricas com menos de quatro palavras por se considerar que elas normalmente não ajudam a avaliar as especificidades de um produto (por exemplo, ‘*não gostei*’).

Para reproduzir melhor o que um sistema automático de anotação faria, algumas informações foram simplificadas. Foram descartadas as divisões de sentenças em trechos, que é uma tarefa difícil de ser executada automaticamente em textos do tipo tratado (isto é, que não necessariamente seguem a norma culta). Os três níveis de polaridade positiva foram unidos em um só; o mesmo ocorreu para opiniões negativas. Trechos subjetivos que não são positivos ou negativos foram identificados como neutros.

A limpeza foi feita para evitar ruídos e propagacão de erros em sistemas que usem o córpus, e a simplificação foi feita para deixá-lo em um estado em que possa simular um conjunto de dados anotado automaticamente, o que seria interessante para projetos de Processamento de Linguagem Natural. É importante ressaltar que a versão original do córpus, antes da limpeza e simplificação, também foi disponibilizada, pois pode ser útil em pesquisas futuras.

3.5. Visão geral do córpus

A Tabela 1 mostra a contagem de aspectos identificados, sentenças e opiniões para cada entidade após a extensão, limpeza e simplificação do córpus.

Dos 1.902 trechos de texto anotados, 89% foram marcados como úteis para fins de avaliação do produto em questão. Desses, 84% são opiniões positivas e 14% são negativas; as opiniões marcadas como fortes ou fracas contribuem com 9% das opiniões.

Os tipos de opinião mais raros no conjunto foram as relativas e as duais⁷, com apenas 2 ocorrências cada. Houve 11 ocorrências de trechos irresolutos, 10 de opiniões medianas, 10 de experiência de uso e 6 de conselhos.

Dos 217 trechos não avaliativos ou sem utilidade, 25% são duplicados, 30% são fora de escopo, 21% são textos irrelevantes ou de contextualização e 24% apresentam problemas de texto (quebrados ou ininteligíveis).

3.6. Disponibilização

O córpus pode ser acessado pela página do projeto OPINANDO (sites.google.com/icmc.usp.br/opinando). Duas versões do córpus estão disponíveis:

⁷Observaram-se alguns casos onde havia opiniões positivas e negativas sobre um mesmo aspecto e os anotadores preferiram separá-las, já que haviam sido orientados a fazer isso sempre que possível. Por exemplo, para a sentença ‘*a tela não tem o melhor contraste de cores, mas a nitidez é imbatível*’, os anotadores identificaram uma opinião positiva sobre a tela e uma negativa sobre a tela.

Tabela 1. Estatísticas do conjunto de dados.

tipo	nome	entidade	aspectos	sentenças	opiniões positivas	opiniões negativas
celular D1	D1a	Motorola Moto G5 Plus	15	269	346	101
	D1b	Galaxy S7	14	253	342	91
câmera D2	D2a	Canon EOS Rebel T5	13	68	77	11
	D2b	Canon PowerShot SX520 HS	15	52	68	8
celular D3	D3a	(subconjunto de D1a)	11	150	143	65
	D3b	(subconjunto de D1b)	10	109	85	65
celular D4	D4a	(subconjunto de D1a)	13	43	56	13
	D4b	(cópia de D1b)	14	253	342	91
câmera D5	D5a	(subconjunto de D2a)	12	39	40	10
	D5b	(subconjunto de D2b)	10	30	37	3
câmera D6	D6a	(subconjunto de D2a)	8	29	37	1
	D6b	(subconjunto de D2b)	11	22	31	5
câmera D7	D7a	(subconjunto de D2a)	4	31	33	6
	D7b	(subconjunto de D2b)	4	25	22	4
celular D8	D8a	(subconjunto de D1a)	12	39	62	10
	D8b	(subconjunto de D1b)	12	32	36	15

- Uma versão limpa e estendida, em formato JSON, recomendada para uso direto em processamento de problemas do mesmo tipo que a sumarização contrastiva.
- Uma versão da anotação manual das quatro entidades, que inclui os comentários não pertinentes devidamente marcados como tal e todas as informações da anotação, bem como o script usado para converter a anotação para o formato final do círculo. Essa versão é recomendada para análise manual e para eventuais trabalhos derivados.

4. Considerações finais

Além do uso prático para avaliar sistemas de Análise de Sentimentos, a construção do círculo proporcionou uma experiência que permitiu sistematizar melhor a identificação de opiniões. Essa tarefa, se feita automaticamente, costuma classificar as opiniões em apenas três classes: positiva, negativa e neutra. Com a análise manual, foi possível refinar melhor essa classificação e obter informações que podem ser úteis em futuros trabalhos.

Os números descobertos com a análise do círculo mostram que quase 90% dos trechos encontrados na fonte são úteis para fins de avaliação de produtos, e 70% de todos os trechos foram marcados com polaridade positiva ou negativa pelos anotadores, número que sobe para 86% se considerarem as positivas e negativas fortes e fracas. Apenas 2% das opiniões são classificadas como outro tipo, o que faz ponderar se realmente vale a pena se preocupar em formas eficientes e sistematizadas de classificar esse tipo de texto em aplicações reais. A preocupação maior talvez seria com os 1% de trechos não opinativos e sem utilidade, dos quais 30% são comentários fora de escopo.

Este projeto permitiu obter um córpus que simule um conjunto de dados automaticamente processado com a vantagem de ele ser livre de ruídos. Espera-se que o córpus construído contribua com outros projetos e que as ideias apresentadas aqui somem valor à pesquisa em Análise de Sentimentos e Processamento de Linguagem Natural em língua portuguesa.

Agradecimentos

Agradecemos ao Otávio Augusto Ferreira Sousa, que colaborou com a anotação do córpus, e à FAPESP (processo 17/12236-0) e à Pro-Reitoria de Pesquisa da USP (PRP N. 668) pelo apoio a este projeto.

Referências

- Condori, R. E. L. e Pardo, T. A. S. (2017). Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, 78:124–134.
- Guo, J., Lu, Y., Mori, T., e Blake, C. (2015). Expert-guided contrastive opinion summarization for controversial issues. In *Proceedings of the 24th International Conference on World Wide Web*, páginas 1105–1110.
- Jin, J., Ji, P., e Gu, R. (2016). Identifying comparative customer requirements from product online reviews for competitor analysis. *Engineering Applications of Artificial Intelligence*, 49:61–73.
- Kim, H. D. e Zhai, C. (2009). Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, páginas 385–394.
- Lerman, K. e McDonald, R. (2009). Contrastive summarization: An experiment with consumer reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, páginas 113–116.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers.
- Liu, B., Hu, M., e Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, páginas 342–351.
- Park, S., Lee, K., e Song, J. (2011). Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, páginas 340–349.
- Vargas, F. A. e Pardo, T. A. S. (2018). Aspect clustering methods for sentiment analysis. In *Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language*, LNAI 11122, páginas 365–374.
- Wang, D., Zhu, S., Li, T., e Gong, Y. (2012). Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data*, 6(3):1–18.

Subsídios Linguístico-Computacionais para a Revisão Gramatical Automática de Redações do Ensino Médio

Ariani Di-Felippo^{1,2}, Dayse Simon³, Milena França⁴, Pedro F. Martins³

¹Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Rod. Washington Luís, 235, Caixa Postal 676 - CEP 13565-905, SP– Brazil

²Núcleo Interinstitucional de Linguística Computacional - NILC
Av. Trabalhador Sao-carlense, 400 - Centro, São Carlos, Brasil

³Letrus
Av. Francisco Leitão, 469 – CEP 05414-025, SP, Brazil

⁴Bacharelado em Linguística – UFSCar

{arianidf,milecardfra}@gmail.com, {pedro,dayse}@letrus.com.br

Abstract. We evaluate the LanguageTool grammar checker in a corpus of essays written by High School students in Brazilian Portuguese. Since LanguageTool is a rule-based open-source grammar checker, the grammatical rules with lowest precision have been improved to automate the process of grammar checking of the mentioned essays.

Resumo. Avalia-se o corretor gramatical LanguageTool em um corpus de redações produzidas por estudantes do ensino médio nos moldes do Enem. Posto que se trata de um corretor simbólico de código aberto, tem-se buscado refinar as regras gramaticais de menor precisão com vistas à revisão gramatical automática das referidas redações.

1. Introdução

A correção gramatical automática (em inglês, *grammar checking*) é uma das aplicações do Processamento Automático das Línguas Naturais (PLN) mais amplamente utilizadas, sobretudo acopladas a editores ou processadores de texto como o Microsoft Word, LibreOffice e OpenOffice. A correção gramatical consiste na detecção de problemas gramaticais (como de concordância, regência, uso de pronomes, etc.) quanto à modalidade escrita formal da língua e, por vezes, na sugestão de correções [Soni e Thakur 2018]. Nesse cenário, destacam-se atualmente as ferramentas *open source*, ou seja, sistemas cujo código-fonte é aberto, o qual, por conseguinte, pode ser adaptado para diferentes tarefas. Para o processamento gramatical do português, citam-se o LanguageTool [Naber 2003]¹, o Vero² [Moura 2011] e o CoGroo [Silva 2013].

Embora a revisão gramatical seja atualmente uma área bem consolidada no PLN, a revisão gramatical automática de textos como as “redações escolares”, por exemplo, é

¹ O LanguageTool é, na verdade, um corretor gramatical multilíngue, que não incluía o português em sua versão original [Naber 2003]. Atualmente, esse sistema já é capaz de processar textos nas diferentes variantes do português (<https://languagetool.org/pt-BR/>).

² O Vero era originalmente um verificador ortográfico [Moura 2011]. A partir de 2009, ele passou a englobar um corretor gramatical (<https://pt-br.libreoffice.org/projetos/vero/>).

um desafio para o PLN, uma vez que esses textos possuem problemas variados quanto à frequência de ocorrência e complexidade de tratamento.

Neste trabalho, apresenta-se uma avaliação do LanguageTool em um *corpus* de sentenças extraídas de redações produzidas por alunos do ensino médio como treinamento para o Exame Nacional do Ensino Médio (Enem). Entre os corretores de livre acesso, o LanguageTool fora selecionado por (i) ser um sistema de PLN simbólico (isto é, a identificação dos problemas é baseada em regras manualmente descritas) e (ii) não ter sido tão amplamente avaliado. Assim, ao se identificar os tipos de problemas gramaticais que o corretor detecta com menor eficiência, este trabalho gera subsídios linguísticos para refinar as regras do sistema, contribuindo para que este possa ser utilizado, por exemplo, na revisão gramatical automática de redações do ensino médio.

2. Avaliação do LanguageTool em um *corpus* de redações

Considerando os objetivos do trabalho, utilizou-se um *corpus* constituído de um conjunto de 82.440 sentenças, as quais compõem redações de alunos do ensino médio produzidas como treinamento para o Exame Nacional do Ensino Médio (Enem). O referido *corpus* foi cedido pela Letrus³, que é um centro de tecnologia e letramento que desenvolve ferramentas de escrita e avaliação de textos para escolas. Especificamente, as sentenças foram coletadas em formato digital da plataforma virtual da própria empresa.

Uma vez selecionadas, as sentenças foram submetidas ao LanguageTool, que identificou um conjunto amplo de problemas nesse *corpus* de sentenças com base em 36 regras (cf. Tabela 1), as quais capturam diferentes tipos de problemas classificados como (i) capitalização, (ii) confusão de palavras, (iii) gramática, (iv) miscelânea, (v) pontuação, (vi) redundância, (vii) repetição, (viii) sintaxe e (ix) tipografia.

Do total de problemas identificados pelo corretor, 3 linguistas computacionais analisaram manualmente uma amostra aleatória de 4.043 casos e classificaram os problemas em duas categorias de detecção, comumente utilizadas no PLN [Rino *et al* 2002, Silva 2013], a saber: (i) *verdadeiros positivos* (VP) (isto é, problemas corretamente detectados pelo corretor grammatical) e (ii) *falsos positivos* (FP) (isto é, problemas identificados equivocadamente pelo corretor grammatical). Na sequência, calculou-se de forma automática a tradicional medida de *precisão* (P) [Resnik e Lin 2010].

Especificamente, a medida P indica o número de problemas corretos (de acordo com os especialistas) que foram detectados pela ferramenta em relação ao total que foi detectado. Para calcular P, tem-se a fórmula: $P = (\text{verdadeiros positivos} / (\text{verdadeiros positivos} + \text{falsos positivos}))$ ⁴. O cálculo de P resulta em um valor entre 0 e 1, sendo que, quanto mais próximo de 1, maior é a precisão obtida pela regra.

Na Tabela 1, tem-se as 36 regras utilizadas pelo LanguageTool para a detecção dos 4.043 casos da amostra. Nessa figura, as regras estão organizadas em função das categorias a que pertencem, segundo as informações extraídas da plataforma *online* LanguageTool Community, especificamente da área que contém as regras relativas ao português⁵. Na última linha da tabela, tem-se a precisão média do LanguageTool (81%).

³ <https://www.letrus.com.br/>.

⁴ Por exemplo, a precisão P da Regra 1 da Tabela 1 foi assim calculada: $P = (66 / 66 + 161) = 0.29$.

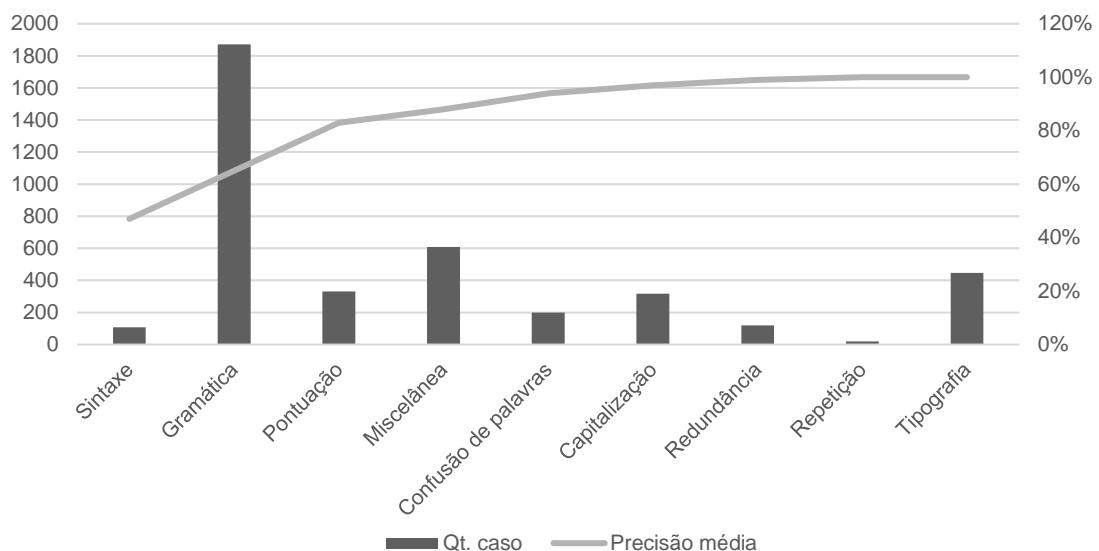
⁵ <https://community.languagetool.org/?lang=pt-PT>.

Tabela 1. Precisão das regras do LanguageTool no *corpus* de estudo.

Categoría	Regra	Qt. casos	VP	FP	P (%)
Capitalização	Capitalização da frase	100	98	2	98
Confusão de palavras	Confusão entre “pratica” e “prática”	100	90	10	90
	Confusão entre “esta” e “está”	217	210	7	96
	Confusão entre “traz” e “trás”	100	99	1	99
Gramática	Concordância de gênero	227	66	161	29
	Concordância verbal	101	31	70	31
	Confusão entre verbo no passado e no futuro	67	24	43	35
	Concordância entre verbo e predicado	100	64	36	64
	Concordância de número	137	93	44	67
	Confusão entre “mau” e “mal”	142	98	44	69
	Confusão entre “a” e “há”	183	138	45	75
	Pronome oblíquo + verbo	155	117	38	75
	Concordância verbo -se + plural	100	82	18	82
	Verbo do tipo “estar” + adjetivo + “de que”	105	90	15	85
	Concordância “ser” + adjetivo	101	89	12	88
	Colocação pronominal	101	95	6	94
	Erro de crase	352	329	14	97
	Expressão prolixia: “mais grande”	12	7	5	58
	Expressão prolixia: “mais bom”	5	3	2	60
Miscelânea	Confusão entre “tem” / “têm”	109	103	6	94
	Remoção de “eu” e “nós”	101	96	5	95
	Ocorrência de “as vezes” (“às vezes”)	100	97	3	97
	Ocorrência de “afim” (“a fim”)	224	224	0	100
	Verbo “estar” + “aonde”	58	58	0	100
	Confusão entre “haver” com “a ver”	24	24	0	100
	Locução entre vírgulas	130	86	44	66
Pontuação	Ausência de pontuação final	100	83	17	83
	Abreviatura “etc”	100	100	0	100
	Conjunção redundante	100	99	1	99
Redundância	Comparativo especial: “mais melhor”	4	4	0	100
	Comparativo especial: “mais pior”	15	15	0	100
Repetição	Palavra repetida	20	20	0	100
Sintaxe	Fragmento: dois artigos seguidos	107	51	56	47
Tipografia	Espaço entre frases	105	105	0	100
	Aspas inteligentes (“ ”)	141	141	0	100
	Ocorrência de espaço antes de pontuação	200	200	0	100
TOTAL		4043	3329	705	81

Com base na Tabela 1, pode-se dizer que, com exceção da categoria “gramática”, as demais englobam problemas bastante pontuais. A categoria “miscelânea”, por exemplo, é composta exclusivamente por regras lexicalizadas. Ademais, ao se cruzar as informações de frequência e precisão média (Figura 1), “gramática” é a categoria mais frequente (1.872 casos) e de menor precisão média (65%).

Figura 1. Frequência e precisão média das regras por categoria.



As regras da categoria “gramática” estão organizadas na Tabela 2 em ordem crescente de precisão, posto que as regras de menor precisão ocupam o topo do ranque. Na última linha da tabela, tem-se a precisão média das regras ditas “gramaticais” do corretor (68%).

Tabela 2. Precisão das regras gramaticais do LanguageTool no *corpus* de estudo.

No.	Regra	Qt. casos	VP	FP	P (%)
1 ^a	Concordância de gênero	227	66	161	29
2 ^a	Concordância verbal	101	31	70	31
3 ^a	Confusão entre verbo no passado e futuro	67	24	43	36
4 ^a	Concordância entre verbo e predicado	100	64	36	64
5 ^a	Concordância de número	137	93	44	67
6 ^a	Confusão entre “mau” e “mal”	142	98	44	69
7 ^a	Confusão entre “a” e “há”	183	138	45	75
8 ^a	Pronome oblíquo + verbo	155	117	38	75
9 ^a	Concordância verbo -se + plural	100	82	18	82
10 ^a	Verbo do tipo “estar” + adjetivo + “de que”	105	90	15	85
11 ^a	Concordância “ser” + adjetivo	101	89	12	88
12 ^a	Colocação pronominal	101	95	6	94
13 ^a	Erro de crase	353	439	14	97
TOTAL		1872	1426	546	68

Partindo-se das regras mais precisas, observa-se, com base na Tabela 2, que a Regra 13, responsável por detectar os problemas de uso de crase, tem precisão de 97%, sendo a mais frequente, com 353 ocorrências na amostra de 4.043 casos. Sobre essa regra, os poucos casos de *falsos positivos* (14) se restringem à ocorrência de formas do verbo *ir* seguidas de “até a” e um substantivo feminino (p.ex.: “[...] olho dentro de casa e vou até”

a porta [...]”), para as quais o corretor sugere “até a”⁶. A Regra 12, que identifica problemas de concordância pronominal também possui uma precisão relativamente alta de 82%. Como exemplo de *falso positivo*, cita-se o caso de “Portanto conclui-se que [...].” Devido à ausência de vírgula depois de “portanto”, o LanguageTool sugere equivocadamente a anteposição do pronome (“Portanto se conclui [...]”).

As Regras 11, 10, 9, 8, 7 e 6 são todas lexicalizadas e possuem precisão mediana, variando de 69% a 88%. Diz-se “lexicalizada” porque a aplicação destas requer a ocorrência de palavras específicas. A Regra 11, por exemplo, aplica-se somente mediante a ocorrência do verbo “ser”. Nesse sentido, pode-se questionar a classificação da Regra 8 como lexicalizada. No entanto, essa regra, ao lidar com pronomes oblíquos, também tem seu espaço de aplicação bastante restrito.

As Regras 5, 4, 3, 2 e 1 tratam de fenômenos mais genéricos, como as concordâncias de gênero (Regra 1) e número (Regra 5). Entre elas, as Regras 3, 2 e 1, apresentam os menores índices de precisão. A Regra 3, com uma precisão sutilmente mais elevada (36%) que 2 e 1, diz respeito especificamente à confusão entre as formas verbais no passado (p.ex.: “andaram”) e no futuro (p.ex.: “andarão”). Esse problema é relativamente pouco frequente, já que houve apenas 67 casos na amostra. Observa-se que a Regra 2, de concordância verbal, e a Regra 1, de concordância nominal de gênero, possuem valores de precisão muito próximos, no caso, 31% e 29%, respectivamente.

Entre os vários *falsos positivos* gerados pela Regra 2 (70) estão casos como o grifado a seguir “A mistura cultural entre eles leva mais conhecimento [...],” para os quais o corretor sugere que haja concordância entre o verbo e o elemento que ocorre imediatamente à sua esquerda. Para essa ocorrência em particular, o LanguageTool sugere “A mistura cultural entre eles levam mais conhecimento [...].” Esse tipo de equívoco parece decorrente da incapacidade do sistema em identificar dependências de mais longa distância, como é o caso da relação entre o verbo “leva” e o núcleo do sintagma nominal sujeito, “mistura”, que ocorre na quarta posição à esquerda do verbo.

Entre os *falsos positivos* gerados pela Regra 1 (161) estão casos como o grifado a seguir “Com a porta dos estudos aberta, [...]”). Para o LanguageTool, “estudos” e “aberta” devem concordar em gênero (e também em número), sugerindo (erroneamente) a reescrita do trecho para “estudos abertos”. No caso, o corretor não reconhece que o adjetivo “aberta” modifica “porta”, que é o núcleo do sintagma nominal “a porta dos estudos”. Isso parece ocorrer porque, embora a correção gramatical do inglês conte com um *part-of-speech tagger* derivado de [Brill 1992], que é um dos melhores da literatura, e um *chunker* (isto é, ferramenta que identifica os sintagmas), o mesmo parece não estar disponível para o processamento do português. Aparentemente, as regras do corretor são compostas por expressões regulares capazes de contemplar de forma mais ampla as concordâncias entre elementos adjacentes, como “determinante + nome” (“a porta”)).

Dos dados da Tabela 2, destaca-se também que as regras, classificadas no sistema como “gramaticais”, detectam no geral tipos de problemas que são cobertos pela tipologia de 17 categorias proposta por Pinheiro [Pinheiro 2008] a partir da análise manual do *corpus* CORVO, composto por 249 redações do Enem [Pinheiro 2008]. Isso indica que a referida tipologia, embora proposta em 2008, é viável para um estudo mais amplo dos tipos de problemas em redações produzidas por estudantes do ensino médio.

⁶ A análise da aplicação dessa regra se baseou em gramáticas e dicionários que não recomendam o uso da preposição “a” após “até”. Assim, entende-se que, em “até a porta”, o “até” está seguido de um artigo.

3. Considerações Finais

O trabalho ora descrito revelou as regras que apresentam os maiores valores de *falsos positivos* nas redações do tipo Enem produzidas por estudantes secundaristas. Diante disso, tem-se estudado a (i) arquitetura do corretor para compreender mais amplamente como se dá a revisão gramatical e, sobretudo, (ii) as informações contidas nas regras, disponíveis *online* em formato *xml*. Na sequência, pretende-se propor refinamentos para as regras, o que pode ser feito por meio do “editor de regras” *online* do LanguageTool⁷. Acredita-se que tais refinamentos podem consistir, por exemplo, em ampliar os padrões previstos pelas expressões regulares e a indicação da necessidade de processos adicionais (p.ex.: *tagging*, *parsing* e correção ortográfica) para o tratamento de outras regras.

Agradecimentos. Ao CNPq, pelo suporte financeiro (Proc. Nº 401175/2018-9).

Referências

- Brill, E. (1992). A simple rule-based part of speech tagger. In Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92), p. 152-155.
- Moura, R. (2011). Vero, the Brazilian Portuguese Spell Checker. Disponível em <http://www.broffice.org/verortografico>, Janeiro.
- Naber, D. A. (2003). Rule-Based Style and Grammar Checker. 2003. Diplomarbeit. Technische Fakultät, Universität Bielefeld. Bielefeld.
- Pinheiro, G. M. (2008). Redações do ENEM: estudo dos desvios da norma padrão sob a perspectiva de corpos. 2008. 152f. Dissertação (Mestrado em Linguística) - Faculdade de Filosofia, Letras e Ciências Humanas - FFLCH, Universidade de São Paulo.
- Resnik, P.; Lin, J. (2010). Evaluation of NLP Systems. In: Clark, A; Fox, C; Lappin, S. (Ed.). The Handbook of Computational Linguistics and Natural Language Processing. Oxford: Wiley-Blackwell, p. 271-295.
- Rino, L.H.M.; Di Felippo, A.; Pinheiro, G.M.; Martins, R.T.; Filié, V.M.; Hasegawa, R.; Nunes, M.G.V. (2002) Aspectos da construção de um revisor gramatical para o português do Brasil. In *Estudos Linguísticos*, v. 31. São Paulo, Brasil. ISSN 1413 0939. 1 CD-ROM.
- Silva, W.D.C.M. (2013). Aprimorando o Corretor Gramatical CoGrOO. 2013. 178f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.
- Soni, M., Thakur, J.S. (2018) A Systematic Review of Automated Grammar Checking in English Language. Submitted to Computational Linguistics. Disponível em <https://arxiv.org/pdf/1804.00540.pdf>

⁷ <https://community.languagetool.org/ruleEditor2/>

Discriminação de palavras e efeitos da variação linguística

Raquel Meister Ko. Freitag¹, Victor Rene Andrade Souza²

¹Departamento de Letras Vernáculas – Universidade Federal de Sergipe (UFS)
São Cristóvão – SE – Brazil

²Departamento de Letras Vernáculas – Universidade Federal de Sergipe
São Cristóvão – SE – Brazil

rko.freitag@uol.com.br, victor.andrade573@gmail.com

Abstract. This paper shows how the degree of social evaluation of phonological variables interferes in its judgment as word in an experimental task of discrimination. The trial effects were tested for the variants of monophthongtion, desnasalization and palatalization of alveolar stops in regressive and progressive contexts in a sample of 70 university students from Sergipe. The results indicate that stereotype and marker type variants are significantly associated with non-word; while the indicator variants do not show this relation.

Resumo. Este estudo investiga como o grau de apreciação social de um fenômeno variável fonológico interfere no seu julgamento como palavra em uma tarefa experimental de discriminação. Foram testados os efeitos de julgamento para as variantes de monotongação, desnasalização e palatalização de oclusivas, em ambientes regressivo e progressivo em uma amostra de 70 universitários de Sergipe. Os resultados apontam que as variantes do tipo estereótipo e marcador são significativamente associadas à não palavra; enquanto as variantes indicador não apresentam essa relação.

1. Introdução

A distinção entre o que é uma palavra ou não de uma língua é uma tarefa importante para o processamento automático da linguagem. Há diferentes critérios a serem adotados, desde critérios lexicais e morfossintáticos, amplamente explorados, até critérios sociolinguísticos que considerem a variação linguística e o papel da avaliação dos falantes no reconhecimento baseado na avaliação social de uma forma linguística.

O nível de avaliação social de uma forma sociolinguística é, classicamente, utilizado para diferenciar variáveis sociolinguísticas quanto à consciência social do falante: do menos consciente (indicador), passando por forma sensível à estratificação estilística (marcador) até o nível consciente (estereótipo) [Labov 1972]. No entanto, o que faz com que uma variável seja sensível ou não à avaliação em uma comunidade pode ser atrelado ao seu grau de saliência; para isso, apenas a identificação da avaliação social das variáveis e variantes não é suficiente; é preciso adentrar no domínio da sociolinguística da percepção (que tem como objeto o julgamento do ouvinte), que correlaciona fatores sociais a traços sociolinguísticos, a fim de contribuir para o desvelamento de um padrão de consciência social na comunidade.

A definição de critérios testáveis para atribuir saliência a formas em situações de contato dialetal e mudança linguística tem sido uma tarefa da sociolinguística e envolve os efeitos externos e internos à língua [Trudgil 1986]. Como efeito externo, é pista da avaliação se uma forma/variável que está passando por mudança sofre estigmatização, e como efeito interno, a saliência medida pela distância fonética e contraste: falantes são mais conscientes das variáveis com variantes que são foneticamente diferentes e de variantes envolvidas na manutenção do contraste fonológico.

Trudgill (1986) argumenta que os fatores de saliência levam a indicadores (formas que não mostram variações estilísticas e falantes não são conscientes) a se tornarem marcadores. Labov (1972) diferencia indicadores e marcadores a partir de estereótipos (formas que são objetos de avaliação social consciente). No entanto, formas estereotipadas, segundo Labov (1972), podem desaparecer por conta da avaliação social consciente, tornando-se cada vez mais distantes do uso real. Assim, uma questão sociolinguística que carece de investigação é: quais formas linguísticas são salientes e que tipos de significado social tais formas podem carregar?

Tradicionalmente, a mensuração do grau de saliência social decorre de pistas distribucionais de produção, com a identificação do perfil sociolinguístico do falante quanto à escolarização, faixa etária, zona de residência. Mais recentemente, têm sido incorporados os resultados de estudos de percepção [Drager 2018, Freitag 2018], considerando os diferentes tipos de tarefas experimentais, que permitem maior controle dos efeitos das variáveis testadas.

A convergência de estudos de produção e percepção possibilita a compreensão mais ampla do processo de variação e mudança linguística [Freitag, Severo, Rost-Snichelotto and Tavares 2016]. Fenômenos sociolinguísticos conscientes e estigmatizados, como estereótipos negativos, são facilmente identificáveis pelos rótulos de avaliação societal que recebem: “errado”, “feio”, “não sabe falar a língua X”, “não é palavra da língua X” [Freitag and Santos 2016]. A abordagem experimental para distinguir palavra e não palavra aplicada a variáveis com diferenças quanto ao nível de consciência social pode trazer pistas da percepção na comunidade, e contribuir para o estabelecimento de escalas de apreciação social de variáveis. E saber o que um falante de uma dada variedade considera como palavra ou não palavra contribui para o processamento automático da língua.

Utilizando a técnica de tarefa de discriminação para fenômenos fonológicos variáveis [Drager 2013], neste estudo apresentamos o resultado de um teste de lexicalidade com estímulos de três fenômenos variáveis na comunidade de fala de Aracaju, Sergipe, Brasil: a monotongação de ditongos decrescentes, com comportamento de indicador, a desnasalização de ditongo átono final, com comportamento de marcador, e a palatalização de oclusivas, em ambientes regressivo e progressivo, com comportamento de estereótipo (Figura 1).

O processo de monotongação consiste no apagamento do glide palatal [j], como em *caixa*, ou velar [w] em ditongo decrescente, como em *cenoura*. No português brasileiro, este fenômeno apresenta condicionamentos distintos a depender da natureza do glide e do contexto fonológico seguinte: o apagamento do glide velar tende a ser uma regra categórica em situações informais, em todos os contextos linguísticos, inclusive na escrita [Araujo and Borges 2019, Gonçalves and Amaral 2014, Simioni and Rodrigues

2014, Toledo 2013, Haupt and Seara 2012, Cristofolini 2011]. O glide palatal apresenta restrições de natureza interna, decorrente do contexto fonológico seguinte: é regra semicategórica, independentemente da formalidade, para contextos em que a sílaba seguinte apresenta o traço palatal, como em *caixa*, *beijo*, e é restrinido por contextos em que a sílaba seguinte é iniciada por oclusiva, como em *leito*, *caibo*. O comportamento é estável em todas as regiões, sem sensibilidade social ou dialetal. Na tipologia de apreciação social, a variante é considerada um indicador.

O processo de desnasalização de ditongo átono final consiste no apagamento do segmento nasal, em nomes, como em *vagem*, e em verbos na terceira pessoa, como em *passaram*. Nos verbos, o apagamento interfere em relações morfossintáticas, e ocorre de modo estável em todo o português brasileiro, com comportamento relativamente sensível ao contexto de monitoramento estilístico. Já nos nomes, a variante desnasalizada está associada a aspectos sociais relativos à escolarização e ruralidade, além de ser relativamente sensível ao contexto de monitoramento estilístico [Silva 2018, Bona and Schwindt 2017, Gomes, Mesquita and Fagundes 2013, Silva, Fonseca and Cantoni 2012, Battisti 2000, Paiva 1996]. Na tipologia de apreciação social, é considerada um marcador.

Processo	Exemplo	Padrão na comunidade	Ocorrência	Tipo de avaliação	
		- monitorado	+ monitorado	dialetal social	
Monotongação -ow	 cen/ow/ra	nunca	frequente	não	Indicador
	 cen/o/_ra	sempre	às vezes	não	
Monotongação -aj, -ej	 c/aj/xá	às vezes	frequente	não	marcador
	 c/a/_xa	frequente	às vezes	não	
Palatalização regressiva	 vestido	frequente	frequente	sim	marcador
	 ves/t/ido	às vezes	às vezes	não	
Desnasalização ditongo final átono	 vag/eN/	frequente	sempre	não	estereótipo
	 vag/e/_	às vezes	raro	sim	
Palatalização progressiva	 oito	frequente	sempre	sim	estereótipo
	 oi/t/o/	às vezes	raro	sim	

Figura 1. Recorrência e avaliação social das variáveis.

O processo de palatalização de /t/ e /d/ ocorre, no português brasileiro, em dois contextos fonológicos distintos, que apresentam distribuição e valoração social distintas. Quando a palatalização é desencadeada por vogal alta [i] posterior à consoante, como em *vestido*, tem-se o processo de palatalização regressiva; a variante resultante, a realização palatal /tʃ/ e /dʒ/, é tida como de prestígio, inclusive abonada em instrumentos normativos. Já a realização oclusiva tem recorrência dialetal (região Nordeste) e social (ruralidade), sendo vista com estigma. Em Sergipe, onde predomina a variante oclusiva, observa-se um processo de mudança incipiente para a implementação da variante palatal, com sinalização de estereótipo positivo [Pinheiro *et al.* 2017, Corrêa and Ribeiro 2018, Pinheiro, Silva and Cardoso 2018, Freitag and Souza 2017, Freitag and Santos 2016, Souza Neto 2014].

Quando a palatalização é desencadeada por glide palatal [j] anterior à consoante, como em *oito*, tem-se o processo de palatalização progressiva; a variante resultante, a realização palatal /tʃ/ e /dʒ/, tem realização dialetal (região Nordeste) e, onde ocorre, é

socialmente demarcada, associada às faixas etárias mais velhas, com menos escolarização e de região rural, o que sinaliza para uma avaliação como estereótipo negativo. Em Sergipe, observa-se, neste contexto, processo de mudança mais avançado para a implementação da variante oclusiva, que tem comportamento de indicador [Freitag 2015, 2019, Pinheiro *et al.* 2017, Corrêa and Ribeiro 2018, Pinheiro, Silva and Cardoso 2018, Freitag and Souza 2017, Freitag and Santos 2016, Souza Neto 2014].

A diferença na avaliação social das variantes pode influenciar os falantes de uma variedade da língua na atribuição de estatuto lexical para as palavras. Enquanto traços indicadores e estereótipos positivos não interferem no julgamento; possivelmente marcadores, e certamente estereótipos negativos, devem interferir no que falantes de uma variedade consideram como palavra ou não palavra. A abordagem experimental para distinguir palavra e não palavra aplicada a variáveis com diferenças quanto ao nível de consciência social pode trazer pistas da percepção na comunidade, e contribuir para o estabelecimento de escalas de apreciação social de variáveis, e essa distinção pode auxiliar nos estudos para o processamento automático da língua.

2. Método

A tarefa experimental desenvolvida constitui-se no julgamento de estímulos linguísticos (Quadro 1) em “palavra” ou “não palavra” do português. Foi constituído um *corpus* de estímulos de palavras lexicais escolhidas pelos critérios de familiaridade (deveriam ser produtivas na língua), de tamanho (mesmo número de sílabas e tonicidade) e de barramento de outros fenômenos sociolinguísticos variáveis que poderiam interferir no julgamento (sílabas travadas por R ou S, alcance vocálico).

Cada palavra foi enunciada por um único locutor (universitário, representativo da variedade de prestígio da fala de Aracaju, Sergipe, segundo seus pares), em uma frase-veículo para favorecer a realização não-final. Posteriormente, o áudio foi recortado para a produção dos estímulos da tarefa de discriminação, que foram divididos em três conjuntos: i) palavras com estímulos-alvo, relativos aos fenômenos variáveis sob análise, contemplando a variante padrão e a não padrão: monotongação, palatalização regressiva, desnasalização e palatalização progressiva; ii) palavras distratoras, ou seja, palavras do português brasileiro que barram a variação linguística; e iii) pseudopalavras, isto é, palavras com a fonotaxe do português brasileiro, mas não se configuraram como item lexical (Figura 2).

Pseudopalavras	Distratores		Monotongação		Palatalização regressiva		Desnasalização		Palatalização progressiva	
			ditongo	monotonogo	occlusiva	palatal	nasal	desnasalizada	occlusiva	palatal
tapi	fala	cinema								
sífo	tabela	crise								
pimada	dúvida	galo								
decato	genro	grupo	l/ow/co	p/o/_co	es/ti/lo	ves/tʃ/ido	passa/ʒeN/	garag/e_/_	coi/t/ado	noi/tʃ/e
navela	saia	céu	f/aj/xā	p/a/_xāo	vin/ti/_	capace/tʃ/e	mensag/ʒeN/	hom/e_/_	i/d/oma	i/dʒ/ota
tonicote	água	coração	am/ej/xā	qu/e/_xo	ban/di/do	men/dʒ/igo	viag/ʒeN/	maquiag/e_/_	ga/t/inho	oi/tʃ/enta
esface	azia	anta	bated/ej/ra	brigad/e/_ro	se/t/e	bo/dʒ/e	malandrag/ʒeN/	paisag/e_/_	pei/t/o	respei/tʃ/o
tixo	lama	lenda	vass/ow/ra	/o/_ro	bal/d/e	/tʃ/ime	vantag/ʒeN/	ont/e_/_	vi/d/eo	doi/dʒ/o
apilca	jiló	conta								
orisia	zero	época								
boreza	seco	data								

Figura 2. Conjunto dos estímulos linguísticos do teste de discriminação com tarefa de decisão lexical.

A tarefa experimental consiste no julgamento binário de uma série randomizada dos itens de i), ii) e iii). O participante, após ouvir cada item, uma única vez, precisa decidir se o item é uma palavra ou não do português. A tarefa foi desenvolvida no software *OpenSesame* [Mathôt, Schreij and Theeuwes 2012]. A aplicação foi realizada com 70 participantes, todos estudantes de diversos cursos de graduação da Universidade Federal de Sergipe, selecionados aleatoriamente e por voluntariedade. A tarefa foi realizada individualmente, em uma cabine acústica, com um computador Dell *Precision* 5400. Após a coleta, os dados foram tratados quantitativamente. Foram realizados testes-t de amostras não pareadas considerando a decisão lexical e o tipo de estímulo, análise da variância entre a resposta e o tempo das respostas, e testes-t para amostras independentes entre as médias de tempo de resposta nas decisões lexicais positivas e os tipos de estímulo. A visualização gráfica dos resultados foi desenvolvida com o pacote *ggstatsplot* [Patil and Powell 2018] para a plataforma R.

3. Resultados e discussões

Os resultados das respostas aos estímulos linguísticos estão apresentados em uma escala: a primeira coluna é a das pseudopalavras; a segunda coluna é a das palavras distratoras, e em seguida, aos pares estão apresentadas as respostas às palavras com estímulos-alvo padrão e não padrão, partindo da escala de apreciação social pré-estabelecida de indicador, marcador e estereótipo (figura 3).

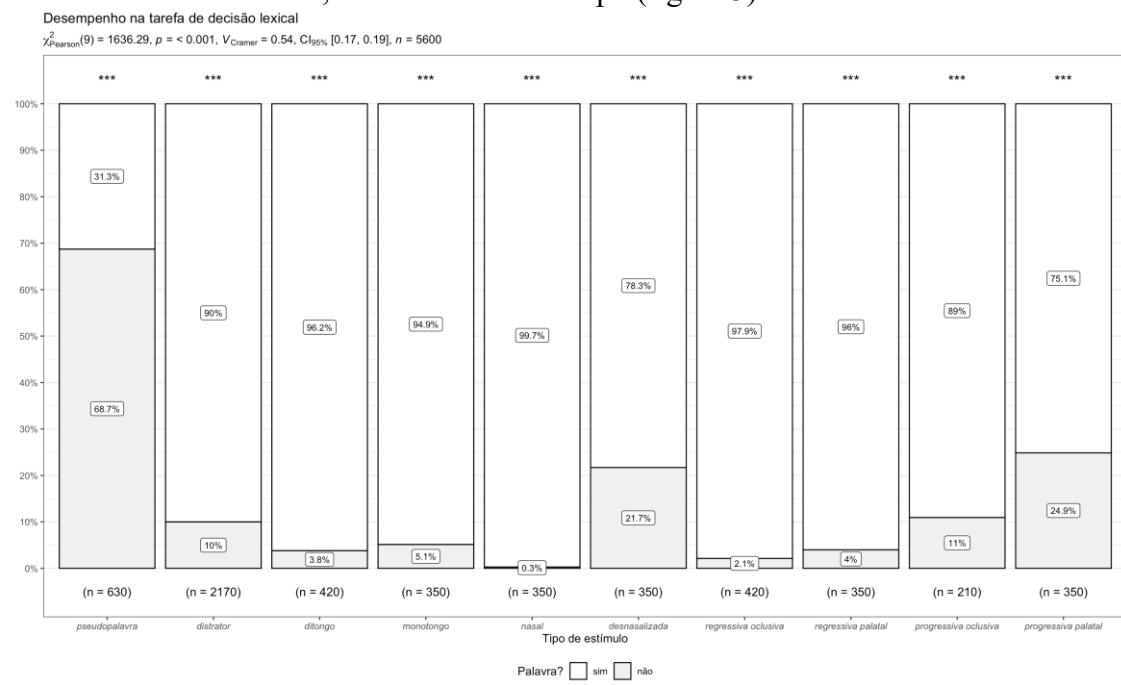


Figura 3. Percentual de identificação como palavra para cada tipo de estímulo.

O percentual de julgamento como palavra para as pseudopalavras é de 31,3%, enquanto para as palavras distratoras é de 90%, diferença estatisticamente significativa ($t(15.578) = 56.865$, $p < .0001$). O teste-t para amostras independentes entre pseudopalavras e os demais tipos de estímulo aponta que a diferença é estatisticamente significativa para todos os outros grupos.

Quando a comparação é entre a variante padrão e não padrão, a realização de ditongo átono final nasal tem 99,7% de decisões como palavra, contra 78,3% da

realização desnasalizada, diferença estatisticamente significativa ($t(69.05) = 6.22, p < .0001$), assim como a comparação entre a realização oclusiva em ambiente antecedido por glide (progressiva oclusiva), 89% de decisões como palavra, contra 75,1% de realização palatal ($t(116.59) = 3.35, p < .0001$). O resultado corrobora a hipótese de que variantes avaliadas como marcador (realização desnasalizada) e estereótipo negativo (realização palatal em ambiente progressivo) tendem a ser consideradas como “não palavras” da língua, enquanto estereótipos positivos (realização palatal em ambiente regressivo) ou com distribuição não saliente, como a realização monotongada, não apresentam diferenças de julgamento que interferem no teste de lexicalidade.

Como medida do desempenho *online* da tarefa, foi considerado o tempo de reação entre o fim da reprodução do estímulo e a resposta, com a hipótese de que o maior tempo despendido para execução da tarefa de decisão lexical pode sinalizar uma maior demanda cognitiva, no caso das pseudopalavras, e também no caso de variantes estigmatizadas [Freitag 2019].

O resultado do teste de decisão lexical aponta que há diferença no tempo de respostas entre uma decisão lexical positiva, ou seja, os participantes consideraram o estímulo julgado como uma palavra (à direita na figura 4) e uma decisão lexical negativa (à esquerda da figura 4). O tempo médio de resposta para uma decisão negativa é superior ao de uma decisão positiva, com exceção dos estímulos de pseudopalavras, e a análise da variância aponta que o efeito principal do tipo de estímulo é significativo no tempo de resposta da decisão lexical ($F(9, 4696) = 25.53, p < .001$) somente para as respostas positivas.

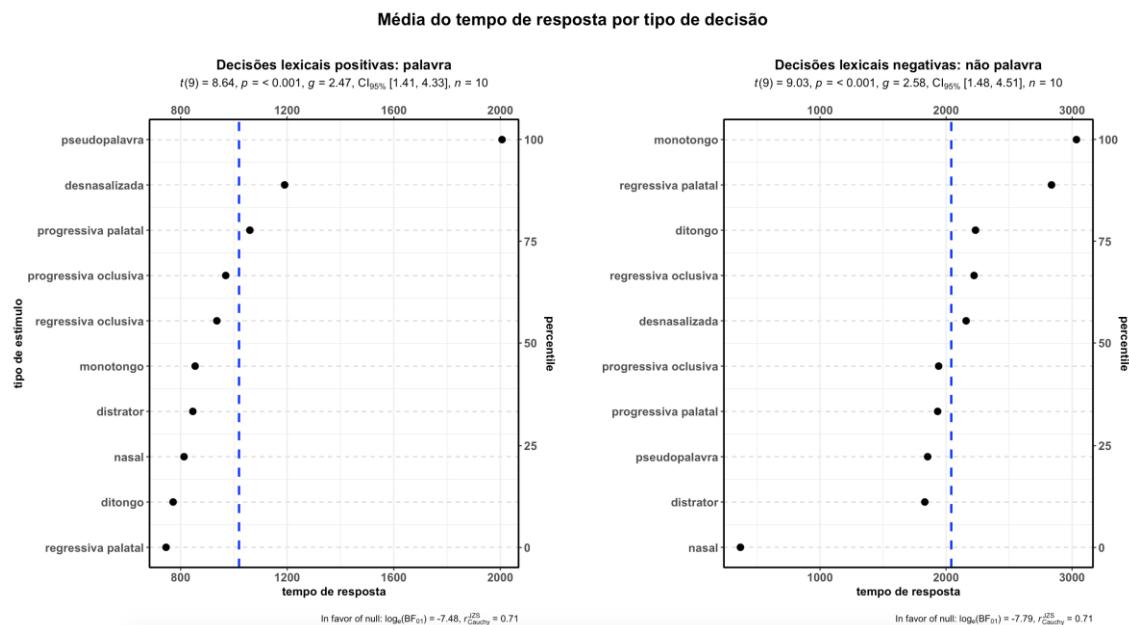


Figura 4. Tempo médio de resposta para cada tipo de estímulo e por tipo de decisão.

Nas decisões lexicais positivas, como esperado, o conjunto de estímulos que demandou maior tempo de resposta foi o das pseudopalavras. Na comparação entre a média do tempo de resposta entre a variante padrão e não padrão em cada processo, todas as variantes tidas como não padrão apresentam tempo de resposta superior à

variante tida como padrão em situações formais na comunidade: a realização de ditongo tem tempo médio de 739.61ms, e a variante monotongada, 815.19ms; a realização da átona final nasal tem tempo médio de 725.83ms, e a variante desnasalizada 1233.34ms; a realização de /t,d/ em ambiente seguinte à vogal alta, regressiva oclusiva, apresenta tempo médio de 892.61ms; e a realização palatal, 716.68ms. Em relação à tendência central do tempo de resposta e o tipo de estímulo avaliado quanto ao julgamento social, o resultado deste tipo de teste difere do resultado obtido em um teste de monitor sociolinguístico [Freitag 2019], no qual estímulos com variantes estigmatizadas (estereótipo negativo) são julgados mais rapidamente do que estímulos com variantes do tipo indicador.

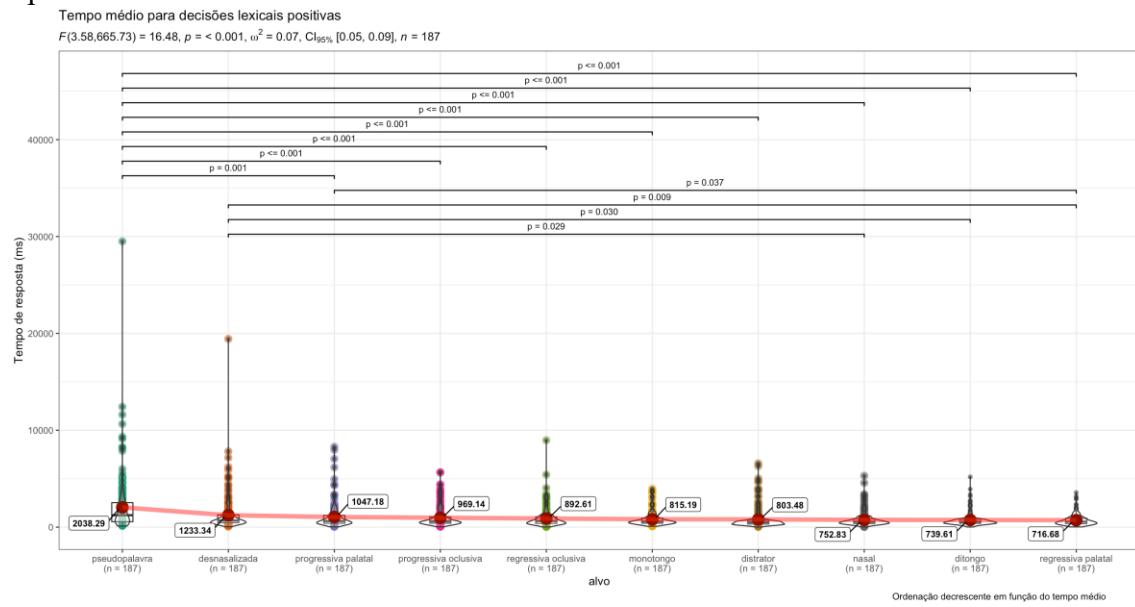


Figura 5. Comparação entre grupos nas decisões lexicais positivas.

Diferenças metodológicas podem ajudar a entender o resultado: enquanto o teste de decisão lexical é consciente, com tarefa explícita, o monitor sociolinguístico é um teste inconsciente, com tarefa indireta. Este resultado contribui para o desenvolvimento de pesquisas em sociolinguística variacionista que fazem uso de técnicas experimentais da psicolinguística para ampliar o poder analítico e preditivo, incorporando mais confiabilidade. Ao mesmo tempo, pesquisas de linguística de *corpus* podem incorporar abordagens experimentais de desenho simples, como a de decisão lexical, para validar o que um falante de uma dada variedade considera como palavra ou não palavra, contribuindo para o processamento automático da língua.

4. Conclusão

O trabalho investigou como o grau de apreciação social de fenômenos variáveis no nível fonológico interfere no seu julgamento como palavra em uma tarefa experimental de discriminação, como o teste de lexicalidade.

Os resultados mostram que palavras sujeitas aos fenômenos estudados recebem mais julgamentos de não palavra do que palavras não afetadas pelos fenômenos. Os resultados do estudo experimental quanto à tarefa de discriminação e ao tempo de resposta apontam que as variantes do tipo estereótipo negativo e marcador são significativamente associadas à não palavra; enquanto as variantes do tipo indicador e

estereótipo positivo não apresentam essa relação. O tempo de decisão lexical positiva e negativa é estatisticamente significativo, também com diferenças no julgamento de variantes do tipo estereótipo negativo e marcador.

Os resultados podem contribuir para a proposição de um critério testável para diferenciar variáveis sociolinguísticas que leve em conta a apreciação social dessas variáveis, a partir da interface com a psicolinguística.

Referências

- Araujo, A. S. and Borges, D. K. V. (2019) “Atitudes linguísticas de estudantes universitários: o fenômeno da monotongação em foco”. In *Tabuleiro de Letras*, v. 12, p. 97-113.
- Battisti, E. (2000) “A redução variável dos ditongos nasais átonos no português do sul do Brasil”. In *Letras de hoje*, v. 35, n. 1.
- Bona, C. D. and Schwindt, L. C. D. S. (2017) “O papel da frequência lexical na desnasalização do ditongo final átono [E᷑N] em não verbos no português do sul do Brasil”. In *Cadernos do ILN*, v. 54, p. 27-46.
- Corrêa, T. R. A. and Ribeiro, C. C. S. (2018) “Avaliação social da palatalização de /t, d/ em Sergipe”. In *A Cor das Letras*, v. 19, p. 109-123.
- Corrêa, T. R. A. (2018) “Estereótipo, estigma e preservação de faces: a realização africada de oclusivas alveolares seguidas de glide palatal em uma comunidade escolar de Aracaju/SE”. In *Caderno Seminal*, v. 30, n. 30, p. 316-344.
- Cristofolini, C. (2011) “Estudo da monotongação de [ow] no falar florianopolitano: perspectiva acústica e sociolinguística”. In *Revista da Abralin*, v. 10, n. 1.
- Drager, K. (2018) “Experimental research methods in sociolinguistics”, Bloomsbury Publishing.
- Freitag, R. M. K. (2015) “Socio-stylistic aspects of linguistic variation: schooling and monitoring effects”. In: *Acta Scientiarum. Language and Culture*, v. 37, n. 2, p. 127-136.
- Freitag, R. M. K. (2018) “Saliência estrutural, distribucional e sociocognitiva”. In *Acta scientiarum. Language and culture*, v. 40, n. 2, p. 41173.
- Freitag, R. M. K. (2019) “Effects of the on-line linguistic processing: palatals in Brazilian Portuguese”. In *Upenn Working Papers in Linguistics*.
- Freitag, R. M. K. and Santos, A. O. (2016) "Percepção e atitudes linguísticas em relação às africadas pós-alveolares em Sergipe. In *A Fala Nordestina: entre a sociolinguística e a dialetologia*, Blucher, São Paulo, p. 109-122.
- Freitag, R. M. K., Severo, C. G., Rost-Snichelotto, C. A. and Tavares, M. A. (2016) “Como os brasileiros acham que falam? Percepções sociolinguísticas de universitários do Sul e do Nordeste”. In *Todas as Letras*, v. 18, n. 2, p. 64-84.
- Freitag, R. M. K. and Souza, G. G. A. (2017) “O caráter gradiente vs. discreto na palatalização de oclusivas em Sergipe”. In *Tabuleiro de Letras*, v. 10, n. 2, p. 78-89.

- Gomes, C. A., Mesquita, C. and Fagundes, T. D. S. (2013) "Revisitando a variação entre ditongos nasais finais átonos e vogais orais na comunidade de fala do rio de janeiro". In *Revista Diacrítica*, v. 27, n. 1, p. 153-173.
- Gonçalves, G. F. and do Amaral, V. S. (2014) "Produções orais e escritas dos ditongos [aj], [ej] e [ow]: dados de São José do Norte/RS". In *Diadorim: revista de estudos linguísticos e literários*, v. 14, p. 127-154.
- Haupt, C. and Seara, I. C. (2012) "Caracterização acústica do fenômeno de monotongação dos ditongos [aj, ej, oj] no falar florianopolitano". In *Revista Linguagem & Ensino*, v. 15, n. 1, p. 263-290.
- Labov, W. (1972) "Sociolinguistics Patterns". University of Pennsylvania, Philadelphia.
- Mathôt, S., Schreij, D. and Theeuwes, J. (2012) OpenSesame: Um construtor de experimentos gráficos de código aberto para as ciências sociais, *Behavior Research Methods*, v. 44, n. 2, p. 314-324.
- Paiva, M. C. (1996) "Supressão das semivogais nos ditongos decrescentes. Padrões sociolinguísticos: análise de fenômenos variáveis do português falado na cidade do Rio de Janeiro", Rio de Janeiro: Tempo Brasileiro, p. 217-236.
- Patil, I. and Powell, C. (2018) ggstatsplot: "ggplot2", Based Plots with Statistical Details.
- Pinheiro, B. F. M., Silva, L. S., Cardoso, P. B. (2018) "Como estudantes do ensino médio acham que falam? Crenças sobre a palatalização de oclusivas e expressão da 1a pessoa do plural". In *A Cor das Letras*, v. 19, n. 41, p. 180-195.
- Pinheiro, B. F. M., et al. (2017) "Processos fonológicos que passam da fala para a leitura". In *Leitura, escrita e literatura: interseções e convergências*, Azevedo, I. C. M. and Roiphe, A. (ed.), São Cristóvão: Editora UFS, p. 10-25.
- Silva, C. C. C. (2018) "Um estudo sobre a redução dos ditongos nasais na fala fluminense". In *Diadorim: revista de estudos linguísticos e literários*, v. 20, p. 409-427.
- Silva, T. C., Fonseca, M. S. and Cantoni, M. (2012) "A redução do ditongo [ãw] postônico na morfologia verbal do português brasileiro: uma abordagem baseada no uso". In *Letras de Hoje*, v. 47, n. 3, p. 283-292.
- Simioni, T. and Rodrigues, É. L. (2014) "Monotongação de ditongos orais decrescentes na escrita de crianças de séries iniciais". In *Letrônica*, v. 7, n. 2, p. 695-712.
- Souza Neto, A. F. (2014) "Realizações dos fonemas /t/ e /d/ em Aracaju-SE", Editora UFS, São Cristóvão.
- Toledo, E. E. (2013) "Estudo em tempo real da monotongação do ditongo decrescente/ej/em amostra de Porto Alegre". In *Letrônica*, v. 6, n. 1, p. 94-107.
- Trudgill, P. (1986) *Dialects in contact*. Oxford, Blackwell.

Análise das relações entre disciplinas do Ensino Médio do Brasil por meio de questões de vestibular com uso de técnicas de PLN

Rafael Telles, Margarethe Steinberger-Elias, André Kazuo Takahata, Luneque Silva Junior,
CECS, Engenharia de Informação, Universidade Federal do ABC,
Santo André, SP, Brasil,
e-mail: rafael.telles@aluno.ufabc.edu, {mborn, andre.t, luneque.junior}@ufabc.edu.br

I. INTRODUÇÃO

As fronteiras entre campos de saber emergentes, tais como Bioengenharia, Nanociência e Ciência Cognitiva, vêm se tornando cada vez mais porosas a partir do avanço do conhecimento e da interdisciplinaridade no século XXI [4] [5]. Tais fronteiras, lexicalmente demarcadas através de um conjunto de termos específicos e exclusivos, tendem a flexibilizar-se e algumas categorias mais rígidas podem abrir espaço para empréstimos lexicais entre disciplinas. A proposta deste trabalho é desenvolver estudo preliminar com apoio da Linguística de *Corpus* [1] [6] sobre como este processo vem afetando as disciplinas do ensino médio no Brasil. Através de um *corpus* composto por questões de vestibular, o objetivo geral deste trabalho é identificar eventuais superposições lexicais entre as diferentes disciplinas, isto é, quais disciplinas são mais similares a outras e quais são mais diferentes. Esta identificação pode ser feita por meio de técnicas de Aprendizado de Máquina, considerando cada disciplina uma classe em um algoritmo de classificação. Assim, é possível medir também o nível de interdisciplinaridade presente nos vestibulares, embora o conceito de interdisciplinaridade ainda necessite ser mais bem definido [8].

Pesquisas bibliométricas têm tratado desse problema usando, por exemplo, co-citações em *corpora* de artigos científicos. No presente trabalho, o objetivo específico é usar técnicas de Processamento de Línguas Naturais (PLN) para identificar uma combinação de métricas e que possam trazer melhores resultados quando utilizados em uma sequência de etapas (*pipeline*) de processamento de textos de um algoritmo de classificação.

Com a pesquisa, espera-se encontrar superposições lexicais entre disciplinas como as seguintes:

- Português e Literatura (superposição muito forte)
- Matemática e Física
- Química e Biologia

Espera-se também que a classe Interdisciplinar tenha superposição com muitas disciplinas diferentes, e que o classificador tenha poucos acertos para essa classe.

As classes “Inglês”, “Espanhol” e “Francês” não deverão ter quase nenhuma superposição por conta das palavras estrangeiras.

II. METODOLOGIA

A. Construção do Corpus

A fonte de dados utilizada (fonte do *corpus*) foi o site Brasil Escola¹, que fornece simulado de vestibulares para que estudantes do Ensino Médio possam se preparar para os vestibulares das universidades ou para o Exame Nacional do Ensino Médio (ENEM).

Os dados extraídos dessa fonte são perguntas de múltipla escolha classificadas em 40 tópicos associados a disciplinas do Ensino Médio. Para fins de simplificação, a lista de classes desconsiderou subtópicos e foi reduzida para 12 disciplinas conforme lista a seguir:

- Biologia
- Física
- Geografia
- História
- Interdisciplinar
- Língua Estrangeira – Espanhol
- Língua Estrangeira – Francês
- Língua Estrangeira – Inglês
- Literatura
- Matemática
- Português
- Química

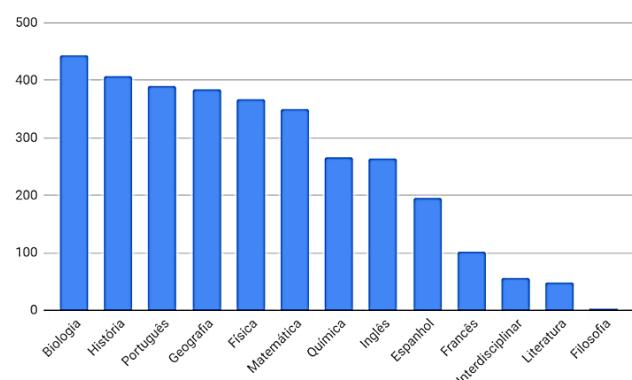


Figura 1 - Distribuição das classes

¹Disponível em:
<https://vestibular.brasilescola.uol.com.br/simulado/questoes/>

A Figura 1 mostra a quantidade de questões por disciplina no *dataset*. É importante notar que:

1. No corpus, as classes estão desbalanceadas. Há apenas 3 questões (0,09% do total) de Filosofia, enquanto há 443 questões de Biologia (13,51% do total). Por este motivo a classe “Filosofia” foi removida do *dataset*;
2. A classe Interdisciplinar é muito vaga e pode não ser útil para o problema de classificação por provavelmente não fornecer meios de separação claros em relação às outras classes;
3. As classes de línguas estrangeiras possuem perguntas com palavras tanto em Português quanto em sua respectiva língua estrangeira, logo é esperado que houvesse uma clara vantagem para reconhecer as palavras estrangeiras que não possuem equivalentes em Português.

B. Pipeline de Processamento de Texto

Foi criado um *pipeline* de processamento de texto composto por etapas de PLN seguidas de um classificador Naive Bayes Multinomial. Este algoritmo tem a capacidade de aprender com os dados de diferentes classes processadas pelas etapas anteriores do *pipeline*. Após um processo de treinamento (aprendizagem supervisionada), o algoritmo está apto a classificar questões de acordo com suas respectivas disciplinas. Outros autores usaram Naive Bayes, mas em *corpora* de textos científicos [7]. Após estabelecer a probabilidade de cada texto pertencer a uma disciplina específica, mediram-se as distâncias entre campos de saber. O grau de interdisciplinaridade foi associado a tais medidas. Neste trabalho optou-se por utilizar as bibliotecas *scikit-learn*² [3] e *Natural Language Toolkit (NLTK)*³ [2] para construir o *pipeline* que tem os seguintes passos e parâmetros:

1. Remover *stopwords* e contar a ocorrência de cada n-grama com o uso do
 - a. Parâmetro: tamanho dos n-gramas a serem considerados, de $n = 1$ até $n = 3$
2. Computar TF-IDF (*term frequency-inverse document frequency*)
 - a. Parâmetro: usar ou não o IDF
 - b. Parâmetro: usar norma L1 (distância de Manhattan) ou normal L2 (distância Euclidiana)
 - c. Parâmetro: suavizar ou não o IDF
 - d. Parâmetro: usar ou não escala logarítmica no TF

²Disponível em: <https://scikit-learn.org/>

³Disponível em: <https://www.nltk.org/>

3. Classificador Naive Bayes Multinomial

- a. Parâmetro: coeficiente de suavização de Laplace, valores 0,1, 0,01 ou 0,001)

A combinação desses parâmetros leva a 144 possíveis *pipelines*. Nesse cenário, a obtenção dos parâmetros do *pipeline* foi realizada por validação cruzada de 5 partições, ou seja: 80% do *dataset* será usado para treinar e 20% será usado para testar o desempenho do classificador. Esse procedimento foi repetido 5 vezes para que todas as partições fossem usadas para teste. A pontuação final do classificador foi a média da acurácia de teste nas 5 partições.

Usando o melhor modelo de classificador encontrado, foram geradas diferentes matrizes de confusão, variando a proporção entre os conjuntos de treinamento e teste. Assim, foi possível verificar o comportamento do *pipeline* e medir as superposições presentes no *dataset*. Também foram extraídos os termos predominantes em cada disciplina para a classificação.

III. RESULTADOS

A. Escolha dos parâmetros do pipeline de processamento de texto

Ao realizar os testes com todas as combinações de parâmetros, foi observado que:

- A normalização L2 se mostrou superior à L1
- O TF com escala sublinear também foi presente em todos os bons classificadores.
- Nos classificadores com maior acurácia, observaram-se n-gramas de tamanho 2 ou 3.
- O parâmetro de suavização de Laplace de 0,01 levou aos melhores resultados.

Para o melhor resultado, obteve-se uma acurácia de 0,86 com os parâmetros: $n=2$, uso do IDF, norma L2, ausência de suavização do IDF, uso de escala logarítmica no TF e suavização de Laplace de 0,01 no Classificador Naive Bayes Multinomial.

B. Análise de matrizes de confusão e dos termos predominantes

Na Figura 2, é mostrada a matriz de confusão quando se utilizou 90% do *dataset* para treinamento e 10% para teste. Já na Figura 3, é mostrada a matriz de confusão em que 10% dos dados foram utilizados para treinamento e 90% para testes.

A partir das matrizes de confusão, observou-se que:

- Há superposição considerável entre as disciplinas de História e Geografia.
- Há uma superposição mínima de Línguas Estrangeiras (Inglês, Francês e Espanhol) com Português, por conta de partes do enunciado que são em Português.

- A classe Interdisciplinar apresentou muitas superposições em diversas disciplinas, apenas as Línguas Estrangeiras e Literatura ficaram com pouca superposição.
- Literatura teve muita superposição com Português, sendo que poucos exemplos foram classificados corretamente.
- Matemática teve superposição considerável com Física.
- Química teve superposição considerável com Biologia.

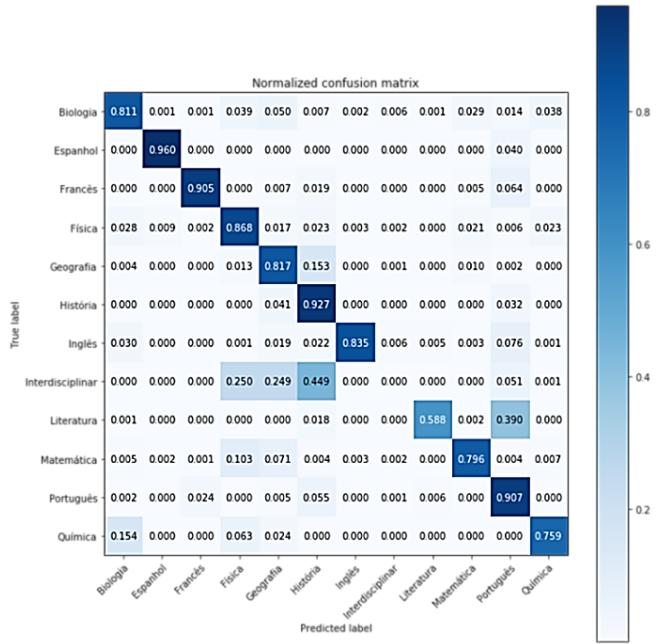


Figura 2: Matriz de confusão para classificação com 90% do dataset para treinamento e 10% para teste.

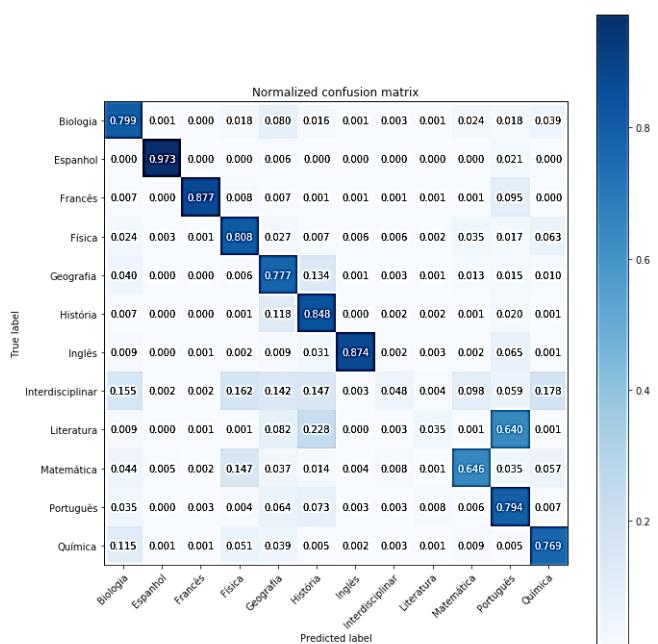


Figura 3: Matriz de confusão para classificação com 10% do dataset para treinamento e 90% para teste.

A taxa de acerto do classificador, em geral, decresceu quando o conjunto de treinamento foi reduzido de 90% para 10% do total. Isto ficou mais evidente para classe Literatura, em que a classificação incorreta como Português cresceu de 0,390 para 0,640. Foram observadas exceções nas categorias Espanhol, Química, Inglês e Interdisciplinar, em que a taxa de acerto aumentou. Como pode ser observado na Figura 1, estas categorias estão entre as que menos possuem questões no dataset. Além disso, tais resultados podem ser relacionados a flutuações devido a diferentes conjuntos de dados selecionados durante o processo de validação cruzada.

QUADRO I: Termos predominantes por disciplina.

Disciplina	Termos predominantes
Biologia	“vírus”, “espécies”, “dna”, “organismos”, “apenas”, “plantas”, “abaixo”, “animais”, “água”, “células”
Espanhol	“al”, “un”, “una”, “con”, “del”, “las”, “los”, “en”, “la”, “el”
Francês	“une”, “un”, “il”, “du”, “des”, “et”, “les”, “est”, “la”, “le”
Física	“força”, “água”, “elétrica”, “constante”, “temperatura”, “energia”, “massa”, “figura”, “velocidade”
Geografia	“país”, “população”, “maior”, “sul”, “região”, “apenas”, “áreas”, “mapa”, “países”, “brasil”
História	“período”, “estados”, “poder”, “estado”, “sobre”, “governo”, “guerra”, “século”, “política”, “brasil”
Inglês	“are”, “on”, “in the”, “that”, “is”, “in”, “and”, “of”, “to”, “the”
Interdisciplinar	“sen cos”, “massa”, “mesma”, “elétron pósitron”, “km2”, “água”, “pósitron”, “cm”, “apenas”, “apenas apenas”
Literatura	“sargento milícias”, “memórias sargento”, “memórias”, “livro”, “obra”, “narrador”, “tempo”, “enredo”, “romance”, “personagens”
Matemática	“nímeros”, “nímero”, “circunferênciia”, “área”, “figura”, “equação”, “então”, “valor”, “igual”
Português	“expressão”, “uso”, “nada”, “pode”, “bem”, “linguagem”, “mundo”, “sentido”, “ser”, “texto”
Química	“gás”, “massa”, “ml”, “oxigênio”, “sódio”, “mol”, “reação”, “ácido”, “água”, “solução”

No Quadro I, são demonstrados os termos predominantes para cada categoria. Pode-se notar que na maioria dos casos os termos são efetivamente ligados aos conteúdos das disciplinas, tais como nos lexemas “vírus” em Biologia e “circunferência” em Matemática. Essa associação não fica bem definida na categoria Interdisciplinar, ocorrendo lexemas associados a diversas áreas, tais como “sen cos” e “massa”. Em algumas categorias foram observados termos relacionados aos enunciados das questões. Por exemplo, em Biologia o termo “abaixo” e, em Física e Matemática o termo “figura”. Outra informação significativa é o fato de que alguns lexemas são frequentes em diversas disciplinas:

- “água” – Biologia, Física, Interdisciplinar, Química
- “massa” – Física, Interdisciplinar, Química
- “Brasil” – História, Geografia

Para as disciplinas de Línguas Estrangeiras, os termos predominantes aparentemente são as *stopwords* de cada língua. Isto é compreensível, uma vez que o *pipeline* não fará a remoção destas palavras (neste trabalho, são considerados apenas *stopwords* em língua portuguesa). Consequentemente, o algoritmo de Aprendizado de Máquina usará as palavras mais significativas de cada língua para diferenciar as classes de Línguas Estrangeiras.

IV. CONCLUSÕES

Os métodos e técnicas empregados obtiveram uma classificação pertinente do léxico por disciplina, com alguma interferência de termos do enunciado das questões. As superposições indicaram que as categorias disciplinares mantêm léxico estável e não afetado por processos de hibridação categorial de campos emergentes da ciência. A categoria Interdisciplinar apresentou as superposições esperadas.

REFERÊNCIAS

- [1] Sardinha, T. B (2004). *Lingüística de Corpus*. Barueri, SP: Editora Manole.
- [2] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc..
- [3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. & Cournapeau, D., (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830
- [4] Leydesdorff, L. & Cozzens, S. (1993). The Delineation of Specialties in Terms of Journals Using the Dynamic Journal Set of the Science Citation Index. *Scientometrics*, 26, 133-154.
- [5] Leydesdorff, L. & Zhou, P. (2009) Nanotechnology as a field of science: its delineation in terms of journals and patents. *Scientometrics*, 70(3), 693-713.
- [6] McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London & New York: Routledge.
- [7] Giannopoulos, T., Foufoulas, I., Stamatogiannis, E., Dimitropoulos, H., Manola, N., Ioannidis, Y. (2014). Discovering and visualizing interdisciplinary content classes in scientific publications. *D-Lib Magazine*, 20(11):4.
- [8] Weingart, P. & Stehr, N. (Eds). (2000). *Practising Interdisciplinarity*. Toronto: University fTorontoPress.

Classificação de subjetividade para a língua portuguesa

Luana Balador Belisário, Luiz Gabriel Ferreira, Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos/SP, Brasil
www.nilc.icmc.usp.br
luana.belisario@usp.br, luizgferreira@usp.br, taspardo@icmc.usp.br

Keywords - análise de sentimentos, classificação de subjetividade, léxico, aprendizado de máquina

I. INTRODUÇÃO

Textos publicados nas redes sociais têm sido uma fonte valiosa de informações para organizações, uma vez que a análise desses textos é uma forma de aprimorar o feedback de produtos e serviços dessas empresas.

Devido ao crescimento do número de usuários em redes sociais e plataformas web, tornou-se possível obter grande volume de dados para esse tipo de análise. Com isso, surgiu a necessidade da criação de ferramentas que pudessem analisar as opiniões dos usuários de forma automática. A área de pesquisa que realiza esse tipo de processamento é a Análise de Sentimentos, também chamada de Mineração de Opiniões.

A análise de subjetividade é uma das primeiras etapas na mineração de opiniões. Nessa tarefa, os documentos de interesse (que podem ser textos completos, sentenças ou mesmo fragmentos menores) são classificados como subjetivos ou objetivos de acordo com sua polaridade: quando classificados como objetivos, assume-se que expressam fatos; quando ditos subjetivos, expressam opiniões e sentimentos (Liu, 2012).

Fatos apontam informações sobre um acontecimento (por exemplo, “Comprei um netbook Philco em nov/2010.”). Opiniões, por sua vez, expressam o ponto de vista de uma ou mais pessoas a respeito de um tema (por exemplo, “Esse livro que li é muito bom, de uma profundidade incrível!”). No exemplo factual, é possível identificar sua objetividade por se tratar de uma notícia e por não ser possível identificar se a notícia é boa ou ruim por princípio. Já no exemplo opinativo, é possível identificar que o autor do comentário gostou no livro lido através das expressões “muito bom” e “incrível”. Essas palavras que denotam evidentemente a opinião e a polaridade dela são chamadas de palavras de sentimento.

Na Tabela 1, pode-se observar algumas outras sentenças classificadas que ocorreram no córpus Computer-BR utilizado pelos autores Moraes et al. (2016). As sentenças subjetivas são ainda subdivididas em polaridade “positiva” e “negativa”, expressando explicitamente sentimento em relação a um notebook, com indicativos como “tô mt feliz” e

“Parece brincadeira de mau gosto”. As sentenças objetivas, apesar de possuírem adjetivos que podem indicar opinião como “boas” e “baum” (bom), não expressam opinião a respeito dos produtos. Cabe destacar ainda a linguagem utilizada, que apresenta abreviações e vocabulário típico do domínio. Essas são considerações que devem ser feitas ao desenvolver métodos para este tipo de classificação.

Tabela 1: Exemplos de sentenças rotuladas do córpus Computer-BR.

Sentença	Polaridade
Sem notebook de novo... Parece brincadeira de mau gosto	Subjetiva/Negativa
Alguem me indica marcas boas de notebook?	Objetiva
Logo um precioso notebook Dell lindíssimo chega aqui em casa tô mt feliz	Subjetiva/Positiva
Esse Not é baum ? Alguem sabe ?	Objetiva

Para a língua portuguesa, há muitos trabalhos na área de análise de sentimentos. Entretanto, no melhor do nosso conhecimento, há apenas uma iniciativa (Moraes et al., 2016) dedicada à tarefa de classificação de subjetividade. Essa tarefa é muito relevante para etapas posteriores de análise textual, sendo responsável por indicar o conteúdo relevante para processamento, ou seja, as sentenças subjetivas, que são de maior interesse em mineração de opiniões. Essa seleção de conteúdo tem potencial para aprimorar os resultados de aplicações relacionadas na área.

Neste artigo, iniciamos por reproduzir os experimentos de Moraes et al. Além disso, estendemos a avaliação dos métodos para outros córpuses, visando avaliar sua robustez. Especificamente, comparamos abordagens para classificação de subjetividade baseada em léxico e baseada em aprendizado de máquina. Mostramos, ao final, que o melhor resultado obtido foi de 77% de acurácia no método baseado em aprendizado de máquina e de 75,2% no método baseado em léxico. Além disso, evidenciamos que fatores como tamanho do córpus, balanceamento entre sentenças de diferentes polaridades e diferentes técnicas de pré-processamento aplicadas nos córpuses influenciam significativamente os resultados.

Na seção seguinte, fazemos uma breve revisão literária. Em seguida, apresentamos os córpus utilizados neste trabalho. Nas Seções IV e V, descrevemos os métodos avaliados. Os resultados principais são relatados na Seção VI. Por fim, fazemos algumas considerações finais na Seção VII.

II. REVISÃO BIBLIOGRÁFICA

Moraes et al. (2016) propõem métodos para efetivamente classificar a subjetividade de textos em nível sentencial para a língua portuguesa. Os autores argumentam sobre a importância dessa área de pesquisa para empresas no que diz respeito a um feedback mais completo e eficiente de produtos ou serviços oferecidos por elas.

Para testar os métodos, os autores do artigo criaram um córpus com 2.317 tweets sobre a área de tecnologia. Esse córpus - chamado Computer-BR - foi anotado manualmente e submetido a um pré-processamento para aumentar a eficiência dos métodos aplicados. Detalhes sobre o córpus e o pré-processamento estarão na próxima seção.

Os métodos abordados e testados pelos autores são baseados em léxico e em aprendizado de máquina. Os melhores resultados com os métodos baseados em léxico atingiram 64% de medida-f, enquanto os resultados dos métodos baseados em aprendizado de máquina chegaram a 75%. Os métodos, implementados neste artigo, são descritos posteriormente.

III. CÓRPUS UTILIZADOS

Neste trabalho, para o teste dos dois métodos, foram usados três córpus compostos por revisões de produtos ou serviços de três domínios distintos.

Como especificado na seção anterior, o córpus Computer-BR contém 2.317 sentenças com mensagens de usuários sobre a área de tecnologia. É interessante observar que esse córpus apresenta desbalanceamento das classes, com quantidade superior de sentenças objetivas.

Além do domínio de tecnologia no córpus Computer-BR, utilizamos um córpus de resenhas de livros, que reúne resenhas curtas de leitores retiradas do córpus ReLi (Freitas et al., 2012), do site de compras Amazon e da rede social Skoob, com 270 sentenças divididas igualmente entre fatos e opiniões. Essas sentenças foram classificadas manualmente.

Por fim, utilizamos um córpus de produtos eletrônicos com 230 sentenças que foram retiradas do conhecido córpus Buscapé (Hartmann et al., 2014) e anotadas posteriormente de forma manual. Assim como o Computer-BR, esse córpus apresenta desbalanceamento, com cerca de 70% de sentenças classificadas como subjetivas.

Os três córpus utilizados são formados por sentenças retiradas da web. Logo, o grande número de abreviações, erros de ortografia e jargões específicos caracterizam esse meio e comprometem o desempenho dos métodos automáticos de classificação. Para amenizar essa dificuldade foram realizadas algumas etapas de pré-processamento.

Primeiramente, as sentenças passaram por um processo de normalização da linguagem com o auxílio do sistema de normalização textual *enelvo*¹ (Bertaglia, 2017). Após isso, as *stopwords* (conforme constam no pacote *Natural Language Toolkit*² - NLTK), a pontuação, os números e os demais caracteres especiais foram removidos. Por fim, os termos foram reduzidos às suas formas canônicas com o auxílio de um lematizador³ desenvolvido no NILC.

A seguir, apresentamos os métodos implementados e avaliados neste artigo.

IV. MÉTODO BASEADO EM LÉXICO

O artigo de Moraes et al. abordou três heurísticas para classificação de subjetividade. As três consistem em pesquisar palavra por palavra da sentença de interesse e verificar a subjetividade e polaridade de cada uma. A Heurística 1 se traduz em somar a polaridade de todas as palavras subjetivas e o resultado já é a classificação; a Heurística 2 classifica uma sentença em subjetiva se ela possui um valor mínimo de palavras subjetivas, e então é feito o cálculo para saber sua polaridade; já na Heurística 3, a sentença é classificada como subjetiva se ela possui uma proporção mínima de palavras subjetivas, ou seja, quantas palavras do total de palavras da sentença são subjetivas. A Heurística 1 foi a que produziu os melhores resultados no artigo de Moraes et al.

A Heurística 1 depende de um léxico de palavras subjetivas (ou palavras de sentimento) pré-classificadas para analisar o texto em nível sentencial. Essas palavras foram associadas com valor 1 se eram positivas, -1 para negativas e 0 para neutras. Basicamente, a heurística consiste em determinar subjetividade da sentença somando as polaridades das palavras que a compõem. A fórmula está representada em (1), onde n é o número de palavras da sentença, ‘term i’ representa cada palavra da sentença e subjetividade(sent) é a subjetividade em nível sentencial:

$$\text{subjetividade}(\text{sent}) = \sum_{i=1}^n \text{polaridade}(\text{term } i) \quad (1)$$

Se a subjetividade(sent) assumir um valor diferente de 0, a sentença é considerada subjetiva, sendo valor positivo (maior que 0) uma sentença “subjetiva positiva” e valor negativo (menor que 0) uma sentença “subjetiva negativa”. Caso o valor seja zero, a sentença é considerada “objetiva” ou “neutra”. Em razão da simplicidade do método, não há tratamento de negação, ironia e advérbios, cujas funções seriam intensificar, neutralizar ou até mudar a orientação das palavras de sentimento.

V. MÉTODO BASEADO EM APRENDIZADO DE MÁQUINA

Para as técnicas baseadas em aprendizado de máquina, foram testados dois algoritmos de classificação, seguindo-se a proposta de Moraes et al.. Ambos utilizam o modelo *bag*

¹ <https://github.com/tfcbertaglia/enelvo>

² <https://www.nltk.org/>

³ <http://conteudo.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar>

of words, que seleciona cada palavra contida na sentença como um atributo distinto. As palavras que poderão ser utilizadas como atributos foram limitadas com base em suas relevâncias para melhor desempenho. A seguir são descritos os métodos de seleção das palavras.

A etapa inicial consiste em quantificar a relevância das palavras em cada classe. Para isso foram utilizadas duas métricas, sendo a primeira a frequência da palavra na classe, como mostrado em (2), onde w_k indica a palavra e c_j indica a classe.

$$\text{freq} = P(w_k | c_j) \quad (2)$$

A segunda métrica utilizou o *Comprehensive Measurement Feature Selection* (CMFS), que busca calcular a relevância da palavra na classe considerando as suas ocorrências nas outras classes. Para isso, realiza-se o produto da probabilidade da palavra pertencer à classe pela sua frequência na classe:

$$\text{CMFS}(w_k, c_j) = P(w_k | c_j) \cdot P(c_j | w_k) \quad (3)$$

Com as métricas descritas, é gerada uma lista para cada classe com as palavras mais relevantes. A partir dessas listas, é gerada uma única lista com as palavras que serão utilizadas como atributos. De cada lista, são selecionadas as n (n menor que 100) palavras mais relevantes e agrupadas de duas maneiras distintas. A primeira simplesmente junta todas as palavras, e a segunda junta as palavras, mas exclui as que aparecem nas duas listas. Em ambos os casos, caso se obtenha o mesmo valor para mais de uma palavra na última posição, é utilizada apenas uma.

Os algoritmos utilizados são o *Naive Bayes* (NB) e o *Sequential Minimal Optimization* (SMO). O primeiro é baseado no teorema de Bayes e calcula previamente as probabilidades dos termos nas classes. Para realizar a classificação, assume-se que os atributos sejam condicionalmente independentes e escolhe-se a classe que tenha a maior probabilidade. O SMO é uma otimização matemática para treinar as “máquinas de vetores de suporte”. O objetivo do método é dividir o espaço descrito pelos atributos em duas regiões, cada uma pertencente a uma classe. Na classificação, analisa-se em qual região do espaço o conjunto de termos da sentença se situa.

VI. RESULTADOS E DISCUSSÃO

Os resultados alcançados para as classes “subjetivo” e “objetivo” são mostrados nas Tabelas 2 e 3, respectivamente. São exibidos valores para as medidas clássicas de precisão, cobertura e medida-f, além da acurácia geral. Exibimos apenas os melhores resultados conseguidos.

Como se pode notar nas tabelas, para os métodos baseados em léxico, fizemos experimentos com os léxicos de sentimento WordnetAffectBR (Pasqualotti e Vieira, 2008) e Sentilex-PT (Carvalho e Silva, 2015), amplamente conhecidos na área.

Tabela 2: Melhores resultados em abordagens baseadas em léxico e aprendizado de máquina para cada córpus para sentenças subjetivas.

Córpus	Computer-BR		Resenha de Livros		Produtos Eletrônicos		
	Método	NB	WordnetAffect	SMO	WordnetAffect	SMO	Sentilex
Precisão	0.580	0.524	0.817	0.736	0.906	0.912	
Cobertura	0.630	0.189	0.703	0.293	0.694	0.768	
Medida-f	0.601	0.278	0.744	0.419	0.782	0.834	
Acurácia	0.771	0.729	0.770	0.600	0.700	0.752	

Tabela 3: Melhores resultados em abordagens baseadas em léxico e AM para cada córpus para sentenças objetivas.

Córpus	Computer-BR		Resenha de Livros		Produtos Eletrônicos		
	Método	NB	WordnetAffect	SMO	WordnetAffect	SMO	Sentilex
Precisão	0.853	0.751	0.734	0.567	0.373	0.405	
Cobertura	0.825	0.934	0.836	0.898	0.729	0.682	
Medida-f	0.838	0.833	0.770	0.695	0.475	0.501	
Acurácia	0.771	0.729	0.770	0.600	0.700	0.752	

Analizando-se as tabelas, é possível inferir que os resultados para o córpus Computer-BR ficaram próximos dos descritos por Moraes et al. (2016), com variações decorrentes, principalmente, das técnicas diferentes de pré-processamento. Para a acurácia, por exemplo, os melhores valores originalmente eram 0.78 para técnicas de aprendizado de máquina e 0.74 para abordagens baseadas em léxico, enquanto os valores obtidos aqui foram 0.77 e 0.73, respectivamente.

É perceptível que o córpus de produtos eletrônicos apresentou piores resultados em relação ao córpus de resenha de livros, principalmente ao utilizar métodos baseados em aprendizado de máquina, evidenciada na classificação de sentenças objetivas. Essa piora pode ser justificada pelo desbalanceamento do córpus, fato que ocorre também no Computer-BR. Além disso, ao comparar com o Computer-BR, que possui aproximadamente dez vezes o número de sentenças, nota-se a influência do tamanho do córpus para o aprendizado de máquina.

Desconsiderando o desbalanceamento, os erros cometidos pelos algoritmos de aprendizado de máquina decorrem, principalmente, da falta de informações sobre alguns termos. Esse fato ocorre com mais frequência nos córpus de avaliação de produtos eletrônicos e Computer-BR, onde a linguagem é mais informal, gerando baixa frequência para diversos termos que possuem significados similares. Como exemplo, temos a sentença retirada do Computer-BR a seguir: “Notebook da Positivo é péssimo, credo! Melhor notebook é Dell”. Os termos “péssimo”, “credo” e “melhor” caracterizam a polaridade da sentença, tornando simples uma classificação que tenha informação semântica a respeito. Entretanto, os termos aparecem poucas vezes no córpus e acabam não sendo selecionados como atributos, o que deixa termos que não possuem polaridade, como “notebook”, “dell” e “positivo”, como possíveis atributos em uma classificação.

Outra limitação do método é por conta do tratamento dos termos fora de contexto. Apesar de ser funcional em casos onde é possível extrair termos que possuem polaridade nítida, em outros casos a polaridade pode ser definida pela

maneira como a sentença foi construída, e o método não é capaz de distinguir.

Em relação ao método baseado em léxico, por se fundamentar simplesmente na busca e contagem das palavras de sentimento, não se tratam com eficiência possíveis desvios na polaridade das sentenças, como o tratamento de sarcasmo e ironia, advérbios de negação e intensidade e desambiguação de palavras. Um exemplo dos problemas citados ocorreria ao classificar a sentença “O processador desse notebook é pouco eficiente”. Nesse caso, o algoritmo encontraria a única palavra de sentimento que seria “eficiente” e classificaria como subjetiva e positiva a sentença, porém, sabe-se que um processador ser pouco eficiente é uma característica negativa para o produto. O advérbio de intensidade “pouco” alterou a polaridade da palavra de sentimento “eficiente”, mas isso não foi detectado pelo método.

Por fim, destaca-se a grande variação de resultados entre os códigos diferentes, corroborando o que se observa na literatura da área. Em particular, para o método baseado em léxico, o léxico WordnetAffectBR foi mais apropriado em dois dos três domínios. Em aprendizado de máquina, o método SMO se destacou em relação ao NB.

A seguir, apresentamos algumas considerações finais.

VII. CONSIDERAÇÕES FINAIS

O trabalho descrito apresentou duas abordagens para a classificação automática de sentenças retiradas da web, tendo como objetivo a distinção entre sentenças subjetivas e objetivas. Os métodos apresentados constituem uma reprodução dos utilizados originalmente por Moraes et al. (2016), estendidos para outros códigos. Apesar dos resultados promissores, ainda há muito a se fazer.

Em aprendizado de máquina, visa-se ainda investigar e desenvolver métodos com o uso de *word embeddings* (Mikolov et al., 2013). Na abordagem baseada em léxico, o próximo passo será implementar o método de classificação de subjetividade baseado em grafos abordado por Vilarinho et al (2018). Há, também, desafios relacionados às classes das sentenças. Por exemplo, é sabido que sentenças objetivas podem carregar opinião. Entretanto, esse fenômeno tem sido deixado de lado nesta pesquisa, podendo ser investigado em trabalhos futuros.

Por fim, observa-se que este trabalho integra o projeto maior OPINANDO (*Opinion Mining for Portuguese: Concept-based Approaches and Beyond*)⁴, em cuja página encontram-se vários dos recursos e ferramentas citados aqui.

AGRADECIMENTOS

À Fundação de Amparo à Pesquisa do Estado de São Paulo (processo nro 2018/11479-9) e à Pró-Reitoria de Pesquisa da USP pelo apoio a este projeto.

REFERÊNCIAS

- Bertaglia, T.F.P. (2017). Normalização textual de conteúdo gerado por usuário. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 160p.
- Carvalho, P. e Silva, M.J. (2015). Sentilex-PT: Principais Características e Potencialidades. Oslo Studies in Language, Vol. 7, N. 1, pp. 425-438.
- Freitas, C.; Motta, E.; Milidiú, R.; Cesar, J. (2012). Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. In the Proceedings of the XI Encontro de Linguística de Corpus (ELC), pp. 1-12.
- Hartmann, N. S.; Avanço, L.; Balage, P. P.; Duran, M. S.; Nunes, M. G. V.; Pardo, T.; Aluísio, S. (2014). A Large Opinion Corpus in Portuguese - Tackling Out-Of-Vocabulary Words. In the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), pp. 3865-3871.
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. 168p.
- Mikolov, T.; Corrado, G.; Chen, K.; Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv: 1301.3781, pp. 1-12.
- Moraes, S.M.W.; Santos, A.L.L.; Redecker, M.; Machado, R.M.; Meneguzzi, F.R. (2016). Comparing Approaches to Subjectivity Classification: A Study on Portuguese Tweets. In the Proceedings of the International Conference on the Computational Processing of the Portuguese Language (PROPOR), pp. 86-94.
- Pasqualotti, P.R. e Vieira, R. (2008). WordnetAffectBR: uma base lexical de palavras de emoções para a língua portuguesa. Revista Novas Tecnologias na Educação, Vol. 6, N. 2, pp. 1-10.
- Vilarinho, G. N.; Ruiz, E. E. S. (2018). Global centrality measures in word graphs for Twitter sentiment analysis. In the Proceedings of the 7th Brazilian Conference on Intelligent Systems (BRACIS), pp. 55-60.

⁴ <https://sites.google.com/icmc.usp.br/opinando/>

Reconhecimento de posicionamentos de natureza moral em textos

Wesley Ramos dos Santos

*School of Arts, Sciences and Humanities
University of São Paulo
São Paulo, Brazil
wesley.ramos.santos@usp.br*

Ivandré Paraboni

*School of Arts, Sciences and Humanities
University of São Paulo
São Paulo, Brazil
ivandre@usp.br*

I. INTRODUÇÃO

A tarefa computacional de análise de sentimentos (AS) representa um conjunto de técnicas - frequentemente fazendo uso de aprendizagem de máquina (AM) - para identificação automática de opiniões, emoções e outros tipos de posicionamentos expressos em língua natural [1], [2]. Sistemas que implementam técnicas de AS - frequentemente com base em textos provenientes de redes sociais e afins - possuem uma ampla gama de aplicações práticas, como avaliação de satisfação e atendimento a consumidores, análise de tendências (políticas, sociais, de mercado, etc.) e construção de perfis de usuários.

De modo geral, pode-se distinguir pelo menos dois tipos de AS: a mineração de opiniões [1], [2] e o reconhecimento de posicionamentos [3]–[9]. A mineração de opiniões é a mais tradicional modalidade de AS. A tarefa computacional neste caso consiste em detectar o alvo de uma opinião, e/ou a polaridade (positiva, negativa, neutra etc.) do sentimento expresso em relação a este alvo [2]. Muitos estudos deste tipo enfocam apenas esta segunda tarefa, que na área de Processamento de Línguas Naturais (PLN) frequentemente acaba sendo chamada também de ‘Análise de Sentimentos’. Em contrapartida, o reconhecimento de posicionamentos consiste em determinar se o autor do texto possui uma atitude (ou posicionamento) favorável, contrária ou neutra em relação ao um alvo de interesse [9].

A distinção entre sentimento e posicionamento é motivada pela observação de que um sentimento, seja ele positivo ou negativo, pode refletir um posicionamento favorável ou desfavorável em relação ao alvo [6]. Por exemplo, dado o alvo ‘Donald Trump’, uma frase como ‘*Jeb Bush é o único candidato mentalmente estável nesta eleição*’ expressa um sentimento positivo (e que seria detectado como tal por sistemas de AS tradicionais), porém reflete uma posição desfavorável em relação ao alvo (Trump) [9].

Este artigo descreve um projeto de pesquisa em nível de iniciação científica atualmente em andamento, e que trata do problema do reconhecimento automático de posicionamentos de natureza moral (tratando de questões como legalização do aborto, maioria penal etc.) a partir de textos em português.

II. TRABALHOS RELACIONADOS

Nesta seção são relacionados alguns dos estudos recentes na área de reconhecimento de posicionamentos em texto para a língua inglesa. Em especial, considera-se a competição SemEval-2016 [6] e os dois principais modelos dela resultantes que, juntamente com o trabalho em [9], são exemplos representativos do atual estado da arte.

A competição SemEval-2016 reuniu - na trilha de modelos supervisionados - 19 sistemas participantes engajados na tarefa de reconhecimento de posicionamentos em *tweets* da língua inglesa. O córpus utilizado para fins de treinamento, descrito em detalhes em [10], contém 2914 *tweets* com posicionamentos sobre cinco alvos (ateísmo, mudanças climáticas, movimento feminista, Hillary Clinton e legalização do aborto). O córpus contém em média 583 instâncias por alvo, mas o conjunto é desbalanceado. Em média há 25,8% de posicionamentos favoráveis e 47,9% desfavoráveis. O conjunto de teste, com 1249 instâncias, é ainda mais desbalanceado, com 24,3% de instâncias favoráveis e 57,3% de instâncias desfavoráveis. O córpus SemEval-2016 e o esquema de anotação empregado serão tomados por base para a proposta de construção de novos conjuntos de dados deste tipo para o idioma português, a serem discutidos nas próximas seções.

O estudo em [3] é um dos pioneiros no reconhecimento computacional de posicionamentos a partir de texto, apresentando uma análise de um córpus de 4873 postagens em fóruns de debate online. O conjunto de dados considerado cobre 14 tópicos de discussão, que vão desde temas de entretenimento a questões ideológicas. Posicionamentos favoráveis e desfavoráveis são reconhecidos com acurácia de até 69%, superando um modelo de *baseline* do tipo unígrafo que obteve acurácia de até 60%.

O reconhecimento de posicionamentos em debates é também o foco do estudo em [5]. Neste caso, entretanto, privilegiou-se a investigação da questão de como o desempenho de um classificador de posicionamentos varia em relação ao volume e qualidade dos dados de treinamento, quanto à complexidade do modelo subjacente, à riqueza do conjunto de características de aprendizagem e ao uso de restrições extralingüísticas em uma ampla gama de cenários. Embora não apresente modelos computacionais radicalmente diferentes de

outros estudos da época, os experimentos realizados deixam uma série de contribuições sobre como construir modelos deste tipo, e sobre quais os tipos de conhecimento a serem considerados.

O estudo em [7] apresenta o modelo de melhor desempenho global na competição SemEval-2016 [6]. A proposta faz uso de uma rede neural recorrente com características aprendidas por supervisão distante a partir de uma extensa base de dados externa. A seguir, são treinados modelos de *embeddings* de palavras e *phrases* usando o método Word2Vec skip-gram [11]. Este conjunto de características é então utilizado para aprendizagem de representações sentenciais por meio de uma tarefa auxiliar de predição de *hashtags*. Finalmente, os vetores de sentenças são otimizados para reconhecimento de posicionamentos com base nos exemplos rotulados do córpus de treinamento.

Também no contexto da competição SemEval-2016, o estudo em [8] apresenta uma abordagem baseada em redes neurais convolutivas que, ao invés de simplesmente predizer quando a acurácia de validação vai atingir seu limite máximo, utiliza um esquema de votação e outras melhorias secundárias. O modelo é treinado individualmente para cada um dos cinco alvos do córpus SemEval-2016 e obtém o segundo melhor resultado global da competição.

O estudo em [4] explora o uso de conhecimento de mundo - na forma de regras sobre amizades e inimizades políticas - para aprimorar a tarefa de reconhecimento de posicionamentos políticos no córpus SemEval-2016. Para este fim específico, propõe-se um modelo enriquecido com uma série de características de aprendizagem relativas a cada alvo, o qual apresenta resultados superiores aos obtidos pelos sistemas participantes originais da competição.

Finalmente, o estudo em [9] apresenta uma avaliação *post-hoc* da tarefa de classificação de posicionamentos SemEval-2016, propondo um modelo muito mais simples do que o do vencedor da competição em [7], e com resultados superiores. O modelo proposto faz uso de SVM com núcleo linear e um conjunto de características computadas a partir dos dados de treinamento, como n-gramas de palavras e caracteres e *word embeddings* computados a partir de um conjunto de dados adicional e não rotulado.

III. EXPERIMENTOS

Ao longo do desenvolvimento do projeto, foram executados uma série de experimentos enfocando dois problemas distintos: (i) o reconhecimento da existência de posicionamento e (ii) o reconhecimento de polaridade. Em (i), o objetivo foi o desenvolvimento de modelos capazes de identificar a existência de posicionamento acerca de um determinado tópico. Nesta tarefa, não há diferenciação entre opiniões favoráveis ou contrárias, mas apenas entre textos que manifestam e que não manifestam opinião sobre o tópico. Em (ii), partindo-se do pressuposto de que os textos que de fato expressam um posicionamento já são conhecidos, o objetivo foi desenvolver modelos capazes de classificar posicionamentos em positivos ou negativos, ou seja, descobrir se o autor de

um texto se coloca como favorável ou contrário ao tópico em questão.

Nas seções a seguir, descrevemos o conjunto de dados utilizado, e os dois experimentos individualmente.

A. Córpus M.Twitter

No contexto do projeto principal ao qual a presente proposta está vinculada, um conjunto de *tweets* brasileiros foi filtrado com uso de cinco palavras-chave, a saber: ‘aborto’, ‘maconha’, ‘maioridade’, ‘ pena de morte’ e ‘cotas raciais’. O objetivo desta tarefa foi o de obter um conjunto de textos potencialmente útil para o reconhecimento automático de posicionamentos sobre temas de natureza moral como os suscitados por estas palavras-chave.

Após a remoção de mensagens duplicadas, em idiomas estrangeiros ou com número excessivo de palavras não encontradas em um dicionário de português [12], chegou-se a um conjunto de textos dividido em cinco categorias, aqui denominado córpus M.Twitter. A distribuição deste córpus é ilustrada na tabela I.

Tabela I
AS CINCO CATEGORIAS TEXTUAIS DO CÓRPUIS M.TWITTER

Palavras-chave	<i>tweets</i>
aborto	11674
cotas raciais	6399
maconha	12201
maioridade	7637
pena de morte	3692

É importante destacar, entretanto, que estas categorias textuais correspondem simplesmente a conjuntos de *tweets* que contêm as palavras-chave consideradas, mas estes textos não necessariamente discutem os temas pretendidos. Por exemplo, uma mensagem como ‘pena de morte para quem coloca uva passa no arroz’ não pode ser interpretada como uma discussão real sobre o tema da pena de morte. Assim, este conjunto preliminar de textos ainda depende de tratamento para que possa ser aproveitado em um estudo sobre o reconhecimento de posicionamentos sobre temas de natureza moral propriamente ditos.

O conjunto de *tweets* foi anotado manualmente até obter um mínimo de 300 instâncias de posicionamentos favoráveis e contrários a cada um dos tópicos em questão, ou até esgotar-se o conjunto de dados daquele tópico. Detalhes adicionais deste córpus, tal qual empregado em cada um dos experimentos a seguir, são discutidos nas próximas seções.

B. Experimento 1: reconhecimento de posicionamento

1) *Visão geral:* O primeiro experimento realizado objetivou o reconhecimento de posicionamentos em texto, modelado como um problema de classificação binária no qual a primeira classe contém todos os textos que não apresentam nenhum posicionamento sobre o tópico, e a segunda classe contém todas as instâncias que se posicionam de forma positiva ou negativa com relação ao tópico. A suposição neste experimento é a de que, embora os textos expressem opiniões opostas,

ainda assim seria possível detectar características comuns aos textos que de alguma forma se posicionam, positiva ou negativamente, de modo a distingui-los de textos que não apresentam posicionamento acerca de um tópico específico.

2) *Dados:* A tabela II apresenta a distribuição do córpus M.Twitter entre as duas classes consideradas (com posicionamento e sem posicionamento) para cada tópico analisado.

Tabela II
QUANTIDADE DE INSTÂNCIAS EM CADA CLASSE

Tópico	com posicio.	sem posicio.
aberto	624	2570
cotas raciais	604	2596
maconha	516	1482
maioridade	483	1433
pena de morte	1045	1518

O córpus foi dividido de forma estratificada em 80% para treino e 20% para teste.

3) *Modelos:* Com base nestes dados, foram comparados três métodos de classificação textual para reconhecimento de posicionamento.

- *BoW-LogReg.* Modelo de contagem *bag-of-words* com algoritmo de regressão logística.
- *NGram-NB.* Modelo de n-gramas de caracteres no intervalo 3 até 16 e classificador *Naive Bayes*.
- *word2vec-MLP.* Modelo baseado em representação distribuída *word2vec* [13] com arquitetura CBOW. O método de aprendizado utilizado neste modelo é a rede *MLP*, com 3 camadas escondidas, cada uma com 150 neurônios.

Os modelos de *embeddings* foram treinados a partir do próprio córpus. Em um experimento piloto, os modelos MLP foram considerados em diferentes configurações de camadas e quantidades de neurônios, mas apenas a melhor combinação será tratada a seguir.

4) *Resultados:* A tabela III sumariza os resultados obtidos para o experimento 1 III-B. Além dos três modelos descritos, é apresentado o resultado de um *baseline* de classe majoritária para cada tópico abordado. A colunas *neu*, *opn* e *avg* representam os valores de medida F, respectivamente, das classes de textos neutra, com opinião e a média ponderada.

Conforme apresentado na Tabela II, existe um desequilíbrio significativo no número de instâncias que são classificadas como neutras e as que expressam alguma opinião. Desta forma, o modelo de *baseline* conseguiu resultados relativamente altos escolhendo sempre a classe majoritária. Os dois modelos que obtiveram o maior destaque foram o de n-gramas de caracteres utilizando um classificador *Naive Bayes* (NGram-NB) e o modelo que utiliza representação distribuída do tipo *word2vec* com um classificador *multilayer perceptron* (*word2vec-MLP*). Ambos obtiveram um destaque sobre o modelo de *baseline* tanto na média de medida F (coluna *avg*), quanto na questão de obter resultados mais equilibrados entre as duas classes mesmo que exista um disparidade entre a quantidade de instâncias.

Tanto no modelo NGram-NB quanto no word2vec-MLP existe a necessidade de processamento adicional que torna

ambos os métodos mais custosos, criação de n-gramas de caracteres e o uso de *word embeddings*, respectivamente. Contudo, esse processamento adicional de fato se mostrou eficaz na tarefa de identificar a existência de posicionamento em que a maioria (com exceção do tópico ‘maconha’) obtiveram um desempenho acima de 0.70 de medida F ponderada.

A relativa vantagem que o modelo baseado em n-gramas obteve pode ser explicada pela sua capacidade de analisar sequências compostas por um conjunto de palavras que, juntas, podem trazer mais significado a análise se comparado a um modelo de contagem de palavras *bag-of-words*. Além disso, o modelo baseado em representação distribuída também obteve resultados bem próximos do desempenho do modelo com n-gramas. Novamente, existe uma vantagem no uso de representação distribuída no que diz respeito a conservação dos significados sintáticos e semânticos, o que explicaria essa vantagem com relação ao modelo de contagem.

C. Experimento 2: detecção de polaridade

1) *Visão geral:* O segundo experimento realizado objetivou a detecção de posicionamentos favoráveis e desfavoráveis a um determinado tópico. Neste caso foi utilizado um modelo de classificação binária em que a primeira classe contém todos os textos que apresentam um posicionamento negativo sobre o tópico, e a segunda classe contém todas as instâncias que se posicionam de forma positiva com relação ao tópico. As instâncias que não apresentam posicionamento não fazem parte do experimento. A suposição neste experimento é a de que, como os textos expressam opiniões opostas, é possível detectar características específicas destes posicionamentos, e assim distinguir uma classe da outra.

2) *Dados:* A tabela IV apresenta a distribuição do córpus M.Twitter entre as duas classes consideradas (com posicionamento e sem posicionamento) para cada tópico analisado.

Tabela IV
QUANTIDADE DE INSTÂNCIAS EM CADA CLASSE

Tópico	contra	a favor
aberto	384	240
cotas raciais	364	240
maconha	181	335
maioridade	240	243
pena de morte	244	801

O córpus foi dividido em 80% para treino e 20% para testes.

3) *Modelos:* Foram comparados três métodos de classificação textual para detecção de polaridade do autor.

- *BoW-LogReg.* Modelo de contagem *bag-of-words* com algoritmo de regressão logística.
- *NGram-NB.* Modelo com representação de n-gramas de caracteres no intervalo 3 até 16 e classificador *Naive Bayes*.
- *NGram-MLP.* Modelo com representação de n-gramas de caracteres com range de 3 até 16. O método de aprendizado utilizado neste modelo é a rede *MLP*, com 3 camadas escondidas, cada uma com 150 neurônios.

Tabela III
RESULTADOS DE MEDIDA F PONDERADA PARA O EXPERIMENTO 1

Tópico	baseline			BoW-LogReg			NGram-NB			word2vec-MLP		
	opn	neu	avg	opn	neu	avg	opn	neu	avg	opn	neu	avg
aborto	0.00	0.89	0.71	0.09	0.89	0.72	0.53	0.82	0.76	0.41	0.85	0.76
cotas	0.00	0.89	0.70	0.01	0.88	0.71	0.41	0.80	0.72	0.33	0.83	0.73
maconha	0.84	0.00	0.61	0.16	0.84	0.65	0.50	0.72	0.66	0.34	0.77	0.65
maioridade	0.00	0.85	0.64	0.30	0.86	0.72	0.59	0.80	0.75	0.50	0.83	0.74
pena de morte	0.74	0.00	0.44	0.45	0.74	0.62	0.67	0.77	0.73	0.63	0.75	0.70

Tabela V
RESULTADOS DE MEDIDA F PARA O EXPERIMENTO 2

Tópico	baseline			BoW-LogReg			NGram-NB			NGram-MLP		
	con	fav	avg	con	fav	avg	con	fav	avg	con	fav	avg
aborto	0.72	0.00	0.41	0.67	0.49	0.59	0.77	0.63	0.71	0.76	0.61	0.70
cotas	0.78	0.00	0.49	0.71	0.31	0.56	0.81	0.67	0.76	0.48	0.77	0.68
maconha	0.00	0.81	0.55	0.39	0.74	0.63	0.51	0.71	0.65	0.48	0.77	0.68
maioridade	0.00	0.63	0.29	0.74	0.68	0.71	0.75	0.68	0.72	0.70	0.61	0.66
pena de morte	0.00	0.80	0.53	0.32	0.80	0.64	0.73	0.84	0.80	0.57	0.83	0.74

4) *Resultados:* A tabela V sumariza os resultados obtidos para o experimento 2 III-C. Além dos três modelos descritos, é apresentado o resultado de um *baseline* de classe majoritária para cada tópico abordado. A colunas *con*, *fav* e *avg* representam os valores de medida F, respectivamente, das classes de textos com opiniões contrárias, opiniões favoráveis e a média ponderada.

Diferentemente do experimento anterior, neste caso ambas as classes possuem uma distribuição mais equilibrada. Embora os 3 modelos analisados tenham obtido um desempenho relativamente melhor do que o *baseline* de classe majoritária, houve um destaque significativo para os dois modelos baseados em n-gramas de caracteres. Novamente, o modelo que utiliza o classificador *Naive Bayes* obteve uma vantagem consideravelmente maior contra o segundo melhor, que utiliza n-gramas e rede *MLP*.

Se por um lado as classes consideradas nesse experimento de fato manifestavam opiniões opostas (enquanto no experimento anterior opiniões contrárias e favoráveis eram agrupadas juntas), por outro lado essa nova forma de categorização não refletiu uma melhora nos resultados, os quais tiveram desempenho muito próximos em ambas as tarefas. Desta forma, mesmo considerando-se que alguns resultados para esse experimento tenham ficado relativamente mais altos, não há indícios de que a tarefa de detectar a polaridade de textos opinativos seja particularmente mais fácil do que detectar a existência de posicionamento ou não.

IV. CONSIDERAÇÕES

Este trabalho apresentou resultados parciais da tarefa de reconhecimento de posicionamentos morais a partir de textos do Twitter brasileiro. Como trabalho futuro, pretende-se expandir o córpus atual e fazer uso de modelos de aprendizado de máquina mais sofisticados - como os baseados em aprendizado profundo - aplicados às tarefas de detecção de posicionamentos e polaridade.

Além da aplicação de técnica mais robustas para as tarefas aqui apresentadas, outros problemas como o desequilíbrio no número de instâncias das classes binárias pode ser tratado a partir da aplicação de técnicas de *resampling* que melhoraram essa distribuição.

AGRADECIMENTOS

Este estudo contou com apoio FAPESP 2017/06828-1 e 2016/14223-0, e da Universidade de São Paulo.

REFERÊNCIAS

- [1] B. Liu, *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.
- [2] M. Tsytarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.
- [3] P. Anand, M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani, and M. Minor, "Cats rule and dogs drool!: Classifying stance in online debate," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011)*. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 1–9.
- [4] M. Lai, D. I. H. Farias, V. Patti, and P. Rosso, "Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets," in *Mexican International Conference on Artificial Intelligence (MICAI-2016). Lecture Notes in Computer Science, vol 10061*. Springer, 2016.
- [5] K. S. Hasan and V. Ng, "Stance classification of ideological debates: Data, models, features, and constraints," in *Proceedings of the International Joint Conference on Natural Language Processing*. Nagoya, Japan: Association for Computational Linguistics, 2013, pp. 1348–1356.
- [6] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 Task 6: Detecting Stance in Tweets," in *Proceedings of the International Workshop on Semantic Evaluation*, San Diego, California, USA, 2016.
- [7] G. Zarrella and A. Marsh, "MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection," in *Proceedings of the International Workshop on Semantic Evaluation*, San Diego, California, USA, 2016.
- [8] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang, "pkudblab at SemEval-2016 Task 6: A Specific Convolutional Neural Network System for Effective Stance Detection," in *Proceedings of the International Workshop on Semantic Evaluation*, San Diego, California, USA, 2016.

- [9] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, “Stance and sentiment in tweets,” *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, vol. 17, no. 3, 2017.
- [10] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, “A dataset for detecting stance in tweets,” in *Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC-2016)*, Portoroz, Slovenia, 2016.
- [11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.
- [12] M. C. M. Muniz, “A construção de recursos linguístico-computacionais para o Português do brasil: o projeto de Unitex-PB,” Master’s thesis, ICMC / USP São Carlos, 2004.
- [13] T. Mikolov, S. Wen-tau, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proc. of NAACL-HLT-2013*. Atlanta, USA: Association for Computational Linguistics, 2013, pp. 746–751.

Melhorias linguísticas no alinhador texto-imagem LinkPICS

1st João Gabriel Melo Barbirato

Departamento de Computação – DC

Universidade Federal de São Carlos – UFSCar

São Carlos, Brasil

joaobarbirato@gmail.com

2nd Helena de Medeiros Caseli

Departamento de Computação – DC

Universidade Federal de São Carlos – UFSCar

São Carlos, Brasil

helenacaseli@ufscar.br

Index Terms—alinhamento, texto, imagem, visão computacional, processamento de língua natural

I. INTRODUÇÃO

O alinhamento texto-imagem é a tarefa que tem o propósito de encontrar as correspondências entre elementos textuais e elementos visuais em um par de texto e imagem associada. Mais especificamente, o alinhamento texto-imagem visa determinar quais são as palavras (ou sequências de palavras) de um texto que estão associadas a uma área específica da imagem que acompanha o texto.

Uma ferramenta computacional implementada com este propósito é o LinkPICS [5]. O LinkPICS tem como objetivo alinhar elementos textuais presentes na notícia com elementos correspondentes da imagem que acompanha o texto. Apesar do LinkPICS ter apresentado uma boa precisão – 98% no alinhamento de pessoas e 72% no alinhamento de objetos, segundo valores relatados em [5] para o idioma inglês – ele ainda apresenta limitações.

Entre as limitações já identificadas, três foram tratadas neste trabalho. São elas:

- 1) **Realiza alinhamento apenas de palavras** – Na versão atual do LinkPICS, cada alinhamento está definido para uma região da imagem e apenas uma palavra no texto. Assim, a ferramenta não detecta entidades representadas por expressões multipalavras, tais como *vacuum cleaner* ou *panda bear*. A Figura 1 ilustra essa limitação.
- 2) **Realiza apenas alinhamento 1:1** – Na versão atual do LinkPICS, uma região da imagem corresponde a uma palavra no texto (alinhamento 1:1). Portanto, se várias regiões da imagem representam uma mesma entidade, elas não são alinhadas com a palavra (alinhamento 1:n). Por exemplo, para a imagem da Figura 2, que possui várias bicicletas, o LinkPICS consegue alinhar apenas uma delas, ao passo que poderia identificar que se trata de diferentes ocorrências de uma mesma entidade e alinhar todas com a mesma palavra no texto.
- 3) **Não considera sinônimos** – Além disso, a versão atual do LinkPICS não identifica sinônimos como sendo ocorrências possíveis da mesma entidade. Por exemplo,

o texto que acompanha a Figura 2 eventualmente pode apresentar a palavra *bike*, sinônima de *bicycle*, e essa não seria considerada uma possível instância da entidade *bicycle*.

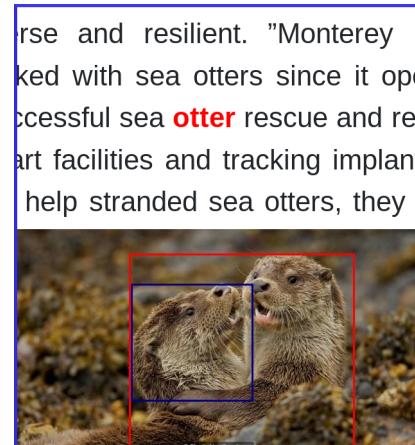


Figura 1. Exemplo de notícia na qual o LinkPICS deveria ter alinhado a região da imagem destacada pela *bounding box* em vermelho com a expressão multipalavra (*sea otter*) e não apenas a palavra *otter* destacada em vermelho no texto.

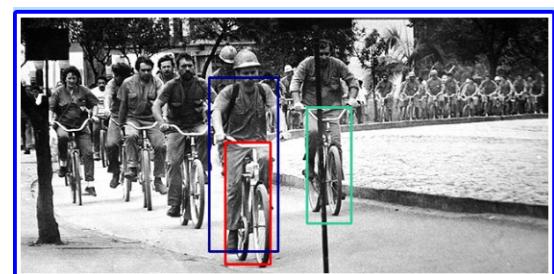


Figura 2. Nessa imagem é visível a presença de várias bicicletas, identificadas por suas *bounding boxes*, contudo a versão atual do LinkPICS é capaz de alinhar apenas uma dessas ocorrências com a palavra correspondente no texto da notícia.

O objetivo do presente trabalho é apresentar algumas estratégias implementadas para lidar com as limitações citadas. Para tanto, a próxima seção (II) descreve a estrutura geral do LinkPICS. As estratégias adotadas e implementadas para

superar as limitações citadas são descritas na seção III. Os experimentos realizados para avaliar essas modificações são descritos na seção IV. Por fim, a seção V traz as considerações finais e propostas de trabalhos futuros.

II. O ALINHADOR TEXTO-IMAGEM LINKPICS

O alinhador automático LinkPICS [5] visa alinhar elementos textuais de sites de notícias com elementos correspondentes da imagem que acompanha o texto. Inicialmente, essa ferramenta foi concebida para analisar textos em inglês¹. Sua arquitetura pode ser dividida em cinco principais etapas, especificadas logo abaixo de II-A a II-E.

A. Extração da notícia

A extração de elementos da notícia se dá a partir de um endereço WEB (URL) da notícia. A versão atual do LinkPICS está preparada para lidar com os formatos de notícias, em inglês, de dois jornais online: Folha de São Paulo International² e BBC³. Após a recuperação da notícia, seus elementos textuais (texto, legenda e título) e visuais (imagem) são enviados para as próximas etapas de processamento, realizadas em paralelo.

B. Processamento de imagem

A etapa de processamento de imagem visa identificar entidades na imagem e destacar suas respectivas regiões utilizando *bounding boxes*. Para isso, o LinkPICS utiliza a DarkNET/YOLO⁴. As *bounding boxes* identificadas são utilizadas na etapa de alinhamento.

C. Processamento de texto

Essa etapa visa processar os elementos textuais da notícia, como título, legenda e corpo do texto. Seu objetivo principal é detectar entidades (substantivos e entidades nomeadas) e organizá-las para a etapa de alinhamento. Entre as ferramentas utilizadas no processamento de texto estão o etiquetador morfossintático TreeTagger⁵, o NLTK⁶ e o reconhecedor de entidades nomeadas da Stanford⁷. Ao final dessa etapa tem-se uma lista de candidatos a entidades presentes no título, legenda e corpo da notícia, que podem aparecer na imagem.

¹Uma versão do LinkPICS também foi produzida para o português do Brasil [6], mas como o foco deste artigo é o idioma inglês, apenas os recursos para o inglês serão mencionados.

²Disponível em: <https://www1.folha.uol.com.br/internacional/en/>. Acesso em: 12 dez. 2018.

³Disponível em: <https://www.bbc.com/news>. Acesso em: 12 dez. 2018.

⁴“Open Source Neural Networks in C.” Disponível em: <https://pjreddie.com/darknet/>. Acesso em: 27 de maio 2018.

⁵Disponível em: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>. Acesso em: 21 maio 2019.

⁶Disponível em: <https://www.nltk.org/>. Acesso em: 21 maio. 2019.

⁷Disponível em: <https://nlp.stanford.edu/software/CRF-NER.shtml>. Acesso em: 12 dez. 2018.

D. Alinhamento de pessoas

O alinhamento de pessoas tem como entrada o conjunto de *bounding boxes* contendo pessoas, identificadas na etapa II-B, e o conjunto de entidades nomeadas, identificadas na etapa II-C. O LinkPICS, então, faz o alinhamento com base na similaridade entre esses dois conjuntos. Veltroni [5] compara vetores de características das faces detectadas na notícia – utilizando a biblioteca de aprendizado de máquina DLIB⁸ – com faces da base de dados LFW [3] por meio da distância euclidiana. Caso a distância entre a face detectada e as faces correspondentes no banco LFW seja menor ou igual a 0.6, o rótulo da face correspondente é atribuído à detectada e, então, a face é alinhada.

E. Alinhamento de objetos

O alinhamento de objetos segue a mesma estratégia do alinhamento de pessoas, só que agora as entradas são o conjunto de *bounding boxes* que não contém pessoas, identificadas na etapa II-B, e o conjunto de substantivos, identificados na etapa II-C. O LinkPICS, então, faz o alinhamento com base na similaridade entre esses dois conjuntos. Veltroni [5] propôs medir a similaridade a partir de *word embeddings* geradas utilizando o GloVe [4] ou com base na estrutura da WordNet usando o método proposto por [7] (medida WUP).⁹ Assim, o par <substantivo, *bounding box*> de maior similaridade é alinhado.

III. METODOLOGIA

Nesta seção são descritas as estratégias adotadas para resolver algumas das limitações da versão atual do LinkPICS. Em especial: (seção III-A) a identificação de expressões multipalavras (para solucionar a limitação de alinhamentos apenas de palavras), (seção III-B) a implementação do alinhamento de uma palavra (ou expressão multipalavra) no texto com várias regiões da imagem para permitir o alinhamento 1:n), e (seção III-C) o reconhecimento de sinônimos (para permitir o alinhamento n:1).

A. Alinhamento de expressões multipalavra

Como solução para a limitação de alinhamento apenas de palavras, adotou-se a definição de padrões de categorias gramaticais que compõem uma expressão multipalavra¹⁰. Para tanto, considerou-se as etiquetas morfossintáticas geradas pelo TreeTagger, como explicado em II-C, e utilizou-se o *parser* de expressões regulares da biblioteca NLTK para identificar as expressões multipalavras que fossem compostos nominais. Nesse processo, foram identificadas a primeira palavra, a última e, caso houvesse mais de duas, as palavras do meio no composto. A Figura 3 demonstra esse processo.¹¹

⁸Disponível em: <http://dlib.net/>. Acesso em: 28 de maio 2018

⁹Nos experimentos apresentados em [5] foram obtidos resultados melhores usando a medida WUP [7] e, por isso, ela foi adotada para cálculo de similaridade nos experimentos apresentados neste artigo.

¹⁰Candidatos a padrões são os definidos em: <http://aim-west.imag.fr/what-are-mwes/>, identificados com o auxílio da ferramenta NLTK. Acesso em: 12 dez. 2018.

¹¹Etiquetas especificadas em <https://courses.washington.edu/hypertxt/csar-v02/penntable.html>. Acesso em: 17 maio 2018

1. Frase extraída da notícia

(...) has worked with sea otters since it opened (...)

2. Frase etiquetada:

(...) has worked with sea otters since it opened (...)

VBZ VBN IN NN NNS IN PP VBD

3. Frase contendo compostos nominais identificados:

(...) has worked with **sea otters** since it opened (...)

VBZ VBN IN { NN NNS } IN PP VBD
composto nominal

Figura 3. Exemplo de identificação de compostos nominais. Em 1, tem-se o trecho de uma sentença extraída da notícia apresentada na Figura 1. Em 2, tem-se o mesmo trecho agora etiquetado pelo TreeTagger no qual é possível identificar a etiqueta *NN* seguida por *NNS*, que configura um candidato a composto nominal. Por fim, 3 mostra, em destaque, a identificação do mesmo com agrupamento das duas etiquetas como um composto nominal.

Para evitar compostos ruidosos (falsos candidatos) utilizou-se o coeficiente de Dice. Assim como mostra Ramisch [2] e ilustra a Equação 1, esse coeficiente se baseia na quantidade de vezes onde o composto $c(\vec{w})$ e cada uma das palavras que o compõem, $c(w_i)$, aparece no texto alvo (cujo número de palavras é N).

$$dice = \frac{N \times c(\vec{w})}{\sum_{i=1}^N c(w_i)} \quad (1)$$

Por exemplo, no caso do candidato a expressão multipalavra apresentado na Figura 3, o coeficiente de Dice foi 0,4878.

O Algoritmo 1 ilustra os passos desse processo. Nesse algoritmo duas funções são chamadas, uma para marcar os candidatos a compostos nominais com base nas expressões regulares (MARCAR_COMPOSTOS) e outra para calcular o coeficiente de Dice (DICE). Um limiar para Dice de 0,065 foi usado, e, portanto, o candidato que obtivesse um coeficiente de Dice maior que o limiar seria classificado como um bom candidato. Esse algoritmo foi executado na etapa de processamento de língua natural do LinkPICS (descrita na seção II-C), de modo que pudesse ter acesso a todos os elementos textuais etiquetados pelo módulo do TreeTagger. Tendo os compostos nominais devidamente identificados, eles foram marcados no texto da notícia e os alinhamentos do substantivo que eles continham foram atribuídos a eles também.

B. Alinhamento 1:n

Notou-se que a limitação de não alinhar uma mesma palavra no texto com várias regiões identificadas na imagem era uma limitação de implementação. Isso porque o rótulo encontrado no processamento de imagens era usado como chave para a identificação da *bounding box* da região da imagem e,

Algorithm 1 expr_multipalavras

Require: *dice, palavras_etiquetadas as etiquetadas*

```

1: function IDENTIFICAR_COMPOSTOS(etiquetadas)
2:   limiar  $\leftarrow 0.065$ 
3:   melhores_candidatos  $\leftarrow$  LISTA()
4:   candidatos  $\leftarrow$  MARCAR_COMPOSTOS(etiquetadas)
5:   for composto  $\in$  candidatos do
6:     if DICE(composto)  $>$  limiar then
7:       melhores_candidatos.add(composto)
8:   return melhores_candidatos

```

portanto, não aceitava repetições. Para a resolução dessa limitação, foram utilizadas chaves incrementais para elementos repetidamente rotulados.

C. Reconhecimento de sinônimos

A solução adotada para reconhecimento de sinônimos foi utilizar a interface da WordNet [1] para recuperar os sinônimos (*synsets*) e utilizar a medida WUP [7] para medir a similaridade entre a palavra alvo e de seus sinônimos. Caso essa distância fosse maior que um dado limiar (utilizou-se o limiar de 0,63), significaria que trata-se de um bom candidato a sinônimo da palavra alvo. O Algoritmo 2 ilustra essa implementação. Nesse algoritmo, duas funções são chamadas: uma para recuperar os *synsets* da WordNet (SYNSETS) e outra para calcular a medida WUP (WUP).

Algorithm 2 n_para_1

Require: *nltk.wordnet.synsets, wup_similarity as wup*

```

1: function OBTER_SINÔNIMOS(palavra_alinhada)
2:   sinônimos  $\leftarrow$  LISTA()
3:   limiar  $\leftarrow 0.63$ 
4:   for sin  $\in$  SYNSETS(palavra_alinhada) do
5:     if WUP(sin, palavra_alinhada)  $>$  limiar then
6:       sinônimos.add(sin)
7:   return sinônimos

```

Exemplos de sinônimos identificados seguindo essa estratégia foram: *aeroplane* (para o termo *plane*), *motorcar* e *automobile* (para *car*).

Diferente do Algoritmo 1, que é executado junto com a etapa de processamento de texto, o Algoritmo 2 foi executado após o alinhamento, uma vez que se deseja encontrar sinônimos apenas de termos alinhados.

IV. EXPERIMENTOS E RESULTADOS

Para avaliar as estratégias adotadas para a solução das limitações da versão atual do LinkPICS, foram realizados experimentos usando a base dedados e as medidas de avaliação descritas a seguir.

A. Base de dados

A base de dados é composta por 81 notícias em inglês extraídas do site internacional da BBC¹², as mesmas utilizadas

¹²<https://www.bbc.com/>

em [5], as quais consistem em pares de texto-imagem. Essa base contém, na maior parte, notícias relacionadas a objetos e outras categorias diferentes de pessoas – como animais, tecnologia e veículos.

B. Medidas de avaliação

Os alinhamentos produzidos pelo LinkPICS após a implementação das melhorias descritas na seção III foram avaliados manualmente com base nos seguintes critérios:

- a palavra é coerente (ou não) com a região da imagem com a qual foi alinhada;
- o sinônimo encontrado é coerente (ou não) com a região da imagem com a qual foi alinhada;
- o composto nominal é coerente (ou não) com a região da imagem com a qual foi alinhada.

Assim, utilizou-se a definição comum de precisão para dados binários, descrita pela equação 2, uma vez que é interessante apenas mensurar se o alinhamento de palavra, sinônimo ou composto nominal com a *bouding box* foi identificado de forma correta ou não.

$$p = \frac{n_{corretos}}{n_{corretos} + n_{incorrectos}} \quad (2)$$

C. Resultados

Das 81 notícias do conjunto de dados, em [5], foram gerados 78 alinhamentos dos quais 56 estavam corretos. Assim, a precisão no alinhamento de objetos obtida em [5] foi de 71,79%.

A versão do LinkPICS incluindo as melhorias relatadas neste trabalho encontrou mais alinhamentos, 106, o que é consequência direta do tratamento da limitação #2 referente ao alinhamento envolvendo apenas uma região da imagem. Desses alinhamentos, 66 foram avaliados como corretos, obtendo, assim, uma precisão de aproximadamente 62,26%. Vale ressaltar, portanto, que as melhorias implementadas trouxeram um aumento no número de alinhamentos gerados, com uma leve perda na precisão.

Em relação ao alinhamento de sinônimos, foram identificados 3 sinônimos de palavras alinhadas e todos esses foram avaliados como possíveis alinhamentos para a região da imagem alinhada à palavra original (precisão de 100%).

Além disso, foram identificados 119 compostos nominais (uma média de aproximadamente 1,12 compostos por alinhamento) dos quais 50 foram avaliados como corretamente alinhados com a região relacionada à palavra que eles continham, apresentando a precisão de aproximadamente 42,02%. Isso mostra que, mesmo utilizando a medida de similaridade Dice, ainda há ruídos na identificação de compostos nominais.

A Figura 4 traz um exemplo de alinhamento obtido com as implementações realizadas neste trabalho.

V. CONSIDERAÇÕES FINAIS

Este trabalho apresentou melhorias linguísticas na ferramenta de alinhamento texto-imagem LinkPICS. Por meio dessas melhorias foi possível encontrar outros elementos de significados semelhantes – como sinônimos de uma palavra e



Patrick the wombat dies and Zara criticised for Pepe skirt

Patrick was the world's oldest and **biggest bare-nosed captive wombat**

Social media users mourn the loss of **celebrity wombat** Patrick, and Zara is criticised for marketing a denim skirt featuring a lookalike of the "hateful" Pepe the Frog meme. The death of Australian "**celebrity wombat**" Patrick, said to be the oldest bare-nosed **wombat** in captivity, has

Figura 4. Exemplo de notícia alinhada. É possível notar que a palavra *wombat* foi alinhada com a região da imagem marcada em vermelho. Também é possível notar que esse alinhamento também foi atribuído aos dois compostos nominais automaticamente identificados: “*biggest bare-nosed captive wombat*” e “*celebrity wombat*”.

compostos nominais que a contém – e identificar instâncias diferentes de uma mesma categoria de objetos – alinhamento de várias regiões da imagem.

Os tópicos aqui abordados fazem parte de um projeto de iniciação científica apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e como próxima etapa está a geração de descrições de imagem utilizando ferramentas que armazenam informações semânticas do texto. Assim, espera-se que essas melhorias contribuam com a elaboração de um modelo de geração mais robusto.

Entre os trabalhos futuros para esta pesquisa está a investigação de abordagens mais refinadas para a identificação de compostos nominais, pois os resultados mostraram que a utilização apenas dos elementos textuais de uma notícia não são suficientes para identificar compostos com alta precisão. Uma possibilidade, neste caso, é acoplar a ferramenta de identificação de expressões multipalavras mwetoolkit¹³ [2] ao LinkPICS.

AGRADECIMENTOS

Este trabalho foi desenvolvido com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) processo nº 2018/11367-6 (Iniciação Científica) e faz parte do Projeto MMeaning (Auxílio Regular FAPESP, processo nº 2016/13002-0).

¹³Disponível em: <https://gitlab.com/mwetoolkit/mwetoolkit3>. Acesso em: 22 maio 2019.

REFERÊNCIAS

- [1] C. Fellbaum, “*Wordnet*,” in *Theory and applications of ontology: computer applications*, Springer, 2010, p. 231-243.
- [2] C. Ramisch, A. Villavicencio e C. Boitet, “*Mwetoolkit: a framework for multiword expression identification.*,” in LREC, Valletta, vol. 10, 2010, p. 662–669
- [3] G. B. Huang, M. Ramesh, T. Berg e E. Leonard-Miller, “*Labeled faces in the wild: a database for studying face recognition in unconstrained environments.*,” in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.
- [4] J. Pennington, R. Socher e C. D. Manning, “*Glove: Global Vectors for Word Representation*,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, p. 1532–1543.
- [5] W. C. Veltroni, “Alinhamento texto-imagem em sites de notícias,” Dissertação de mestrado, DC, UFSCar, São Carlos, BR, 2018.
- [6] W. C. Veltroni e H. de Medeiros Caseli, “*Text-Image Alignment in Portuguese News Using LinkPICS*,” in *International Conference on Computational Processing of the Portuguese Language*, 2018, p. 125–135,
- [7] Z. Wu e M. Palmer, “*Verbs Semantics and Lexical Selection*,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics*, 1994, p. 133–138.

Investigação do uso de *word embeddings* para cálculo de similaridade em memórias de tradução

1st Karina Mayumi Johansson
Departamento de Computação – DC
Universidade Federal de São Carlos – UFSCar
São Carlos, Brasil
kahjohansson@gmail.com

2nd Helena de Medeiros Caseli
Departamento de Computação – DC
Universidade Federal de São Carlos – UFSCar
São Carlos, Brasil
helenacaseli@ufscar.br

Palavras-Chave—*Word Embeddings, Memórias de Tradução, Ferramentas CAT, Português do Brasil, Inglês*

I. INTRODUÇÃO

Os avanços tecnológicos que ocorreram nos últimos anos relacionados à popularização da internet tornaram possível a disponibilização cada vez maior de conteúdo multilíngue, principalmente em sites e repositórios de informação acessíveis e editáveis por todos, como as wikis. Nesse cenário, a tradução desse conteúdo (para entendimento ou disponibilização de uma versão em outro idioma) tornou-se uma demanda não apenas de pesquisadores, cientistas e membros do governo, mas do público em geral. A internet proporcionou às pessoas o acesso à informação de um modo nunca antes visto pela sociedade e a disponibilidade crescente de informação multilíngue se contrapõe à dificuldade dos usuários da internet, espalhados por todo o mundo, em entender esse conteúdo.

Embora existam, atualmente, diversos sistemas e ferramentas para a tradução automática (TA), comerciais ou disponíveis online, os mesmos ainda não são capazes de produzir TA de qualidade comparável a do humano, em domínios irrestritos.

Além dos sistemas completamente automáticos, devido à qualidade contestável das traduções por eles produzidas, outras frentes de pesquisa se desenvolveram com o intuito de criar ferramentas e recursos para auxiliar o humano na tarefa de traduzir ou de editar a tradução antes, durante ou depois de sua realização.

Especificamente em relação à primeira frente, é nela que se inserem as ferramentas investigadas neste trabalho, as *Computer-Assisted Translation (CAT) tools*, que utilizam como principal recurso as Memórias de Tradução (MT). Uma MT pode ser entendida como um repositório que armazena os segmentos originais (geralmente frases) e as respectivas traduções humanas, com o objetivo de serem manipulados e reaproveitados em traduções futuras. Com o passar do tempo de uso, a MT agrupa um grande conjunto de segmentos acompanhados de suas respectivas traduções e esses podem ser recuperados pela ferramenta CAT por meio de um casamento (*matching*), completo ou parcial, sempre que um segmento igual ou semelhante aparecer novamente. Assim, uma MT pode ser considerada um *córpus paralelo*, pois contém seg-

mentos alinhados de textos já traduzidos, que têm a função de serem reaproveitados em uma futura tradução.

Entre as principais vantagens do uso de memórias de tradução estão: a consistência e a agilidade. As MT asseguram que os documentos traduzidos são consistentes, incluindo terminologia, estruturas, expressões e definições comuns. Além disso, elas também aceleram o processo de tradução já que dividem com o tradutor humano a pesada carga de tradução recordando para ele o que já foi previamente traduzido e deixando-o focado na tradução de novos trechos.

Ferramentas CAT podem oferecer, ainda, funcionalidades extra, como: análises estatísticas, exportação e importação de MT, conversão de formatos, trabalho colaborativo e alinhamento. Por meio do alinhamento é possível estabelecer a correspondência entre segmentos (ou palavras) de um arquivo original com os segmentos (ou palavras) de um arquivo que foi traduzido sem o uso das MT permitindo, assim, a criação “automática” de uma MT ou um glossário. Apesar de serem bastante úteis, essas funcionalidades extras não são o foco principal deste trabalho.

Neste trabalho apresentamos um projeto cujo objetivo é investigar a aplicabilidade de *word embeddings* para implementar a funcionalidade principal relacionada ao uso de MT em uma ferramenta CAT, que é o casamento (*matching*) entre segmentos da sentença sendo traduzida e os segmentos presentes na MT. A estratégia tradicionalmente usada para implementar o casamento é considerar a intersecção (ou sobreposição) nas sequências de palavras (n-gramas) presentes nos segmentos de texto em comparação, o que pode ser descrito como casamento de n-gramas. Essa estratégia tradicional, contudo, não é capaz de capturar similaridade semântica além do nível trivial [2].

Como alternativa ao casamento de n-gramas, o projeto apresentado neste trabalho propõe investigar como as *word embeddings* podem ser usadas para encontrar a similaridade entre segmentos de uma MT. As *word embeddings* (ou vetores de palavras) são representações vetoriais de palavras, geradas de modo não supervisionado a partir de córpus. Elas representam as palavras em um espaço vetorial no qual a similaridade semântica entre duas palavras é calculada com base na proximidade dos vetores que as representam [2].

Neste trabalho, portanto, apresenta-se um projeto de investigação do uso de *word embeddings* mono e bilíngues

para implementar a funcionalidade de casamento de segmentos de texto fundamental nas ferramentas CAT. Para tanto, a seguir, são descritos brevemente os dois tópicos principais desta pesquisa: memórias de tradução e *word embeddings*. Em seguida, são descritos os objetivos e a metodologia a ser seguida, bem como as duas ferramentas CAT selecionadas como candidatas a alteração nesse projeto. Por fim, são apresentados os resultados esperados e as possíveis contribuições do projeto em questão.

II. MEMÓRIAS DE TRADUÇÃO

Segundo [5], as memórias de tradução (MT) são uma das principais fontes de conhecimento que dão suporte à tradução humana nas ferramentas CAT. Uma MT é uma base de dados que armazena segmentos fonte e alvo chamados de unidades de tradução (do inglês, *translation units* ou TU). Esses segmentos podem ser fragmentos sub-sentenciais, sentenças completas ou mesmo parágrafos em duas línguas e, idealmente, eles são traduções perfeitas uns dos outros.

Para que esses segmentos sejam usados em uma ferramenta CAT é necessário calcular uma pontuação de casamento entre uma sentença de entrada sendo traduzida e o lado fonte de cada TU armazenada na MT. Se a pontuação atingir um valor mínimo, o lado alvo da TU é sugerido como uma tradução para o usuário [5].

Novas TUs são inseridas na MT a todo momento, ou seja, sempre que o usuário realiza a tradução de um segmento fonte gerando um segmento alvo equivalente para ele, este par (essa TU) pode ser armazenado na MT para uso futuro. Esse crescimento constante da MT garante sua utilidade para o tradutor humano que não precisará, no futuro, re-traduzir segmentos previamente traduzidos garantindo a consistência nas traduções e agilizando seu trabalho. Como bem colocado por [5], juntamente com a quantidade, a qualidade das TUs armazenadas é um fator determinante na utilidade da MT.

Exemplos de sistemas de MT são OmegaT¹, Wordfast², Déjà Vu da Atril Solutions³, memoQ⁴ e SDL Trados⁵ entre outros.

A Figura 1 ilustra o uso de uma MT em uma ferramenta CAT, o SDL Trados. Na tabela presente na parte inferior da figura são apresentados, na coluna da esquerda, os segmentos do texto fonte e, na coluna da direita, as traduções correspondentes a cada segmento. Na parte superior da figura são apresentadas as semelhanças e diferenças do segmento atual do texto fonte com o segmento da MT com melhor pontuação de casamento.

Ainda na parte superior, à direita, é apresentada a tradução recuperada da MT e sua pontuação de casamento. A primeira sentença teve uma pontuação de casamento (do inglês, *context match* ou CM) de 100% uma vez que seu contexto era semelhante ao da TU. Nesse caso, o usuário optou por recuperar a

tradução da TU e utilizá-la sem modificações. Para as segunda e terceira sentenças, é possível observar que os valores de CM foram, respectivamente, 88% e 77%. Nesses casos, o usuário optou por realizar modificações na tradução recuperada.

Assim, a partir das informações apresentadas na ferramenta, o tradutor humano verifica a sentença fonte selecionada e caso tenha sido encontrada uma TU semelhante na MT, ele pode optar por: (1) utilizar a tradução da TU e modificá-la se necessário, ou (2) ignorar a tradução sugerida e criar uma nova tradução para o segmento. Ao confirmar a tradução e prosseguir para o novo segmento, a nova TU é guardada na MT, e pode ser utilizada nos demais segmentos do texto, assim como em outros projetos, por meio da importação dessa MT. Deste modo, o tradutor segue esse processo iterativo até o fim do documento.

III. WORD EMBEDDINGS

As *word embeddings* (ou vetores de palavras) são representações vetoriais de palavras, geradas de modo não supervisionado a partir de córpus, que representam as palavras em um espaço vetorial. A similaridade semântica entre duas palavras é, então, calculada com base na proximidade dos vetores que as representam [2].

Essas similaridades e disparidades semânticas também podem ser estendidas para duas ou mais línguas. Em [3], por exemplo, os autores propõem a construção de modelos de língua monolíngues para inglês, espanhol e tcheco usando grandes córpus e, em seguida, usam um dicionário bilíngue pequeno para aprender uma projeção linear entre os vetores que representam cada língua. Desse modo, a tradução de uma palavra fonte é obtida projetando seu vetor para a língua alvo e buscando pelo vetor mais similar na língua alvo.

Exemplos de ferramentas que geram *word embeddings* são word2vec⁶, GloVe⁷, fastText⁸ e MUSE⁹.

Como as *word embeddings* capturam as similaridades sintáticas e semânticas em uma ou várias línguas, elas têm sido aplicadas, entre outros, para encontrar a similaridade entre textos. Em [1], por exemplo, foram usadas *word embeddings* bilíngues para recuperar sentenças semanticamente similares. Foram utilizados três pares de línguas: inglês-espanhol, inglês-italiano e inglês-croata. Como explicado pelos autores, a similaridade semântica em textos é estimada com base no quanto dois textos estão semanticamente relacionados ou associados, o que varia desde a equivalência semântica (na qual o significado dos dois textos é exatamente o mesmo) até a completa disparidade (na qual o significado de um texto está completamente dissociado do significado do outro). Assim, segundo esses autores, a similaridade semântica entre dois textos é dada pela sobreposição de significado entre os dois

⁶Disponível em: <http://deeplearning4j.org/word2vec>. Acesso em: 29 maio 2019.

⁷Disponível em: <https://nlp.stanford.edu/projects/glove/>. Acesso em: 29 maio 2019.

⁸Disponível em: <https://fasttext.cc/>. Acesso em: 29 maio 2019.

⁹Disponível em: <https://github.com/facebookresearch/MUSE>. Acesso em: 29 maio 2019.

¹Disponível em: <https://omegat.org/>. Acesso em: 28 maio 2019.

²Disponível em: <https://www.wordfast.net/>. Acesso em: 28 maio 2019.

³Disponível em: <https://atril.com/>. Acesso em: 28 maio 2019.

⁴Disponível em: <https://www.memoq.com/en/>. Acesso em: 28 maio 2019.

⁵Disponível em: <https://www.sdltrados.com/>. Acesso em: 28 maio 2019.

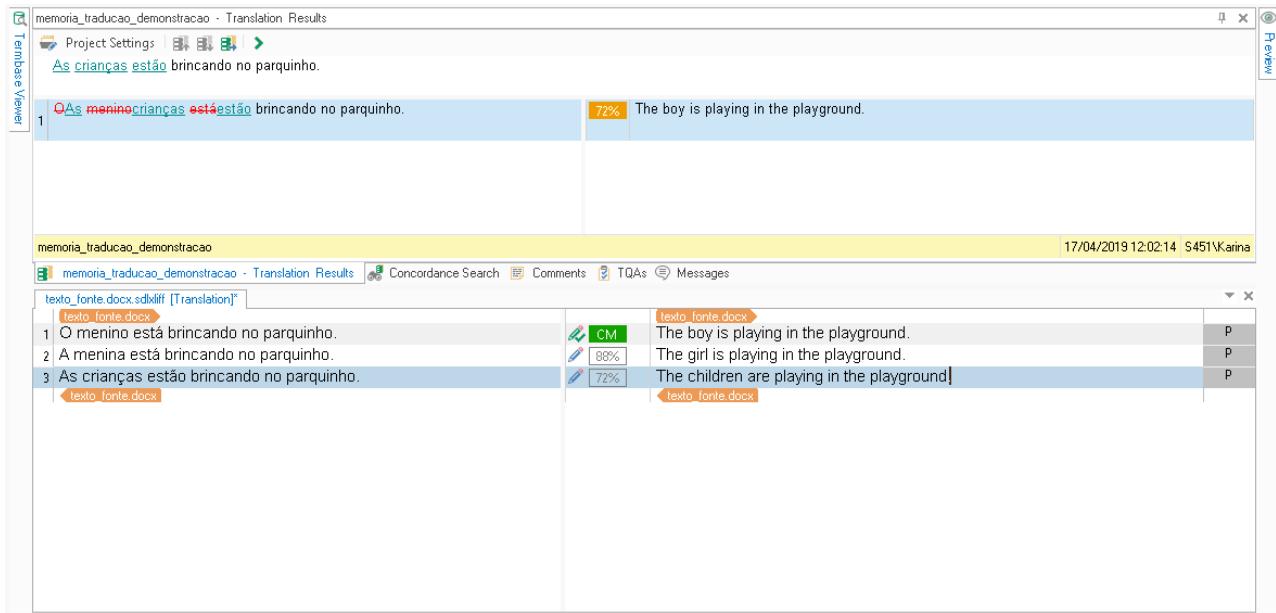


Figura 1. Exemplo de utilização de sistema de memória de tradução.

textos, por exemplo, as sentenças “o gato está ronronando” e “o cachorro está latindo” são mais relacionadas do que “a tartaruga está correndo”.

No âmbito das MTs, *word embeddings* bilíngues do par de línguas inglês-italiano foram usadas por [5] como uma das *features* no processo de limpeza das MTs, no qual TUs de baixa qualidade foram eliminadas da MT.

IV. OBJETIVOS E METODOLOGIA

A partir da contextualização apresentada anteriormente, pode-se estabelecer como objetivo do projeto aqui apresentado: verificar a aplicabilidade de *word embeddings* no cálculo da similaridade entre segmentos de uma sentença sendo traduzida e os segmentos de uma MT.

Assim, as seguintes questões de pesquisa são levantadas para este trabalho:

- 1) O uso de *word embeddings* para casamento de segmentos de uma sentença sendo traduzida e os segmentos de uma MT (TUs) é melhor do que a estratégia tradicional baseada em n-gramas?
- 2) O uso de *word embeddings* bilíngues em conjunto com *word embeddings* monolíngues é efetivo?

Para atingir o objetivo apresentado anteriormente e responder às questões de pesquisa levantadas, as seguintes etapas serão realizadas:

- 1) Levantamento das principais ferramentas CAT de código aberto para seleção daquela que será alterada nesse projeto,
- 2) Levantamento das memórias de tradução para o português do Brasil e o inglês que poderão ser usadas como córpus base para a avaliação,
- 3) Levantamento das *word embeddings* disponíveis para o português do Brasil e o inglês, bem como o estudo das

principais ferramentas de geração de *word embeddings* mono e bilíngues, caso seja necessário gerar *word embeddings* específicas para esse projeto,

- 4) Proposição e implementação da estratégia de cálculo de similaridade de segmentos de MT usando separada e conjuntamente *word embeddings* monolíngues e bilíngues,
- 5) Avaliação das estratégias propostas com base na comparação da qualidade do *matching* no sistema *baseline* (versão do sistema CAT sem qualquer alteração) e nas versões do sistema nas quais será implementado o *matching* usando *word embeddings* mono e bilíngues separada e conjuntamente.

O projeto sobre o qual trata este trabalho está no início de seu desenvolvimento e, até o presente momento, apenas a primeira atividade foi finalizada, para a qual os resultados são apresentados na seção V.

V. FERRAMENTAS CAT

A primeira atividade do projeto ao qual este trabalho se refere trata do levantamento e estudo das ferramentas CAT de código aberto disponíveis livremente. Nesse sentido, como resultado de um primeiro levantamento foram selecionadas as seguintes ferramentas: MateCat [7] e OmegaT [8].

As ferramentas citadas possuem recursos em comum, como: (i) utilização de memórias de tradução com *fuzzy matching*¹⁰ e (ii) pré-tradução por meio de *plugins* que acessam tradutores automáticos como Apertium¹¹ e Google Tradutor¹².

¹⁰Fuzzy matching é um casamento com taxa de similaridade menor que 100% entre o segmento a ser traduzido e um segmento da memória de tradução. Cada ferramenta CAT possui um valor padrão de limitante inferior, que pode ser alterado pelo usuário.

¹¹Disponível em: <https://www.apertium.org/>. Acesso em: 29 maio 2019.

¹²Disponível em: <https://translate.google.com/>. Acesso em: 29 maio 2019.

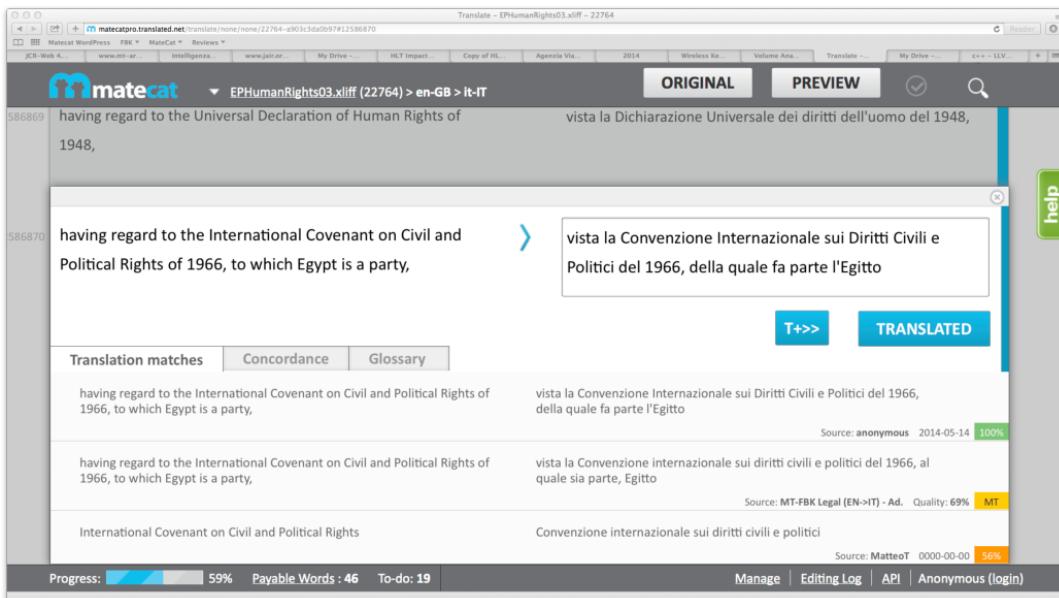


Figura 2. Interface da ferramenta CAT MateCat. [6]

Uma diferença existente é que OmegaT foi desenvolvido em plataforma *desktop* e MateCat em plataforma *web*.

MateCat é um acrônimo para *Machine Translation Enhanced Computer Assisted Translation*. Trata-se de uma ferramenta CAT com funcionalidades de: (i) pós-revisão, (ii) outsourcing de gerenciamento de projetos, traduções e revisões e (iii) trabalho em equipe, por meio de divisão e gerenciamento de projetos. A Figura 2 apresenta sua *interface*, contendo em seu lado esquerdo os segmentos fonte, e à direita, suas respectivas traduções.

OmegaT é uma ferramenta desenvolvida em Java, com *plugins* em diversas linguagens, como: Javascript, Groovy e Python. Seus principais diferenciais são: (i) propagação de correspondência (*match propagation*), ou seja, assim que um segmento é traduzido, a mesma tradução é inserida automaticamente em todos os segmentos idênticos e (ii) suporte de dicionários mono e multilíngue. A Figura 3 apresenta sua tela de tradução, que possui no canto inferior esquerdo os segmentos originais e abaixo de cada um deles, suas traduções. No canto superior esquerdo são apresentadas as *fuzzy matches*, e ao seu lado direito, são apresentadas as sugestões do tradutor automático.

A Tabela I apresenta um comparativo dessas ferramentas. Além da análise de características pontuais, como as listadas na Tabela I, uma análise qualitativa das funcionalidades e facilidade de uso e alteração de código ainda será realizada para determinar qual delas será a escolhida para alteração nesse projeto.

VI. CONTRIBUIÇÕES E RESULTADOS ESPERADOS

Este trabalho apresentou um projeto, em estágio inicial de desenvolvimento, que visa investigar a aplicabilidade de *word*

Tabela I
COMPARATIVO DAS FERRAMENTAS CAT

Categorias	Ferramentas	
	MateCat	OmegaT
Plataforma	Web	Desktop
Linguagem de programação utilizada	PHP	Java
Licença	LGPL license	GNU General Public License v3.0
MT públicas integradas	MyMemory	MyMemory
Formatos de MT suportados	TMX	TMX, TTX, TXML, XLIFF e SDLXLIFF

embeddings no cálculo da similaridade entre segmentos de uma sentença sendo traduzida e os segmentos de uma memória de tradução.

Ao final desse projeto, espera-se obter uma estratégia para cálculo da similaridade semântica que seja uma alternativa à estratégia de casamento tradicional baseada em n-gramas. Para validar e avaliar a estratégia proposta ela será implementada/incorporada em uma ferramenta CAT de código aberto permitindo casamentos mais semanticamente motivados e, como tal, mais abrangentes.

Embora *word embeddings* tenham sido usados para detecção de similaridade textual [1], [2] e para limpeza de MT [5], não se tem notícia de um trabalho que tenha investigado a aplicação de *word embeddings* mono e bilíngues para o casamento de segmentos nas MTs. Assim, este trabalho surge como a primeira iniciativa de investigação neste contexto.

Uma outra contribuição deste projeto refere-se ao principal idioma sob investigação, o português do Brasil, que ainda é carente de pesquisas e desenvolvimento de recursos/ferramentas

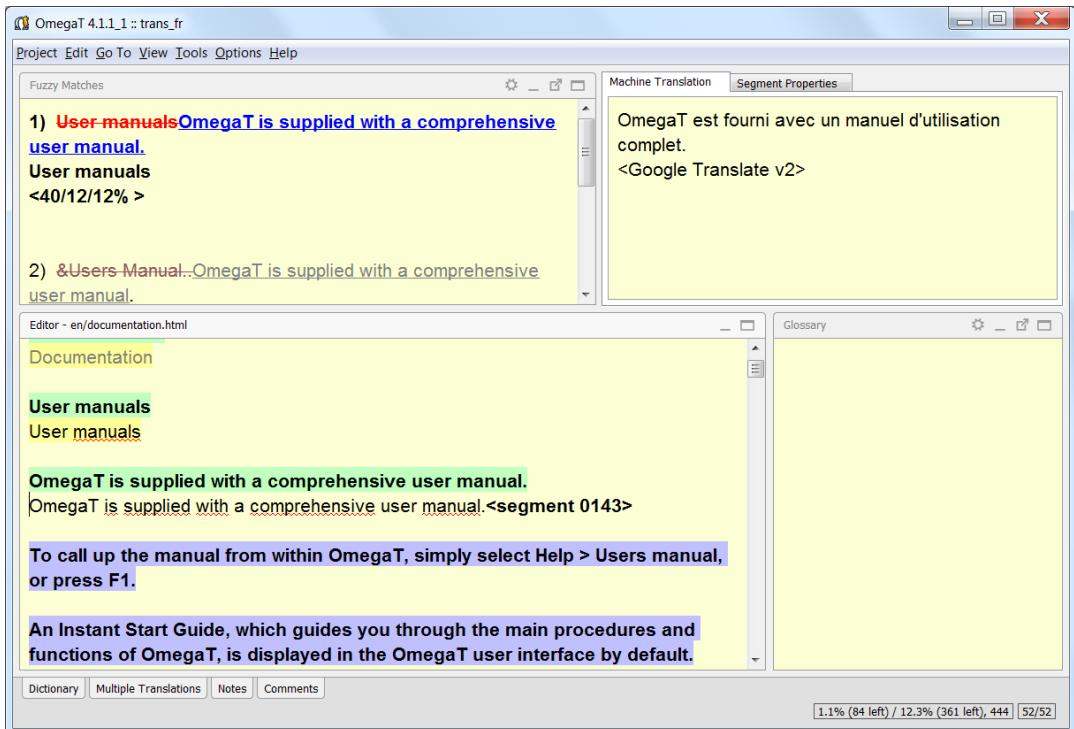


Figura 3. Interface da ferramenta CAT OmegaT. [8]

para a tradução (humana e automática) quando comparado ao cenário de outros idiomas, como o inglês.

REFERÊNCIAS

- [1] G. Glavas, M. Franco-Salvador, S. Ponzetto and P. Rosso, “A Resource-Light Method for Cross-Lingual Semantic Textual Similarity,” 2018.
- [2] T. Kenter and M. Rijke, “Short Text Similarity with Word Embeddings,” in Proc. 24th ACM International Conference on Information and Knowledge Management (CIKM ’15), 2015, pp. 1411-1420.
- [3] T. Mikolov, Q. Le and I. Sutskever, “Exploiting Similarities among Languages for Machine Translation,” 2013.
- [4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems (NIPS), 2013, pp. 3111-3119.
- [5] M. J. Sabet, M. Negri, M. Turchi and E. Barbu, “An Unsupervised Method for Automatic Translation Memory Cleaning,” in Proc. 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 287–292.
- [6] M. Frederico et al., “The Matecat Tool,” in 25th International Conference on Computational Linguistics, 2014.
- [7] “MateCat,” matecat.com [Online]. Available: <https://www.matecat.com>. [Accessed May 29, 2019].
- [8] “OmegaT,” omegat.org [Online]. Available: <https://omegat.org/>. [Accessed May 29, 2019].

Preparação para Leitura Distante em português: diálogos entre PLN e Humanidades Digitais

Luísa Rocha

*Departamento de Letras
PUC-Rio & Linguateca
Rio de Janeiro, Brasil
l.rocha7@globo.com*

Cláudia Freitas

*Departamento de Letras
PUC-Rio & Linguateca
Rio de Janeiro, Brasil
claudiafreitas@puc-rio.br*

Diana Santos

*Linguateca & ILOS
Universidade de Oslo
Oslo, Noruega
d.s.m.santos@ilos.uio.no*

Index Terms—Distant Reading, Leitura à distância, Corpus, Literatura, Humanidades Digitais

I. INTRODUÇÃO

Humanidades Digitais é uma área de atividade acadêmica que junta ciências humanas e recursos digitais para criar novos pontos de vista e métodos para esses estudos. Leitura Distante, ou "ler à distância", ou "distant reading" [1], é uma das técnicas nas humanidades digitais, criada por Franco Moretti, para estudar literatura de uma forma diferente. Consiste em distanciar-se das obras literárias, não como um obstáculo, mas sim como uma forma de obter novos ângulos, vendo menos detalhes e mais relações, padrões e formas em uma obra ou entre obras. Para Moretti, entender padrões entre os romances de determinado período de tempo, pode revelar uma regularidade do período, o que ele chama de "estrutura temporária", e o gênero do romance seria seu reflexo literário.

Com o intuito de uma maior familiarização com a área, tendo em vista a utilização da leitura distante em obras literárias de língua portuguesa, foi feito um levantamento de alguns trabalhos para se reunir exemplos e ideias de pesquisas para se fazer no corpus OBras [2]. Composto atualmente por 246 obras de literatura brasileira em domínio público e em permanente expansão, o Corpus OBras foi criado para ser a contraparte brasileira do Corpus Vercial. Todo material está anotado com informação morfossintática e semântica, e está disponível para consulta e para ser baixado [3]. No entanto, para que as explorações aconteçam da melhor maneira possível, são necessários ajustes e cuidados no tratamento das obras, de modo a não comprometer as pesquisas, como a atribuição do gênero correto dos nomes próprios, e o acréscimo de informação semântica (semas), local, pessoa ou religião [4]. A adição de novas obras, principalmente de autoras, ao corpus, também é uma tarefa em andamento.

II. FAMILIARIZAÇÃO COM A ÁREA: ALGUNS TRABALHOS RELACIONADOS

Os três trabalhos resumidos abaixo são exemplos de estudos que utilizam Leitura Distante para analisar algumas obras. As leituras foram importantes para futuros trabalhos com obras brasileiras.

O trabalho de [5] propõe-se a detectar diferentes emoções nos textos literários. Detecção de sentimentos é um método muito usado para comentários de usuários em redes sociais e avaliação de produtos online, então os autores se perguntaram se sua aplicação em textos literários seria tão útil quanto para a outra área. Para isso, foi criado um dicionário de alemão das palavras associadas às sete emoções de Plutchik e Ekman, e o aplicaram em um estudo de caso com dois romances de Franz Kafka. Utilizando-se do dicionário, 300 palavras foram anotadas como sentimentos pertencentes a uma das sete emoções, correspondendo aproximadamente a: raiva, desgosto, medo, felicidade, tristeza, surpresa e satisfação. Dois entre três anotadores concordaram na anotação de 85% das palavras, enquanto todos os três só concordaram em 46%, o que indica como pode ser difícil a anotação de sentimentos. Com os resultados, os autores tentaram detectar mudanças como aumento ou diminuição de sentimentos ao longo dos romances, assim como quais sentimentos são mais presentes em cada personagem.

Já o trabalho de [6] pretende examinar as redes sociais dos personagens em nove romances de Jane Austen e Charles Dickens. A análise de redes sociais (Social Network Analysis - SNA) permite pesquisas com um nível único de abstração, ao mesmo tempo em que mantém a estrutura coletiva dos personagens dentro do enredo.

Os textos vêm do Projeto Gutenberg e foram anotados manualmente por estudiosos de literatura com uma metodologia "radicalmente inclusiva". Para anotar cada obra, primeiro foi criado um dicionário de personagem contendo uma única entrada para cada um e seus apelidos. Cada personagem se torna um nó e cada coocorrência entre nós forma uma aresta no gráfico. Foram contabilizadas todos os tipos de coocorrências e não só conversas diretas, pois essa maneira permitia capturar um maior número de interações e associações entre personagens.

Um dos resultados foi observar que as sociedades construídas por Austen parecem ser mais compactas do que as de Dickens. Outro ponto observado foi o uso de muitas micronarrativas por Dickens, apesar dos autores acreditarem que tal técnica não afeta o enredo central, sendo só um estilo pessoal do autor.

Por fim, [7] utilizam-se dos romances ingleses do século

XIX para mapear as emoções da cidade de Londres. Os autores empregaram um programa de REM (Reconhecimento de Entidade Mencionada) para selecionar as passagens que se referiam à cidade, como ruas ou bairros. Essas passagens são compostas por uma sequência de duzentas palavras, tendo a localização de Londres no centro. Em seguida, ‘voluntários’ deveriam identificar (anotar) as emoções expressas ali. Houve muita discordância entre os anotadores ‘voluntários’ (foi um experimento de anotação *crowdsouce*) e a anotação feita por alunos de graduação, que serviram como grupo de controle. Por isso, os autores decidiram reduzir as opções de emoções para os opositos: “Frightening”(assustador) e “Happy”(feliz). Essas etiquetas só foram atribuídas às passagens se pelo menos metade dos anotadores concordasse na resposta. Essas mesmas passagens também foram analisadas pelo “Sentiment Analysis Program” de Standford, que utiliza um dicionário de termos positivos e negativos.

Antes de discutir os resultados, o artigo apresenta um contexto da expansão geográfica da cidade de Londres, e como era esperado que tal crescimento fosse representado na literatura. Contudo, só o centro e a parte oeste da cidade, que crescia para longe do rio, foram as mais citadas. Os bairros mais frequentes foram Westminster e The City, o que reflete a parcialidade nas representações da cidade de Londres.

Com isso, o primeiro achado da pesquisa foi a ausência do crescimento da cidade na Londres Ficcional, já que o foco nos dois bairros permanecia o mesmo. A característica heterogênea do bairro The City é um dos motivos apontado para sua permanência como local dos romances, enquanto, por outro lado, a homogeneidade de Westminster, bairro das classes mais altas, pode ser apontado como o principal motivo para sua permanência.

Um ponto interessante, ressaltado rapidamente pelos autores, é a “semântica do espaço”, ou seja, o léxico que mais foi associado àquele lugar. Referente a Westminster, o léxico mostrava opulência (parques, praças e jardins), patriarcado e servidão (condes, servos, ordem) e também rituais de “socialização formal” (encontros e visitas).

Ao finalmente apresentar a “geografia do massustador ou da felicidade” na cidade de Londres, se descobriu, na verdade, uma cidade sem emoções. A maior parte das passagens foi anotada como “emocionalmente neutra” (67% com anotadores humanos e 78% com a ferramenta de *Sentiment Analysis*). Todavia, naquelas passagens marcadas com emoções, novamente West End e The City foram as regiões que mais apareceram marcadas com *feliz* e *assustador*, respectivamente.

III. CORREÇÕES FEITAS NO OBRAS

Os textos lidos oferecem múltiplas possibilidades de se trabalhar com a técnica de leitura distante e são inspiração para pesquisas com o corpus OBras [2], composto por obras de literatura brasileira dos séculos XVII até XX. Em pesquisas preliminares, percebemos a necessidade de fazer algumas modificações e correções na anotação para que as buscas obtivessem resultados mais precisos.

A. Correção de gênero gramatical dos Nomes Próprios sem gênero definido

Utilizando a expressão de busca [pos=”PROP.*hum”& gen=”M/F”] no serviço AC/DC [8] – ferramenta de acesso e interrogação de corpus –, a pesquisa retornou 4185 palavras, todos nomes próprios que o parser PALAVRAS [9] classificou como “humanos” [10] e para os quais atribuiu um gênero gramatical indefinido (Masculino/Feminino). O objetivo desta etapa foi revisar essa lista para atribuir o gênero gramatical correto e revisar a classificação semântica do nome próprio, caso fosse um erro (isto é, se o nome próprio não fosse “humano”). Abaixo, em negrito, estão exemplos de palavras anotadas como nome próprio e com gênero indefinido.

Para tal revisão, as 4185 palavras foram transformadas em um arquivo de lista tipo texto simples (txt), nomeado *lista-ACDC* e depois foram retirados todos os nomes terminados em “-a” – formando um novo arquivo de lista, nomeado “*terminados em a* – e, depois, foram retirados todos os nomes terminados em “-o” – formando um terceiro arquivo de lista, nomeado *terminados em o*. Em seguida, o arquivo *lista-ACDC* com as alterações foi renomeado para *nomes M-F*. A lista foi examinada para separar manualmente aqueles que eram erros (Exemplo 1), e nomes facilmente reconhecíveis (Exemplo 2 e Exemplo 3). Aqueles que possivelmente fossem sobrenomes não foram retirados da lista. Os erros formaram um quarto arquivo chamado *lista de erros*.

- (1) id=”Os_Sertões_II Prosa:prosa EdC 1902 ”: **Adiante** recuava o sertanejo, recuando pelos cômodos escusos.
- (2) id=”A_falência Prosa:romance JLdA 1901 naturalismo_realismo”: indagou **D.Joana** .
- (3) :id=”O_gaúcho Prosa:romance JdA 1870 romantismo, regionalismo”: Enfim estava **Juca** um mancebo.

Conferimos que todos os nomes no arquivo *terminados em a* eram femininos para torná-lo o arquivo dos *nomes femininos*, assim, adicionamos os outros nomes posteriormente identificados como femininos. Também conferimos todos os nomes no arquivo *terminados em o* com o mesmo intuito, depois renomeado para *nomes masculinos*.

Uma vez que, no arquivo *nomes M-F*, só restasse nomes desconhecidos ou dúvidas, retornamos ao AC/DC para procurar as passagens em que esses nomes apareciam, no intuito de encontrar algum indicador de gênero: pronome, determinante, adjetivo ou contexto (Exemplo 4 e Exemplo 5). Às vezes, só as ocorrências não eram suficientes para determinar o gênero, então esses nomes foram separados para uma pesquisa mais detalhada a ser feita depois que o arquivo *nomes M-F* tivesse sido todo pesquisado no AC/DC. Os nomes separados foram digitados junto ao nome da obra na internet para procurar um resumo ou relação de personagem (Exemplo 6).

- (4) id=”Ubijarara Prosa:romance JdA 1874 indianismo_romantismo”: Murinhém atravessou rápido a campina e apresentou-se em frente de **Canicrã**, chefe dos tapuias.

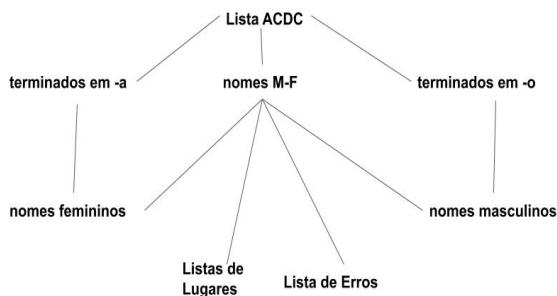


Figura 1. Fluxograma sobre as listas de nomes próprios

- (5) id="Macunaíma Prosa:romance MdAndrade 1928 modernismo": Zozoiaça riu bem por causa que não podia dar **Taina-Cã** de casamento pra filha velha não.
- (6) id="Os_trabalhadores_do_mar Prosa:romance MdA 1866 ": Nessa noite, **Clubin** ceou à mesa dos guardas das costas, e, contra o costume, saiu logo depois de cear.

Ao final, o arquivo *lista-ACDC* se repartiu em quatro arquivos de lista: *Nomes Masculinos*; *Nomes Femininos*; *Lista de Erros*; *Listas de Lugares*.

Nessa correção, optamos por deixar os pronomes de tratamento como “Vossa Excelência”, “vosmice”, “Vossa senhoria”, “Majestade” etc. com a pos=“PROP.*hum”, pois nosso foco era nomes próprios de pessoas. No caso de nomes próprios não humanos, como o cachorro chamado Black (Exemplo 7), também não foram considerados Pessoa .

Os nomes indígenas nas obras “UbirajaraMacunaíma”e ”O ermitão do Muquém”foram interessantes de se observar, pois eram muito diferentes (Exemplo 8 e Exemplo 9). Nesse caso foi necessário uma pesquisa com contexto mais amplo e optou-se por usar a marcação pessoa ficcional.

- (7) id="Contos_fora_da_moda Prosa:conto ArtA 1894 ": A ventura de Leandrinho tinha um único senão: havia na casa um cãozinho de raça, um bull-terrier, chamado **Black**, que latia desesperadamente sempre que farejava a presença daquele estranho.
- (8) id="Macunaíma Prosa:romance MdAndrade 1928 modernismo": Era **Emoron-Pódoile**, o Pai do Sono .
- (9) id="O_ermitão_do_Muquém Prosa:romance BG 1866 romantismo,regionalismo": – Já que Tupã, exclama ele enfurecido, obstinadamente me persegue, e me acabrunha ao peso de amarguras e infortúnios, sem que eu nada tenha feito para merecê-los, eu te invoco, ó **Anhangá**, e de hoje em diante ponho em tuas mãos o meu destino, ó manitó do mal!

1) Atribuição do sema e correções associadas: Além do gênero, o tipo semântico de cada nome próprio também foi adicionado, em uma revisão da anotação automática fornecida pelo parser PALAVRAS [9]. Marcado pela tag *sema*, a classe semântica é utilizada para indicar variadas informações semânticas. As classes semânticas podem ser várias, e em nosso caso nos interessam os tipos pessoa, lugar, obra ou religião [4]. A classe das pessoas pode ser subdividida em pessoa ficcional, usada para personagens ficcionais, e pessoa histórica, usada para pessoas que existiram. Sobre os nomes próprios referentes a religião, que foram poucos, optou-se por colocar todos com sema Religião. Assim tanto entidades e deuses quanto santos estariam com o mesmo sema. Contudo, mitologias e seres mitológicos (Exemplo 8) foram marcados com sema pessoa ficcional, [sema=Pessoa:ficc].

2) Correção de segmentação: A busca pelos nomes próprios evidenciou também muitos erros de segmentação. Encontramos diversos casos de D. ou S. que, por causa do ponto final, atrapalhavam a análise automática (Exemplo 10 e Exemplo 11), ou de nomes compostos, que eram separados, mas deveriam constituir uma unidade, apenas (Exemplo 12)

- (10) id="A_Marquesa_de_Santos Prosa:romance PS 1925 histórico": D. Pedro devia conduzi-la até o salão onde estava a Imperatriz .
- (11) id="Um_dístico Prosa:conto MdA 1886 ": Mentalmente nunca soube o que era; talvez refletia no concílio de Constantinopla, nas penas eternas ou na exortação de **S. Basílio** aos rapazes .
- (12) id="A_Marquesa_de_Santos Prosa:romance PS 1925 histórico": **D. Ilda Mafalda de Sousa Queiroz**, a rutilante Marquesa de Valença, vestido de gorgorão negro, cadeia de ouro e mitenes de seda, corre pelos grupos o seu lorgnon de madrepérola .

B. Correção de pré-processamento

Erros de pré-processamento acontecem quando o texto não é tratado da forma correta. Os erros mais comuns eram nomes de capítulos tratados como se fizessem parte do texto do capítulo. (Exemplo 13),

- (13) id="Os_trabalhadores_do_mar Prosa:romance MdA 1866 ": CAPÍTULO II CLUBIN DESCOBRE ALGUÉM

De forma semelhante, uma pesquisa com nomes próprios do tipo lugar está em andamento. Novamente, tiramos proveito da anotação automática feita pelo PALAVRAS [9], e as expressões de busca utilizadas até o momento foram [pos=“PROP.*top”] e [pos=“PROP.*civ”], retornando o resultado de 3530 palavras, aproximadamente.

C. Adição de obras

As novas obras já tratadas e já adicionadas ao corpus são os cinco romances: “O Coruja”, de Aluísio Azevedo, “O Bom-Crioulo”, de Adolfo Caminha, “Os Dois Amores”, de

joaquim Manuel de Macedo, “A Carne”, de Júlio Ribeiro, e “O Homem”, de Aluísio Azevedo. Foram pré-processados seguindo as instruções na página do projeto. Apenas uma obra trouxe um caso não abordado nas instruções.

No romance “Os Dois Amores”, um dos personagens lê uma história, de forma que está escrita para que o leitor também leia. Optou-se por não tratar esse caso como uma canção ou poema (ou seja, como “tipos especiais de texto”), mas sim como parte do capítulo a que pertence.

A pesquisa para encontrar obras escritas por mulheres, a fim de deixar o corpus mais balanceado quanto ao gênero da autoria e comparar estilos, está em andamento, porém há algumas dificuldades. As obras precisam pertencer ao domínio público e a digitalização, preferencialmente, não deve ser do tipo imagem, o que dificularia o trabalho do pré-processamento.

IV. CONCLUSÃO

Os textos estudados são fonte de várias ideias para se trabalhar com a leitura distante em português, especificamente com o corpus OBras, e para que os resultados sejam mais precisos, correções e melhorias são necessárias. Mais de mil correções envolvendo atribuições de gênero, inclusão de sema e correções de segmentação e pré-processamento foram aplicadas no corpus OBras. As correções do sema lugar serão implementadas em breve, assim como adição de obras escritas por mulheres. A melhoria e expansão do corpus sintaticamente anotado é bom para toda e qualquer pesquisa de PLN, podendo também servir de treino para modelos de aprendizado automático.

AGRADECIMENTOS

Luísa Rocha é bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico, no âmbito do projeto “Recursos para o ‘distant reading’ em português: diálogos entre PLN e Humanidades Digitais”. Número do processo da Bolsa: 101815/2019-0

REFERÊNCIAS

- [1] F. Moretti, *A Literatura Vista de Longe*. Porto Alegre: Arquipélago, 2008.
- [2] D. Santos, C. Freitas, and E. Bick, “OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain.” OpenCor, 2018.
- [3] Linguateca. [Online]. Available: "<https://www.linguateca.pt/OBRAS/OBRAS.html>"
- [4] D. Santos and C. Freitas, “Estudando personagens na literatura lusófona,” in *Anais do STIL’2019*, Salvador, Brasil, 2019.
- [5] R. Klinger, S. S. Suliya, and N. Reiter, “Automatic Emotion Detection for Quantitative Literary Studies – A case study based on Franz Kafka’s “Das Schloss” and “Amerika”,” in *Digital Humanities 2016: Conference Abstracts*. Kraków, Poland: Jagiellonian University and Pedagogical University, July 2016, pp. 826–828. [Online]. Available: <http://dh2016.adho.org/abstracts/318>
- [6] S. Grayson, K. Wade, G. Meaney, J. Rothwell, M. Mulvany, and D. Greene, “Discovering Structure in Social Networks of 19th Century Fiction,” in *Proceedings of the 8th ACM Conference on Web Science*, ser. WebSci ’16. New York, NY, USA: ACM, 2016, pp. 325–326. [Online]. Available: <http://doi.acm.org/10.1145/2908131.2908196>
- [7] Heuser, Ryan and Moretti, Franco and Steiner, Erik, “The emotions of london,” in *Literary Lab Pamphlet 13*, 2016. [Online]. Available: <https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf>

- [8] D. Santos and E. Bick, “Providing Internet access to Portuguese corpora: the AC/DC project,” in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*. Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: <http://www.lrec-conf.org/proceedings/lrec2000/pdf/85.pdf>
- [9] E. Bick, “The Parsing System “PALAVRAS”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework,” Ph.D. dissertation, 2000.
- [10] ———, “Functional Aspects in Portuguese NER,” in *Proceedings of the 7th International Conference on Computational Processing of the Portuguese Language*, ser. PROPOR’06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 80–89. [Online]. Available: http://dx.doi.org/10.1007/11751984_9

ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente

Elvis de Souza

Departamento de Letras

PUC-Rio

Rio de Janeiro, Brasil

elvis.desouza99@gmail.com

Cláudia Freitas

Departamento de Letras

PUC-Rio

Rio de Janeiro, Brasil

claudiafreitas@puc-rio.br

Palavras-chave—revisão de corpora, métodos de avaliação, anotação automática, processamento de linguagem natural, linguística computacional

I. INTRODUÇÃO

Sistemas de anotação morfossintática (pos-taggers e parsers) que utilizam tecnologia de aprendizado de máquina demandam corpora volumosos e bem anotados para aprenderem a anotar textos adequadamente. De modo geral, visando melhorar a qualidade dos anotadores automáticos, fez-se muito em relação à tecnologia subjacente aos sistemas, o que elevou drasticamente, no decorrer dos anos, a qualidade da anotação grammatical. No entanto, há um gargalo na qualidade do material que serve de treino, o que acaba impactando na qualidade da aprendizagem [1]. No corpus Bosque-UD [2], de textos jornalísticos em Língua Portuguesa adaptados para o formato Universal Dependencies [3], as métricas de POS, quando o modelo é treinado utilizando a ferramenta UDPipe [4], são relativamente confortáveis: 96,46% para a versão 2.3 do Bosque-UD. Contudo, esse mesmo material em outros níveis de análise linguística, como o de relações sintáticas, leva a resultados de 81.82%, indicando que ainda há um grande espaço para melhorias.

Acreditamos, junto com Manning [1], que um caminho para superar esses gargalos deve ser pela via linguística: melhorando a anotação dos corpora que servem de treino para os sistemas de anotação, tornando-os mais consistentes e eliminando possíveis erros humanos.

Nesse contexto, este trabalho relata a construção de uma Estação de Trabalho (ET) desenhada a partir da perspectiva linguística, com o objetivo de facilitar a revisão, a edição e a avaliação de corpora anotados, alinhando o trabalho feito pelos especialistas em língua, de um lado, e os resultados práticos, isto é, o desempenho de sistemas de PLN, de outro. Desse modo, discussões teóricas sobre as categorias gramaticais podem ser embasadas não apenas quanto à adequação linguística a certas teorias, mas também nos resultados empíricos de sistemas de aprendizado artificial.

Nossa Estação de Trabalho comprehende, até o momento, dois eixos centrais: 1) a edição e revisão do corpus anotado; e 2) a avaliação do resultado da revisão, feita a partir da análise da anotação automática. Apresentaremos a arquitetura

das ferramentas participantes da nossa ET tendo em vista que foram desenvolvidas objetivando duas tarefas complexas, em momentos distintos: primeiro, a contabilização de sujeitos ocultos no corpus Bosque-UD (seção 2), e posteriormente, o lançamento de uma nova versão, intensamente revisada, do mesmo corpus (seção 3).

É importante ressaltar que, embora estejamos apresentando as ferramentas publicamente, nosso objetivo, no lugar de simplesmente compartilhar os códigos empacotados¹ e descrever como os implementamos, é relatar o que nos levou a desenvolvê-los, destacando a importância de estruturar os fenômenos nos corpora sob diferentes perspectivas que possam motivar o trabalho linguístico no PLN, fomentando o diálogo entre as duas áreas.

II. CONTEXTUALIZAÇÃO

Começamos o desenvolvimento da ET, inicialmente, a partir de uma tarefa complexa, porque envolve contar algo que não tem materialidade: a contabilização de sujeitos ocultos no corpus Bosque-UD.

Para a quantificação de sujeitos ocultos em um corpus, precisamos, antes, tomar decisões qualitativas, isto é, foi necessário decidir quais frases consideraríamos “sujeito oculto”. Toda contagem é, antes, uma forma de qualificar os fenômenos que se nos apresentam [5], e decisões importantes como essas demandam, além de conhecimento teórico/gramatical, ferramentas que nos permitam visualizar os dados de que dispomos a fim de uma tomada de decisão com respaldo nas ocorrências da língua em uso. Evidentemente, nosso ponto de partida para a discussão foram as gramáticas, no entanto, sozinhas, elas não apresentam respostas para todos os fenômenos com que precisamos lidar em um corpus real, mesmo em casos simples como os sujeitos ocultos.

Nesse contexto, desenvolvemos o Interrogatório, um ambiente de busca em corpora no formato CoNLL-U escrito em Python e Javascript. À medida que nos aprofundamos na caça às sentenças com sujeito oculto, novas utilidades eram demandadas, de tal maneira que o ambiente foi sendo construído tendo em mente que, a qualquer momento, uma nova necessidade poderia emergir.

¹Disponíveis em <https://github.com/alvelvis/ACDC-UD>

CF19-1

Disse que não conseguia vislumbrar artifícios fraudulentos ou prática de peculato no protocolo assinado por Quérica.

[Mostrar contexto](#) [Mostrar anotação](#) [Mostrar opções](#) [Abrir inquérito](#)

Figura 1. Frase encontrada como resultado da primeira busca por sujeitos ocultos

Nossa primeira tentativa foi a de criar uma regra geral que respondesse por grande parte das sentenças com sujeito oculto, tal como nos informam as teorias gramaticais. Desenvolvemos, então, um padrão de busca que procurasse por sentenças em que um token X não tivesse Y como seu filho — no nosso exemplo, nenhuma “raiz” de sentença poderia ter “sujeito” como seu filho. E ter como primeiro problema uma busca que envolve a ausência foi um ótimo desafio.

Diferentemente da maioria das ferramentas que lida com dependências sintáticas (como por exemplo [6] e [7]), preferimos ver nossos resultados (as frases devolvidas) como linhas de concordâncias “simples”, com o trecho buscado em negrito (com cores diferentes para cada elemento indicado na expressão de busca), sem evidenciar as relações/anotações de dependência sintática, como ilustra a figura 1.

O negrito nas palavras encontradas e a possibilidade de visualizar diferentes palavras com cores é essencial para que possamos focalizar a nossa atenção no que de fato importa. Clicando em “Mostrar anotação”, ainda, é possível visualizar como todos os tokens da sentença estão anotados morfossintaticamente no corpus, inclusive com suas relações de dependência. Com essa funcionalidade, conseguimos elaborar os próximos passos da tarefa, pois podemos identificar, morfossinteticamente, quais as características que queremos considerar e quais podemos descartar em buscas futuras.

A partir dos resultados da primeira busca, adicionamos duas utilidades ao Interrogatório, e de grande relevância quando se deseja buscar elementos específicos não apenas para a busca e contagem, mas também para a revisão de corpus: (a) a possibilidade de filtrar os resultados de uma busca realizando outras buscas dentro da primeira, e (b) a capacidade de fazer pesquisas por dependência sintática.

Estudando os resultados e realizando os filtros necessários, chegamos à porcentagem de 16,04% de frases com sujeito oculto no corpus Bosque-UD (todos os procedimentos e resultados relativos à pesquisa com sujeitos ocultos estão descritos em [8]). Além do resultado, tínhamos em mãos uma ferramenta pronta para ser usada em outros projetos, desenvolvida com base no uso e aberta às mudanças que quiséssemos implementar a qualquer momento.

III. AVALIAÇÃO DO CORPUS

Posteriormente, nossa ET, que já contava com o Interrogatório — um ambiente em expansão e ainda incipiente no momento —, se viu diante da tarefa de nos auxiliar a identificar pontos em que nosso corpus poderia melhorar — seja na

Métricas oficiais					
Metric	Precision	Recall	F1 Score	AligndAcc	
Tokens	100.00	100.00	100.00		
Sentences	100.00	100.00	100.00		
Words	100.00	100.00	100.00		
UPOS	96.45	96.45	96.45	96.45	
XPOS	100.00	100.00	100.00	100.00	
UFeats	94.81	94.81	94.81	94.81	
AllTags	92.94	92.94	92.94	92.94	
Lemmas	96.94	96.94	96.94	96.94	
UAS	86.46	86.46	86.46	86.46	
LAS	82.42	82.42	82.42	82.42	
CLAS	75.33	74.89	75.11	74.89	
MLAS	67.40	67.01	67.20	67.01	
BLEX	72.05	71.63	71.84	71.63	

Figura 2. Exemplo de métricas de avaliação do CoNLL 2018 Shared Task

qualidade da anotação ou nas categorias linguísticas propostas. O Bosque-UD estava se preparando para o lançamento de uma nova versão [9], e enquanto o Interrogatório já era um ambiente que nos ajudaria na tarefa de visualizar as sentenças, precisávamos ainda de uma motivação empírica para guiar nossas correções — ou, para guiar nossa busca por uma estratégia sistemática de correção.

Em [10] exploramos a estratégia de revisão por divergência de modelos, usando dados das matrizes de confusão. Para dar continuidade a essa forma de revisão, nossa ET passou a integrar a visualização das métricas de avaliação da anotação feitas por algum modelo à visualização e correção das sentenças no corpus em si. Como ferramenta de anotação, escolhemos o UDPipe [4] e o tomamos como régua para comparar os modelos treinados a partir de diferentes versões do nosso corpus.

Para avaliar a qualidade das revisões — medidas a partir do desempenho do UDPipe na partição teste do corpus — usamos as métricas oficiais do CoNLL 2018 Shared Task [11]. A partir dessas métricas (Figura 2), conseguimos avaliar globalmente se o material de treino, isto é, nosso corpus, está mais consistente em relação a uma versão anterior, estabelecendo comparações entre as métricas anteriores e posteriores às nossas alterações. Para cada alteração substancial no corpus, realizamos um novo treinamento e avaliação do sistema.

Embora nos informem sobre a qualidade da anotação automática (e, de maneira indireta, sobre a qualidade/consistência da anotação manual), essas métricas pouco nos ajudam sobre o que podemos fazer para melhorar a anotação. Para nos indicar quais categorias precisam de uma revisão atenta, então, elaboramos uma outra maneira de visualizar a anotação, que envolve a distribuição de acertos por categoria. Isto é, além dos números globais, vemos, para cada tipo de anotação linguística (dependência sintática, POS etc.) o quanto o modelo acertou ou errou, o que nos diz, indiretamente, o quanto a anotação daquela categoria está consistente. A Figura 3 nos mostra a tabela para os casos de erros de POS, e vemos, pela tabela, que se a identificação de

Acurácia por categoria gramatical		GOLDEN	ACERTOS	PORCENTAGEM
UPOS				
ADJ		444	386	86.93693693693693%
ADP		1623	1616	99.5686999383857%
ADV		364	350	96.15384615384616%
AUX		283	272	96.113074204947%
CCONJ		210	203	96.66666666666667%
DET		1561	1543	98.8468930172966%
NOUN		1925	1837	95.42857142857143%
NUM		248	230	92.74193548387096%
PART		17	14	82.35294117647058%
PRON		318	295	92.76729559748428%
PROPN		851	812	95.41715628672151%
PUNCT		1343	1343	100.0%
SCONJ		100	84	84.0%
SYM		30	29	96.66666666666667%
VERB		853	821	96.24853458382181%
X		30	3	10.0%
-		742	742	100.0%

Figura 3. Acurácia de parte das classes gramaticais ao se comparar a anotação do corpus (Golden) e a previsão do parser UDPipe

UD[2]	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X	_	All
UD[1]																		
ADJ	386	1	2	0	0	0	27	0	0	0	5	0	0	0	23	0	0	444
ADP	0	1616	1	0	0	4	0	0	0	0	2	0	0	0	0	0	0	1623
ADV	2	3	350	0	0	2	3	0	0	1	1	0	1	0	1	0	0	364
AUX	0	0	0	272	0	0	1	0	0	0	0	0	0	0	10	0	0	283
CCONJ	0	1	3	0	203	0	2	0	0	0	0	0	1	0	0	0	0	210
DET	1	1	3	0	0	1543	0	1	0	6	6	0	0	0	0	0	0	1561
NOUN	29	2	3	2	0	3	1837	0	0	0	41	0	0	0	8	0	0	1925
NUM	2	2	0	0	0	1	3	230	0	0	10	0	0	0	0	0	0	248
PART	0	3	0	0	0	0	0	0	14	0	9	0	0	0	0	0	0	17
PRON	0	0	6	0	0	2	2	1	0	295	2	0	5	0	0	0	0	318
PROPN	4	0	1	0	0	0	29	2	0	1	812	0	0	0	2	0	0	851
PUNCT	0	0	0	0	0	0	0	0	0	0	0	1343	0	0	0	0	0	1343
SCONJ	0	1	3	0	0	0	0	0	0	12	0	0	84	0	0	0	0	100
SYM	0	0	0	0	0	0	0	0	0	0	1	0	0	29	0	0	0	30
VERB	13	1	1	2	0	0	5	0	0	0	5	0	0	0	821	0	0	853
X	1	2	0	0	0	0	2	0	0	0	17	0	0	0	0	3	0	30
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	742	742	
All	438	1633	373	281	203	1560	1916	234	14	315	982	1343	91	29	865	3	742	10942

Figura 4. Matriz de confusão de POS (classes gramaticais). Nas linhas, a anotação do corpus original, e nas colunas, a previsão do parser UDPipe

preposições (ADP) não traz preocupações (99% de acertos), o mesmo não pode ser dito da classe dos adjetivos (ADJ), com quase 87% de acertos.

Uma vez identificada uma categoria com porcentagem insatisfatória, lançamos mão da análise de matrizes de confusão (Figura 4), continuando a estratégia já utilizada em [12], para identificar quais são as divergências entre a anotação no corpus e a previsão do parser. As informações são disponibilizadas em páginas de extensão .html geradas automaticamente, possibilitando hiperlinks: em qualquer um dos números na matriz, visualizamos uma relação das frases que se encontram naquela interseção, com a palavra divergente em negrito para direcionar nossa atenção.

4 / 5

sent_id = CF852-4

text = Chocante.

ADJ PROPN - Comentários:

Mostrar UD[1] Mostrar UD[2]

Figura 5. Frase encontrada na análise da divergência entre ADJ e PROPN. Neste caso, o sistema previu erroneamente, pois “Chocante” é adjetivo, e não nome próprio

Por exemplo, na interseção entre ADJ e PROPN (casos em que, no corpus, a palavra estava anotada como ADJ, e o sistema previu PROPN), encontramos a frase da Figura 5 (de apenas uma palavra). Na mesma página podemos visualizar, ainda, a anotação completa que consta no corpus (UD[1]) e a anotação realizada pelo sistema (UD[2]).

É importante ressaltar que tanto a tabela de “Acurácia por categoria gramatical” quanto a matriz de confusão podem ser geradas para todas as informações morfossintáticas no formato UD, tais como classe grammatical e relação sintática.

IV. CORREÇÃO DO CORPUS

Sendo possível analisar caso a caso quais foram as divergências entre golden e modelo (ou entre modelos distintos), podemos elaborar generalizações que nos ajudem a avançar com a revisão do material, fazendo uma revisão linguisticamente motivada e tornando o material mais consistente.

Com o Interrogatório (seção 2), podemos realizar pesquisas nos corpora a partir de 5 critérios de busca (descritos na Figura 6). Essas pesquisas são guiadas pelo nosso conhecimento linguístico e pelo que generalizamos serem erros comuns ao analisar as matrizes de confusão.

Vejamos, por exemplo, mais uma divergência numerosa: entre adjuntos adverbiais e adjuntos adnominais. Frases como a da Figura 7 são muito comuns de figurarem entre as divergências. No corpus, acertadamente, “manhã” está anotado como adjunto adverbial dependente de um verbo, “ligar”. A anotação automática, porém, previu “manhã” como adjunto adnominal dependente de “televisão”, como sendo uma característica do objeto, assim como falamos “televisão de plasma” ou “televisão de LED”. Parece haver uma lógica no erro da máquina, e, de fato, trata-se de uma situação conhecida na linguística: ambiguidade do sintagma preposicionado. Devemos, portanto, procurar explorá-la para facilitar nossa correção, fazendo do uso de expressões específicas e dos filtros, já mencionados na seção 2.

No Interrogatório, para todas as frases que buscamos é possível “Abrir inquérito”, isto é, editá-las (Figura 8). Na interface, basta clicar na coluna que se deseja alterar e digitar a alteração. Então, o inquérito é salvo no corpus e as alterações são anexadas a um relatório com todos os “inquéritos” feitos pelo usuário.

Critério	Expressão	Resultados da busca
1 Regex	(.*dizer.*)\n(.+PROPN)	<p>É uma situação absurda», disse Taylor, 49.</p> <p>«Um monte de artistas vai querer ver os concertos», diz Ohtake.</p>
2 Ausência de B apontando para A	root#?#nsubj csubj#?	<p>Já não há o império do mal para combater.</p> <p>«Não damos conta de atendê-los.</p>
3 Regex Independentes	VERB.*root::!nsubj::!csubj	<p>Diariamente, está promovendo desfiles de moda para seus consumidores.</p> <p>Não houve acordo para uma trégua durante a Copa.</p>
4 País e filhos (dependê ncia)	\NOUN\l.*nsubj :: \tADV\l	<p>O árbitro Pinto Correia esteve bem durante toda a primeira parte e durante quase toda a segunda.</p> <p>Está longe a constituição no Nou Camp de um novo «dream team», como o de Romário e Stoichkov, o que já enerva a direcção.</p>
5 Sintaxe em Python	token.upos == "NOUN" and token.head_token.upos == "ADJ"	<p>O problema é político porque envolve, por exemplo, a gratuidade da educação, da saúde, da previdência mínima.</p> <p>Mulher morre em rio presa ao cinto do carro</p>

Figura 6. Exemplificação de critérios de busca no Interrogatório

# text = Os três cortadores de cana eram de Alagoas e estavam na cidade havia 15 dias.																				
# source = CETENFolha n=269 cad=Contidiano sec=soc sem=94a																				
# sent_id = CF269-2																				
# id = 1131																				
1 Os	o	DET	_	Definite=Def Gender=Masc Number=Plur PronType=Art	3	det	_	_	3	nummod	_	3	nsubj	_	8	case	_	5	cc	_
2 três	três	NUM	_	NumType=Card	8	nsbj	_	_	8	mod	_	8	root	_	0	advcl	_	0	obj	_
3 cortadores	cortador	NOUN	_	Gender=Masc Number=Plur	13	det	_	_	13	case	_	13	dep	_	10	obj	_	10	advcl	_
4 de	de	ADP	_	-	14	mod	_	_	14	nummod	_	14	obj	_	8	SpaceAfter=No	_	8	punct	_
5 cana	cana	NOUN	_	Gender=Fem Number=Sing	15	mod	_	_	15	nummod	_	15	obj	_	8	SpaceAfter=No	_	8	punct	_
6 eram	ser	AUX	_	Mood=Ind Number=Plur Person=3 Tense=Imp VerbForm=Fin	16	mod	_	_	16	nummod	_	16	obj	_	8	SpaceAfter=No	_	8	punct	_
7 de	de	ADP	_	-	17	mod	_	_	17	nummod	_	17	obj	_	8	SpaceAfter=No	_	8	punct	_
8 Alagoas	Alagoas	PROPN	_	Gender=Masc Number=Sing	18	mod	_	_	18	nummod	_	18	obj	_	8	SpaceAfter=No	_	8	punct	_
9 e	e	CCONJ	_	-	19	mod	_	_	19	nummod	_	19	obj	_	8	SpaceAfter=No	_	8	punct	_
10 estavam	estar	VERB	_	Mood=Ind Number=Plur Person=3 Tense=Imp VerbForm=Fin	20	mod	_	_	20	nummod	_	20	obj	_	8	SpaceAfter=No	_	8	punct	_
11-12 na	-	-	_	-	21	mod	_	_	21	nummod	_	21	obj	_	8	SpaceAfter=No	_	8	punct	_
11 em	em	ADP	_	-	22	mod	_	_	22	nummod	_	22	obj	_	8	SpaceAfter=No	_	8	punct	_
12 a	o	DET	_	Definite=Def Gender=Fem Number=Sing PronType=Art	23	mod	_	_	23	nummod	_	23	obj	_	8	SpaceAfter=No	_	8	punct	_
13 cidade	cidade	NOUN	_	Gender=Fem Number=Sing	24	mod	_	_	24	nummod	_	24	obj	_	8	SpaceAfter=No	_	8	punct	_
14 havia	haver	VERB	_	-	25	mod	_	_	25	nummod	_	25	obj	_	8	SpaceAfter=No	_	8	punct	_
15 15	15	NUM	_	NumType=Card	26	mod	_	_	26	nummod	_	26	obj	_	8	SpaceAfter=No	_	8	punct	_
16 dias	dia	NOUN	_	Gender=Masc Number=Plur	27	mod	_	_	27	nummod	_	27	obj	_	8	SpaceAfter=No	_	8	punct	_
17 .	.	PUNCT	_	-	28	mod	_	_	28	nummod	_	28	obj	_	8	SpaceAfter=No	_	8	punct	_

Figura 8. Interface de edição de sentenças no Interrogatório

A partir de uma página com resultados de busca, também, é possível aplicar regras de transformação escritas em Python, conforme descrito na documentação².

Voltando à etapa primeira da ET, compararmos os resultados das versões e, então, começamos novos experimentos, de tal modo que as duas etapas da ET se retroalimentam.

V. CONCLUSÃO

Apresentamos uma Estação de Trabalho para trabalhar com corpora anotados composta por duas frentes interligadas: (1) a busca, revisão e edição de frases; e (2) a avaliação da qualidade da anotação do corpus.

Em termos gerais, a ideia é que por meio da visualização de divergências entre golden e modelo (ou entre diferentes modelos) seja possível depreender padrões que possam nos ajudar a buscar mais consistência. Além disso, por meio do Interrogatório, temos à disposição um sistema de busca e correção de corpora motivado pelo tipo de interrogações que linguistas podem vir a fazer ao conjunto de sentenças, de maneira tal que ajude pesquisas linguísticas com base em corpus, por um lado, e torne mais eficiente o processo de correção, manual ou automática, por outro.

AGRADECIMENTOS

Elvis de Souza é bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) no projeto *Construção de datasets para o PLN de língua portuguesa*. Número do processo da bolsa: 128693/2019-3.

REFERÊNCIAS

- [1] C. D. Manning, “Part-of-speech tagging from 97% to 100%: is it time for some linguistics?” in *International conference on intelligent text processing and computational linguistics*. Springer, 2011, pp. 171–189.
- [2] A. Rademaker, F. Chalub, L. Real, C. Freitas, E. Bick, and V. de Paiva, “Universal dependencies for portuguese,” in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 2017, pp. 197–206.

²Disponível em <https://github.com/alvelvis/Interrogat-rio/wiki>

1 / 48

```
# sent_id = CP981-4

# text = Talvez o Presidente, quando liga a televisão de manhã e
descobre que apesar de Hillary continuar ao seu lado o país insiste
em discutir se ele deve ou não abandonar o cargo porque teve uma
relação extraconjugal com Monica Lewinsky, pense, como uma das
personagens de «Happiness»:

 obl  nmod - Comentários:



```

Figura 7. Divergência comum entre adjunto adverbial e adnominal. No corpus, “manhã” está anotado como adverbial, mas o sistema previu adnominal

- [3] R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, T. Oscar *et al.*, “Universal dependency annotation for multilingual parsing,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 92–97.
- [4] M. Straka, J. Hajic, and J. Straková, “Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing,” in *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*, 2016, pp. 4290–4297.
- [5] D. Santos, “Podemos contar com as contas?” *quot; In Sandra Maria Aluísio; Stella E O Tagrin (ed) New Language Technologies and Linguistic Research: A Two-Way Road Cambridge Scholars Publ 2014; 2014*, 2014.
- [6] F. M. Tyers, M. Sheyanova, and J. N. Washington, “Ud annotatrix: an annotation tool for universal dependencies,” 2017.
- [7] M. Janssen, “Dependency graphs and teitok: Exploiting dependency parsing,” in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2018, pp. 470–478.
- [8] C. Freitas, E. de Souza, and L. Rocha, “Quantificando (e qualificando) o sujeito oculto em português,” in *VI Jornada de Descrição do Português, STIL 2019*, 2019.
- [9] J. Nivre, M. Abrams, Ž. Agić, L. Ahrenberg, G. Aleksandravičiūtė, L. Antonsen, K. Aplonova, M. J. Aranzabe, G. Arutie, M. Asahara, L. Atayah, M. Attia, A. Atutxa, L. Augustinus, E. Badmaeva, M. Ballesteros, E. Banerjee, S. Bank, V. Barbu Mititelu, V. Basmov, J. Bauer, S. Bellato, K. Bengoetxea, Y. Berzak, I. A. Bhat, R. A. Bhat, E. Biagetti, E. Bick, A. Bielinskienė, R. Blokland, V. Bobicev, L. Boizou, E. Borges Völker, C. Börstell, C. Bosco, G. Bouma, S. Bowman, A. Boyd, K. Brokaitė, A. Burchardt, M. Candito, B. Caron, G. Caron, G. Cebiroğlu Eryiğit, F. M. Cecchini, G. G. A. Celano, S. Čéplő, S. Cetin, F. Chalub, J. Choi, Y. Cho, J. Chun, S. Cinková, A. Collomb, Ç. Çöltekin, M. Connor, M. Courtin, E. Davidson, M.-C. de Marneffe, V. de Paiva, A. Diaz de Ilarrazoza, C. Dickerson, B. Dione, P. Dirix, K. Dobrovoljc, T. Dozat, K. Droganova, P. Dwivedi, H. Eckhoff, M. Eli, A. Elkahky, B. Ephrem, T. Erjavec, A. Etienne, R. Farkas, H. Fernandez Alcalde, J. Foster, C. Freitas, K. Fujita, K. Gajdošová, D. Galbraith, M. Garcia, M. Gärdenfors, S. Garza, K. Gerdes, F. Ginter, I. Goenaga, K. Gojenola, M. Gökkirmak, Y. Goldberg, X. Gómez Guinovart, B. González Saavedra, M. Grioni, N. Gržūtis, B. Guillaume, C. Guillot-Barbance, N. Habash, J. Hajic, J. Hajic jr., L. Hà Mý, N.-R. Han, K. Harris, D. Haug, J. Heinecke, F. Hennig, B. Hladká, J. Hlaváčová, F. Hociung, P. Hohle, J. Hwang, T. Ikeda, R. Ion, E. Irimia, O. Ishola, T. Jelínek, A. Johannsen, F. Jørgensen, H. Kaşikara, A. Kaasen, S. Kahane, H. Kanayama, J. Kanerva, B. Katz, T. Kayadelen, J. Kenney, V. Kettnerová, J. Kirchner, A. Köhn, K. Kopacewicz, N. Kotsyba, J. Kovalevskaitė, S. Krek, S. Kwak, V. Laippala, L. Lambertino, L. Lam, T. Lando, S. D. Larasati, A. Lavrentiev, J. Lee, P. Lê H`ong, A. Lenci, S. Lerpradit, H. Leung, C. Y. Li, J. Li, K. Li, K. Lim, Y. Li, N. Ljubešić, O. Loginova, O. Lyshevskaya, T. Lynn, V. Macketanz, A. Makazhanov, M. Mandl, C. Manning, R. Manurung, C. Máränduc, D. Mareček, K. Marheinecke, H. Martínez Alonso, A. Martins, J. Mašek, Y. Matsumoto, R. McDonald, S. McGuinness, G. Mendonça, N. Miekka, M. Misirpashayeva, A. Missilä, C. Mititelu, Y. Miyao, S. Montemagni, A. More, L. Moreno Romero, K. S. Mori, T. Morioka, S. Mori, S. Moro, B. Mortensen, B. Moskalevskyi, K. Muischnek, Y. Murawaki, K. Müürisepp, P. Nainwani, J. I. Navarro Horňiacek, A. Nedoluzhko, G. Nešpore-Běžkalne, L. Nguyêñ Thi, H. Nguyêñ Thị Minh, Y. Nikaido, V. Nikolaev, R. Nitisoroj, H. Nurmi, S. Ojala, A. Oltókun, M. Omura, P. Osenova, R. Östling, L. Øvreliid, N. Partanen, E. Pascual, M. Passarotti, A. Patejuk, G. Paulino-Passos, A. Peljak-Łapińska, S. Peng, C.-A. Perez, G. Perrier, D. Petrova, S. Petrov, J. Piitulainen, T. A. Pirinen, E. Pitler, B. Plank, T. Poibeau, M. Popel, L. Pretkalniņa, S. Prévost, P. Prokopidis, A. Przepiórkowski, T. Puolakainen, S. Pyysalo, A. Rääbis, A. Rademaker, L. Ramasamy, T. Rama, C. Ramisch, V. Ravishankar, L. Real, S. Reddy, G. Rehm, M. Rießler, E. Rimkutė, L. Rinaldi, L. Rituma, L. Rocha, M. Romanenko, R. Rosa, D. Rovati, V. Roșca, O. Rudina, J. Rueter, S. Sadde, B. Sagot, S. Saleh, A. Salomoni, T. Samardžić, S. Samson, M. Sanguinetti, D. Särg, B. Saulīte, Y. Sawanakunanon, N. Schneider, S. Schuster, D. Seddah, W. Seeker, M. Seraji, M. Shen, A. Shimada, H. Shirasu, M. Shohibussirri, D. Sichinava, N. Silveira, M. Simi, R. Simionescu, K. Simkó, M. Šimková, K. Simov, A. Smith, I. Soares-Bastos, C. Spadine, A. Stella, M. Straka, J. Strnadová, A. Suhr, U. Sulubacak, S. Suzuki, Z. Szántó, D. Taji, Y. Takahashi, F. Tamburini, T. Tanaka, I. Tellier, G. Thomas, L. Torga, T. Trosterud, A. Trukhina, R. Tsarfaty, F. Tyers, S. Uematsu, Z. Urešová, L. Uria, H. Uszkoreit, S. Vajjala, D. van Niekerk, G. van Noord, V. Varga, E. Villemonte de la Clergerie, V. Vincze, L. Wallin, A. Walsh, J. X. Wang, J. N. Washington, M. Wendt, S. Williams, M. Wirén, C. Wittern, T. Woldemariam, T.-s. Wong, A. Wróblewska, M. Yako, N. Yamazaki, C. Yan, K. Yasuoka, M. M. Yavruyan, Z. Yu, Z. Žabokrtský, A. Zeldes, D. Zeman, M. Zhang, and H. Zhu, “Universal dependencies 2.4,” 2019, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: <http://hdl.handle.net/11234/1-2988>
- [10] L. Rocha, I. Soares-Bastos, C. Freitas, and A. Rademaker, “Scavenger hunt: what do we find when look for confusions,” in *International Conference on the Computational Processing of Portuguese, PROPOR 2018*, 2018.
- [11] D. Zeman, J. Hajic, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, “Conll 2018 shared task: multilingual parsing from raw text to universal dependencies,” in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018, pp. 1–21.
- [12] C. Freitas, L. F. Trugo, F. Chalub, G. Paulino-Passos, and A. Rademaker, “Tagsets and datasets: Some experiments based on portuguese language,” in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2018, pp. 459–469.

Um estudo sobre desidentificação de evoluções clínicas

Thaila Elisa Quaini, Henrique D. P. dos Santos, Sandra C. de Abreu, Bernardo S. Consoli and Renata Vieira

Pontifical Catholic University of Rio Grande do Sul, School of Technology

Email: {thaila.quaini, henrique.santos.003, sandra.abreu, bernardo.consoli}@acad.pucrs.br, renata.vieira@pucrs.br

I. INTRODUÇÃO

O prontuário eletrônico tem um papel importante no ambiente hospitalar trazendo benefícios para segurança do paciente e qualidade no sistema de saúde. [1]. Esses sistemas produzem enormes quantidades de dados, informações que podem revelar relações intrínsecas entre os sintomas, doenças, interações medicamentosas e o diagnóstico, que podem ser usadas para diferentes propósitos. [2]. Entretanto, para utilizar tais dados para pesquisa, primeiro é necessário remover os dados que possam permitir a identificação dos pacientes.

A desidentificação é um processo de retirada de dados e informações de uma pessoa, é a ação de retirar a identidade de algo ou alguém. Portanto com este processo, além de mantermos a privacidade dos pacientes, possibilitamos à pesquisadores e estudantes o uso de dados reais para desenvolvimento de suas pesquisas em saúde.

Neste trabalho, nós propomos um processo de desidentificação de prontuários eletrônicos a partir de uma lista de nomes especialmente construída para a tarefa e o comparamos com métodos usuais de Reconhecimento de Entidades Nomeadas.

O restante deste artigo está organizado da seguinte forma: A seção II apresenta os trabalhos relacionados. Os recursos utilizados são apresentados na seção III. A seção IV descreve a abordagem proposta, seguida pelos resultados e limitações encontradas durante o processo. Um comparativo com algoritmos de Reconhecimento de Entidades Nomeadas é apresentado na seção V. Por fim, na seção VI apresentamos as conclusões e trabalhos futuros.

A aprovação ética do uso do conjunto de dados do hospital nesta pesquisa foi concedida pelo Comitê de Ética em Pesquisa do Grupo Hospitalar Conceição sob o número 71571717.7.0000.5530.

II. TRABALHOS RELACIONADOS

Alguns trabalhos anteriores já trataram do problema de desidentificação de textos, assim como a criação de lista de nomes para a tarefa de identificação de entidade nomeada.

[3] descreve que os nomes não são retirados em sua totalidade, ou seja, alguns nomes são retirados parcialmente, como por exemplo, os sobrenomes. Diferentemente do nosso algoritmo proposto que procura retirar todo e qualquer nome, salvo exceções. Além disso, os autores aplicam Machine Learning para identificar os nomes que devem ser retirados ou não. Ocorre que em vários momentos existem erros, onde

é identificado automaticamente um falso positivo, ou seja, palavras comuns que não devem ser retiradas.

[4] fez uma pesquisa bibliográfica sobre artigos que citavam desidentificação e os resultados foram os mais diversos possíveis. Dentre mais de 200 artigos, somente 18 foram selecionados para a segunda fase da pesquisa, uma análise detalhada sobre os artigos que falavam sobre a desidentificação automática de textos. Alguns dos artigos retiravam somente conceitualmente as classes do PHI, alguns retiravam só os nomes dos pacientes, outros dos pacientes e do plano de saúde, e 10 deles retiravam todas as categorias. O artigo dividiu as classificações dos artigos analisados, das categorias as quais eles referenciavam as suas retiradas. Uma coisa que diminui a confiança no trabalho deles é que foram usados somente resumos de alta e relatórios de patologia, que são textos normalmente com padrões de escrita, e assim, de fácil desidentificação enquanto nós possuímos e trabalhamos em cima de todos os tipos de documentos.

[5] criou sua lista de nomes baseado na lista restrita existente em Portugal. Nessa lista existe um conjunto pré-definido de nomes legalmente aceitos para os recém nascidos, não é permitido o uso de nomes estrangeiros, variações de escrita ou qualquer outra fora da lista divulgada pelo Instituto dos Registos e do Notariado de Portugal¹. A lista ganha novos nomes todos os anos, e por isto não estão todos na base Repantino. Como no Brasil não existe nenhuma lei que nos impeça de colocar algum nome, a retirada dos mesmos acaba sendo mais trabalhosa, contando com todas as variações existentes, realmente o trabalho manual não daria conta de limpar todas as evoluções.

Portanto, esse trabalho visa tanto a desidentificação de prontuários quanto a criação de uma lista de nomes utilizando não só a leitura e adição manual de nomes às listas, mas também utilizando o Word Embedding. Além disso, essa lista se destina a tarefa de desidentificação de prontuários médicos, tratando alguns problemas específicos dessa área.

III. MATERIAIS E MÉTODOS

Nesta seção, os recursos utilizados no desenvolvimento do algoritmo são apresentados.

A. Evoluções Clínicas

Evolução clínica são os dados recebidos sobre cada paciente durante a sua estadia no Hospital, como por exemplo,

¹<http://www.irn.mj.pt>

os exames que o paciente realizou, os resultados desses exames, a informação de qual médico cuidou do paciente, se houve participação de algum residente, todas as informações são detalhadas. Neste artigo, os prontuários utilizados foram cedidos pelo Hospital Nossa Senhora da Conceição, a fim de estudar e melhorar a estadia do paciente. Existem 10 anos de evoluções registradas em prontuário eletrônico, o que resulta em um total de 5 milhões de evoluções. Entretanto, neste trabalho foram utilizadas 3 mil evoluções.

B. Repentino

Repentino é uma lista de Entidades Nomeadas composta por locais, nomes e organizações Portuguesas. Conforme descrito na Seção II, o Repentino foi criado em 2006 por [5].

C. Word Embeddings

O Word Embedding (WB) [6] compreende um conjunto de técnicas de modelagem de linguagem e de aprendizado de recursos, em que palavras ou frases do vocabulário são mapeadas para vetores de números reais. Os métodos para gerar este mapeamento incluem redes neurais, redução de dimensionalidade na matriz de coocorrência de palavras, modelos probabilísticos, método base de conhecimento explicável e representação explícita em termos do contexto no qual as palavras aparecem.

a) *Lista_WB*: Lista de palavras do Word Embeddings que apareceram em no mínimo 100 evoluções. Essa lista é utilizada para verificar se o termo é uma palavra existente em um dicionário “comum”. Exemplos de termos: enfermeira, emergência, paciente.

IV. ABORDAGEM PROPOSTA

Nesta seção, a abordagem proposta é apresentada em etapas, para um melhor entendimento do processo de construção das listas de nomes.

a) *Etapa 1: Uso do Repentino*: Iniciamos nosso trabalho utilizando a lista de nomes Repentino, a qual foi de grande importância no processo. O Repentino contém 3797 nomes, sem contar as Organizações e os Locais que nele constam, os quais não foram utilizados neste trabalho. A partir dessa lista de nomes, conseguimos ir adaptando-a e criando a nossa própria lista.

b) *Etapa 2: Nomes Novos e Removidos*: Nesta etapa criamos duas novas listas: a primeira é a NOMES_NOVOS, a qual teve como base a lista Repentino e foram adicionados nomes que apareciam ao longo da leitura das evoluções, tanto nome de pacientes quanto de médicos e enfermeiras. A partir dessa lista, criou-se a segunda lista de NOMES_REMOVIDOS em que nomes identificados como doenças, itens cirúrgicos, ou palavras que normalmente são utilizadas para fins normais, como “clara”, que é usada tanto como um nome como para examinar a funcionalidade dos rins do paciente (“urina clara”) foram removidos. Como a maior parte da criação das listas foi manual, eram necessários estes procedimentos, pois a criação das listas se deu a partir da leitura dos prontuários e da observação de nomes existentes nos mesmos.

Tendo em vista que a lista Repentino é feita de nomes portugueses de Portugal, haviam múltiplos “nomes” que no Brasil não são utilizados para a mesma função, como por exemplo, “paredes” e “motorista”. Como essas palavras não são utilizadas aqui como nomes, elas entram na lista NOMES_REMOVIDOS. Como dito anteriormente, pela lista ser composta de nomes portugueses, haviam nomes muito comuns entre os brasileiros que não se encontravam na lista, por exemplo, “Lúcia”, “José” e “Luís”, os quais foram adicionados à lista NOMES_NOVOS.

A Tabela I apresenta exemplos de nomes novos e nomes removidos.

Nomes Novos	Nomes Removidos
José	Clara
Wevillim	Ramos
Santa	Lobo

Tabela I
NOMES NOVOS E REMOVIDOS.

c) *Etapa 3: Word Embedding*: Nesta etapa, temos a incorporação do Word Embedding (WB) ao processo, porém constatamos que o mesmo estava adicionando muita “sujeira” para a nossa lista e assim dificultando a desidentificação correta das palavras. Com a descoberta deste problema, resolvemos interromper o uso do Word Embedding, mas antes disso criamos 2 listas para filtrar as palavras indesejadas e incorretas. A primeira lista foi a LISTA_NOMES_EXCEÇÃO que possuia algumas “palavras” que não eram nomes, como por exemplo, “tidra” ou “tabada”; “carolinarafael” que até é um nome, mas está com erro de digitação. Essa lista possui um total de 70 nomes. A segunda lista foi a LISTA_TERMOS_EXCEÇÃO que foi criada para retirar alguns termos que não estavam na Lista_WB e acabaram sendo adicionados pelo WB, como por exemplo, “noradrenalina”, “tylex” e “Microvlar” que são medicamentos e não devem ser retirados de um prontuário médico. Essa lista possui um total de 63 nomes. A Tabela II apresenta exemplos de nomes e de termos removidos.

Após o uso de filtros para o nosso WB, obtivemos resultados, partidos de um percentual de 89% de similaridade semântica. Em alguns casos existiam mais de um nome similar com percentual acima do estipulado, conforme ilustrado na Tabela III.

d) *Etapa 4: Nomes WB*: A partir das listas criadas na Etapa 2 e na Etapa 3 conseguimos iniciar nosso processo de desidentificação das evoluções, em que incluímos nomes como “Maria”, “Joana” e “Gabrielly”. Com a lista NOMES_REMOVIDOS, corrigimos alguns termos que não

Nomes Exceções	Termos Exceções
Carolinarafael	Piurica
Tidra	Pacientes
Tabada	Noradrenalina
Francielledda	Microvlar
Felix	Tylex
Caneda	Lasegue

Tabela II
NOMES E TERMOS DE EXCESSÕES DO WB.

Nome	Nomes similares	Similaridade
Mariana	Jade	0.92
	Paola	0.90
Claudio	Gilberto	0.90
	Josefa	0.93
Josefina	Edi	0.93
	Olga	0.92
Karoline	Olga	0.90
Karoline	Calvett	0.94
Renata	Vanessa	0.90

Tabela III
SIMILARIDADE SEMANTICA NO WB

deveriam ser retirados, como por exemplo, “lobo”, “clara” e “equipe”, uma vez que “lobo” foi retirado por ser comumente encontrado como “lobo frontal”, “clara” foi retirada por se referir a resultado de exames de urina. A lista NOMES_WB possui um total de 612 nomes.

Por fim, o resultado de todas estas etapas e procedimentos realizados geraram a LISTA_FINAL que é a soma das listas Repentino, Nomes Novos e Nomes WB. A Tabela IV apresenta os resultados de cada etapa.

Etapas	Listas	Total
1	Repentino	3797 nomes
2	Nomes Novos	3867 nomes
2	Nomes Removidos	56 nomes
3	Nomes Exceção	70 nomes
3	Termos Exceção	63 nomes
4	Nomes WB	612 nomes
Resultado	Final	4530 nomes

Tabela IV
RESULTADOS DE CADA ETAPA

O resultado da lista final pode ser observado na Tabela V, que mostra a diferença da quantidade de nomes existente na base Repentino e na base adaptada. Como a base Repentino faz uso da lista de nomes fornecida pelo Instituto dos Registos e do Notariado de Portugal², ela tem um grande número de nomes, visto que é a base de dados que os portugueses possuem para dar nome aos seus filhos. Lembrando que Portugal e Brasil, mesmo usando a mesma língua, possuem diferenças de escrita e de entendimento, algumas mudanças e muitos acréscimos foram realizados, resultando na nova lista, e nos novos números, da base adaptada.

Description	#
Repentino	3797
Base adaptada	4466

Tabela V
TOTAL DE NOMES

A. Processo de remoção baseado em listas

Para a realização da desidentificação de um prontuário é necessário aplicar um removedor de acentos em todas as palavras que o compõem, já que as listas dos nomes estão sem acentos. Na sequência, para cada palavra do prontuário é verificado se a mesma existe na LISTA_WB, lista de

²<http://www.irn.mj.pt>

palavras que não devem ser retiradas (veja seção 3.3). Em caso positivo, a palavra é descartada e a próxima palavra é automaticamente analisada. Caso contrário, a LISTA_FINAL é percorrida na busca da palavra em foco e se encontrada é substituída por “NoInfo”. Além disso, aplica-se mais um processo de verificação de parte da palavra, pois como já foi dito anteriormente, existem palavras sem espaçamento, como “EmegenciaMaria”. Então é verificado se parte desta palavra é um nome, caso seja, essa parte é substituída por “NoInfo” e o restante da palavra permanece no local devido. Após o processo, os acentos são devolvidos às palavras.

Para um melhor entendimento do processo proposto, um exemplo de prontuário médico desidentificado é apresentado a seguir:

“R104 - (Dor abdominal e pelvica) Outras dores abdominais e as nao especificadas Evolução HNSC: *****enfermagem****paciente MUITO AGITADA PELA MANHÃ. RECEBEU TODAS MEDICAÇÕES PRESCRITAS SEM EFEITO CONSIDERÁVEL SIC NA PASSAGEM DE PLANTÃO. MÃE ENTENDEU NÃO SER POSSÍVEL LEVÁ-LA A CONSULTA NA SANTA CASA PELA AGITAÇÃO ATUAL. LIGO PARA GASTRO E DEIXO RECADO COM RESIDENTE NoInfo. Evoluído por: XXXXX em XX/XX/XX às XX:XX.”

O processo mais demorado de todo nosso trabalho foi a parte de análise manual dos prontuários, na qual podemos analisar as palavras que foram ou não retiradas, se estão corretas ou se precisam de alguma correção. Como previsto, no início do nosso trabalho o que mais ocorriam eram erros como a retirada de termos médicos e a não retirada de alguns nomes. Entretanto, ao transcorrer do desenvolvimento do processo, conseguimos corrigir a maioria dos problemas, mas nem todos foram resolvidos, como podemos observar na próxima seção.

B. Limitações

Algumas das limitações do nosso algoritmo são as palavras que contém nomes, como por exemplo, a palavra “periférico” em que o algoritmo retiraria o nome “érico”, ou em “oscarbazepina” em que é retirado o nome “oscar”. Nesse último caso, o erro é causado por grafia incorreta, pois o nome original do remédio seria “oxcarbazepina”. A solução encontrada para esses problemas foi inserir tais palavras no dicionário, porém é humanamente impossível prever ou inserir todas as palavras que vão sofrer erros de grafia, ou seja, é um ponto que não temos como considerar concluído.

Outra limitação foi a retirada erroneamente de informações dos leitos e números de registros dos pacientes, pois nosso algoritmo seleciona 5 ou mais caracteres e os retira, consequentemente, algumas palavras que se encontram agrupadas com os números são retiradas. Por exemplo, “1234567Lista” em que o algoritmo substitui tudo por “NoInfo”.

Alguns nomes continuam nas evoluções, como “Clara”, devido ao fato de que ao mesmo tempo a palavra “clara” se referia tanto à resultados de exames de urina quanto a algum paciente, porém não temos como retirar resultados de exames sem prejudicar a evolução, então optou-se por manter as duas

variações do nome. Existem outros nomes que não foram retirados das evoluções, pois estão agrupados com outras palavras, como por exemplo, “EmergênciaMaria”, que acaba formando uma nova palavra (a junção de duas palavras) e que, por recorrentes erros de digitação acaba obtendo uma frequência alta, e consequentemente entra no dicionário, o que impossibilita a retirada dos nomes.

V. COMPARAÇÃO COM ALGORITMO DE RECONHECIMENTO DE ENTIDADES NOMEADAS

Como forma de observação de eficiência do algoritmo proposto, realizamos um comparativo dos nossos resultados com o NERP-CRF, que é um sistema de Reconhecimento de Entidades Nomeadas do Português descrito em [7]. O NERP-CRF aplica o método probabilístico Conditional Random Fields (CRF) para a identificação e posterior classificação das entidades nomeadas do Português, como nomes de Pessoas, Organizações, Locais, entre outros. Neste trabalho, o NERP-CRF foi utilizado para desidentificação do corpus da Coleção Dourada do Segundo HAREM³ e de uma amostra de evoluções médicas.

Para realizar o comparativo com as evoluções médicas, separamos uma amostra de 50 evoluções entre as 3 mil evoluções existentes, de todos os tipos, e executamos nessas evoluções o algoritmo proposto e o NERP-CRF. Dentro destas 50 evoluções, tivemos um total de 106 nomes identificados manualmente que serviram de referência para a avaliação. Uma análise dos resultados foi realizada e a partir dela foram contabilizados os acertos (nomes que deveriam ser retirados), os erros (palavras retiradas incorretamente) e os nomes não identificados. Os resultados são ilustrados na Tabela VI.

A execução da nossa ferramenta obteve 80 acertos, 9 erros e 26 nomes não identificados. Para exemplificar, temos alguns casos de acertos como “Gustavo”, “Priscila”, “Antonio”; um exemplo de erro como “Lutero” que é de “fisio utero” que é um tipo de fisioterapia porém existe um “l” entre as palavras que pode se referir tanto a fisiologia como pode ser somente um erro de digitação, e alguns exemplos de nomes não identificados como “Daphnise”, “Ayres”, “Zaleski”.

Quando utilizamos o NERP-CRF nas evoluções, ele apresentou os seguintes resultados: 61 acertos, 170 erros e 37 nomes não identificados. Alguns exemplos de erros comuns nessa etapa foram: “Dor”, “Tuberculose”, “DESCRIÇÃO”. Como os textos jornalísticos possuem um formato diferente em relação as evoluções, em que a maioria dos nomes estão bem sinalizados com letra maiúscula no inicio da palavra, este seria um dos motivos do NERP-CRF ter identificado erroneamente esses casos já que as evoluções não possuem esse padrão.

O algoritmo proposto foi também executado nos textos da Coleção Dourada do Segundo HAREM, porém como o formato dos textos são diferentes, foi necessário desenvolver um pequeno código para conseguir coletar todos os dados e verificar os resultados. O total de nomes existentes nos textos era de 3060 nomes sendo eles da categoria Pessoa e

tipo Individual. Como resultado, obtivemos 1297 acertos, 1763 nomes não identificados e 1367 erros (veja Tabela VI).

O NERP-CRF foi também aplicado nos textos da Coleção Dourada do Segundo HAREM. Entretanto, ao realizar a verificação dos textos do HAREM no NERP-CRF verificou-se a necessidade do desenvolvimento de uma ferramenta para realizar a comparação entre os textos.

Obtivemos vários problemas ao executar o NERP-CRF devido ao formato dos textos jornalísticos terem uma organização e os prontuários terem outra totalmente diferente. Por exemplo: os caracteres "<" e ">" no NERP-CRF são tidos como um "parenteses" para informações descartáveis, já nas evoluções eles são usados para demonstrar resultados de exames, como por exemplo, "pH > 3". Este fator gerou muitos problemas, mas o que mais gerou erro foi o fato de que o NERP-CRF identifica um nome sempre sendo inicializado por letra maiúscula, e algumas vezes até a palavra inteira em maiúscula, como por exemplo, muitas das identificações erradas foram da palavra “Dor”, ou então “CAT”, porém, pensa-se que todos os nomes inicializados da forma esperada foram identificados corretamente mas isso não aconteceu. Alguns dos nomes eram precedidos de “#”, como por exemplo, “##Luis”. Outros casos de não identificação foram “Antonio”, “BIANCA”, “Francisco”.

Textos	Algoritmo	NERP-CRF
Evolução	80 acertos	61 acertos
	9 erros	170 erros
	26 não ID	37 não ID
HAREM	1297 acertos	1810 acertos
	1367 erros	1246 erros
	1762 não ID	1250 não ID

Tabela VI
RESULTADOS COMPARATIVOS.

Com estes resultados, conseguimos observar que o propósito e o desenvolvimento do algoritmo realmente conversaram e foi obtido um bom resultado em relação aos objetivos do trabalho, já com os textos que não faziam parte do projeto, os resultados foram ruins, porém como a ferramenta não foi desenvolvida com esse objetivo, era de se esperar um resultado similar a este.

VI. CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho, nós desenvolvemos uma lista de nomes e um algoritmo para a desidentificação de prontuários médicos, a partir de listas de nomes pré-existentes como a criada por Sarmento [5]. As evoluções, provenientes do Hospital Nossa Senhora da Conceição, foram usadas no estudo.

Na comparação de ferramentas, percebemos que quando utiliza-se uma ferramenta que não foi desenvolvida para tal trabalho, os resultados muito provavelmente serão ruins, e que isto só prova o quanto este trabalho foi necessário.

Como trabalho futuro, temos o deslocamento de datas. Já que o leito, registro e os nomes, tanto de pacientes quanto de médicos, já estão sendo retirados, somente com a data e hora do atendimento, seria muito difícil de acessar os dados do

³<https://www.linguateca.pt/HAREM/>

hospital e descobrir quem era o paciente da evolução, e sendo assim, o deslocamento de datas ficou como segundo objetivo.

REFERÊNCIAS

- [1] M. B. Buntin, M. F. Burke, M. C. Hoaglin, and D. Blumenthal, “The benefits of health information technology: a review of the recent literature shows predominantly positive results,” *Health affairs*, vol. 30, no. 3, pp. 464–471, 2011.
- [2] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” *Nature reviews. Genetics*, vol. 13, no. 6, p. 395, 2012.
- [3] I. Neamatullah, M. M. Douglass, H. L. Li-wei, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford, “Automated de-identification of free-text medical records,” *BMC medical informatics and decision making*, vol. 8, no. 1, p. 32, 2008.
- [4] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, “Automatic de-identification of textual documents in the electronic health record: a review of recent research,” *BMC medical research methodology*, vol. 10, no. 1, p. 70, 2010.
- [5] L. Sarmento, A. S. Pinto, and L. Cabral, “Repentino—a wide-scope gazetteer for entity recognition in portuguese,” in *International Workshop on Computational Processing of the Portuguese Language*. Springer, 2006, pp. 31–40.
- [6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [7] D. O. F. do Amaral and R. Vieira, “Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields.” *Linguamática*, vol. 6, no. 1, pp. 41–49, 2014.

Do PDF ao TXT: Desafios na extração de informação em textos técnico-científicos

Aline Silveira

Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
silveira26aline@gmail.com

Elvis de Souza

Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
elvis.desouza99@gmail.com

Tatiana Cavalcanti

Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
tatiana.shc@hotmail.com

Cláudia Freitas

Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
claudiafreitas@puc-rio.br

Index Terms—Extração de informação, Documentos técnico-científicos, Pré-processamento, Processamento de Linguagem Natural

I. INTRODUÇÃO

Extração de Informação (EI) é o processo de examinar um texto automaticamente visando a capturar informações relevantes para determinado interesse. Em outras palavras, o objetivo da extração de informação é transformar textos, que são fonte de dados não estruturados, em informação organizada de acordo com certos critérios de importância, deixando o que não interessa para trás [1]. No tocante ao modelo de texto científico, tal técnica se mostra relevante porque pode revelar tendências de pesquisa científica ao longo do tempo, concatenar informações de fontes muito vastas em volume e diversidade, evidenciar citações ausentes e proximidade temática entre os autores, e auxiliar na construção de bancos de dados a partir de grandes corpora textuais, entre outras aplicações. Nossa objetivo é construir um grande corpus de textos técnico-científicos de domínio específico com suas informações estruturadas e prontas para processamento futuro, tendo em vista facilitar buscas semânticas no conteúdo; para isso, EI se torna fundamental.

O Processamento de Linguagem Natural (PLN) viabiliza a conversão de um texto cru, essencialmente uma sequência aleatória de bits digitais, em uma sequência bem definida de unidades linguísticas detentoras de sentido [2]. Essa definição de caracteres, palavras e sentenças é fundamental porque, uma vez feita, serve de entrada para uma outra tarefa crucial dentro do PLN, a anotação morfossintática. A anotação morfossintática, por sua vez, pode facilitar a Extração de Informação, porque oferece um guia prévio de generalizações para o sistema ao dar o primeiro passo de reconhecimento de padrões linguísticos.

Nosso interesse na EI é resultado da necessidade de criar um corpus para extrair informações de documentos técnico-científicos em português, no domínio de petróleo. Neste artigo, relatamos o que se tem feito recentemente no campo da EI voltada ao domínio técnico-científico e quais são as nossas questões práticas. Esperamos, futuramente, poder relatar nossos resultados empíricos e, assim, contribuir de forma

relevante para o estado da arte em extração de informação em textos técnico-científicos.

II. DO PDF AO TXT: DESAFIOS DO PRÉ-PROCESSAMENTO

Para que qualquer texto de língua natural seja legível por uma máquina, são necessários alguns passos iniciais. O primeiro é garantir que todos os textos de que dispomos têm seus caracteres codificados da mesma forma (em Latin-1, UTF-8, CP1252, etc.). A essa primeira tarefa dá-se o nome de identificação de codificação de caractere. O segundo passo é a identificação da língua em que o texto está escrito. E o último grande passo é descartar informações indesejáveis como imagens, tabelas, cabeçalhos, links e notas de rodapé. Ao final desse processo, o que se tem é um texto bem definido, organizado pela sua língua, pronto para ser segmentado e analisado em níveis mais complexos [3].

Como já foi visto, alguns detalhes podem atrapalhar a leitura da ferramenta computacional utilizada e, por conseguinte, influenciar no resultado final da análise do texto. Alguns deles, por exemplo, no caso dos textos técnicos-científicos da categoria dissertação ou tese, são o sumário, os agradecimentos, a folha de aprovação, listas em geral, figuras e tabelas. Esses elementos, juntamente com paginação e notas de rodapé, quando transformados em texto plano (formato .txt), apresentam deformação e rompem com a linearidade dos textos. A figura 1, obtida a partir de um dos documentos de nosso corpus, ilustra a questão das notas de rodapé. Como podemos observar, a palavra “natural” se transforma em outra palavra (“natural2”). Ou seja, se “natural” for uma palavra relevante, perderíamos essa ocorrência e esse contexto. A figura 2 (oriunda do corpus Brasileiro e acessada pelo serviço AC/DC [4]), ilustra que, sem o cuidado necessário no pré-processamento, o que seria uma frase se transforma em blocos de texto contendo números e tabelas que, mais tarde, serão erroneamente processados como períodos únicos, dificultando etapas posteriores do processamento automático de texto.

Isso acontece pois esses segmentos são, normalmente, caracterizados pelos elevados custos de constituição das redes de gasodutos o que, na maioria das vezes, torna o monopólio a solução econômica mais viável. Isso significa que a atividade é um monopólio natural².

Isso acontece pois esses segmentos são, normalmente, caracterizados pelos elevados custos de constituição das redes de gasodutos o que, na maioria das vezes, torna o monopólio a solução econômica mais viável. Isso significa que a atividade é um monopólio natural¹²

Figura 1. Exemplo de distorção relativa a notas de rodapé.

<p>: Gêneros N°de ocorrências 1 Freqüência 2 Penicillium spp 50 53,161 % Aspergillus spp 19 41,333 % Cladosporium spp 10 4,534 % Rhizopus spp 1 0,343 % Fusarium spp 5 0,215 % Aureobasidium spp 1 0,143 % Chrysosporium spp 5 0,100 % Alternaria spp 4 0,057 % Epicoccum spp 1 0,029 % Helmimthosporium spp 1 0,029 % Geotrichum spp 1 0,014 % Gliocladium spp 1 0,014 % Nigrospora spp 1 0,014 % Rhizomucor spp 1 0,014 % 1 Refere-se ao número de vezes que um determinado gênero foi identificado independentemente das contagens de UFC .

Figura 2. Exemplo de distorção relativa a tabelas.

O trabalho de [5], também voltado para a constituição de um corpus de documentos técnico-científicos, enfrentou - e resolveu - vários dos problemas mencionados aqui, ainda que nem todas as soluções utilizadas se adequem aos nossos propósitos. Reconhecida como a etapa "mais importante e mais trabalhosa" na constituição de um corpus que, posteriormente, será anotado morfossintaticamente, as autoras utilizaram editores de texto e expressões regulares, associados à supervisão humana, para eliminar ruídos associados a representações numéricas, o uso de ponto final em nomes próprios como *Dra.*, e asteriscos associados a palavras. Por outro lado, elementos que nos interessa manter, como títulos e referências bibliográficas ao longo do texto, foram excluídas. E um dos nossos principais problemas, em função do tipo de texto, não parece ter sido um problema: as enumerações itemizadas, retratadas na figura 2.

Em suma, não faltam desafios ao pré-processamento de textos. Tendo essa etapa sido concluída corretamente, ou seja, uma vez que tenhamos um corpus de qualidade, as etapas subsequentes do processamento automático podem acontecer da melhor maneira possível.

III. EI EM DOCUMENTOS TÉCNICO-CIENTÍFICOS: SEMEVAL 2017 E 2018

Para otimizar a eficiência em qualquer trabalho de investigação, é imprescindível fazer uma busca para que se conheça o que já foi feito no campo e que tipo de conclusões já foram tiradas. Considerando que o nosso interesse de pesquisa reside em textos técnico-científicos, fez sentido analisar, especificamente, as tarefas apresentadas nas avaliações SemEval de 2017 e 2018. Por meio dessas competições e a partir de suas várias tarefas e subtarefas, que abarcam diversos temas e objetivos específicos, os participantes têm a oportunidade de testar seus sistemas e avaliar sua posição em relação aos outros da mesma área. Percebemos que, ao observarmos os métodos, as escolhas, as conclusões e os resultados obtidos na avaliação, ganhamos um melhor entendimento acerca de nosso próprio trabalho, podendo aproveitar o que deu bons frutos e descartar o que falhou.

Em particular, vale descrever as tarefas de 2017 e 2018 individualmente. No SemEval de 2017, a tarefa 10 (ScienceIE - Extracting Keyphrases and Relations from Scientific Publications [6]) foi a de identificar termos-chave, classificá-los e relacionar os de mesma classificação a partir de um único parágrafo de publicações científicas das áreas de Ciência da Computação, Ciência de Materiais e Física. Um dos maiores objetivos dos desenvolvedores da tarefa era o de atender aos editores de publicações científicas, tendo em vista que com essa tarefa eles poderiam recomendar artigos aos leitores, identificar revisores em potencial para submissão e analisar tendências de pesquisa ao longo do tempo.

Para realizar tal tarefa foi necessário subdividi-la em três subtarefas: a primeira foi a de identificação dos termos-chave - subtarefa A; a segunda, de classificação desses termos em 3 categorias diferentes (PROCESS, TASK e MATERIAL) - subtarefa B, e a terceira, de classificação do tipo de relação entre os termos de mesma categoria em sinônimo ou hiperônimo - subtarefa C. É importante ressaltar que havia três cenários de avaliação diferentes, um com o texto cru, simplesmente – necessitando que as três subtarefas fossem realizadas sem nenhum ponto de partida; outro com a identificação manual prévia dos termos-chave - ou seja, com a subtarefa A já realizada previamente; e um terceiro cenário com a identificação manual prévia dos termos-chave e suas classificações – ou seja, com as subtarefas A e B dadas previamente. Os sistemas que participaram da competição variaram muito em técnicas, desde redes neurais até métodos baseados em regras. Isso mostra a diversidade de maneiras de resolver esta tarefa. Para a anotação do corpus que serviu de treinamento e avaliação aos sistemas foram recrutados estudantes de graduação dos mesmos domínios dos textos e professores dessas mesmas áreas. A concordância entre eles oscilou entre 45% e 85%, sendo metade dos casos com concordância maior ou igual a 60%. Tal fato mostra a dificuldade humana em detectar termos-chave, suas classificações e relações, mesmo quando os anotadores são especialistas no domínio. Do mesmo modo, a observação da concordância entre os anotadores nos informa o grau de dificuldade que a máquina enfrenta ao realizar essas tarefas, evidentemente complexas.

Vejamos os 3 cenários diferentes em que as avaliações ocorreram. No primeiro cenário de avaliação, com o texto cru, participaram 17 sistemas, dos quais o maior pontuador teve um desempenho de 43% de acerto (F1) e utilizou o método de rede neural recorrente. Já no segundo cenário de avaliação, em que apenas as subtarefas B e C deveriam ser resolvidas, somente 4 competidores participaram. O ganhador teve um acerto (F1) de 64%, e fez uso de classificadores com características lexicais, características ortográficas e n-gramas para a subtarefa B. Para a subtarefa C, o ganhador fez uso de um sistema baseado em regras na parte de sinônimos e de padrões da Hearst [7] para detectar hiperônimos. No último cenário, no qual a única subtarefa a ser realizada foi a C - relacionar os itens de mesma classificação -, a melhor atuação - dentre as 5 participantes - foi do sistema de redes neurais convolucionais que alcançou a marca de 64%, tal como no cenário anterior. Realça-se aqui

o salto de desempenho entre os cenários, que se relaciona diretamente com a importância da precisão da subtarefa A para o sucesso das subtarefas consecutivas: isto é, identificar quais são os termos-chave parece ser a parte mais difícil de toda a tarefa.

Essa tarefa da competição SemEval de 2017 mostra a variação metodológica com que sistemas podem operar. Por outro lado, mostra também que ainda há um grande espaço para avanços, já que 64%, embora não seja um índice baixo, pode ser insatisfatório do ponto de vista prático, isto é, de quem irá tirar proveito dos resultados.

No ano seguinte, no SemEval 2018, a tarefa 7 (Semantic Relation Extraction and Classification in Scientific Papers [8]) foi a de identificar e classificar – entre 6 categorias pré-determinadas (USAGE, RESULT, MODEL, PART_WHOLE, TOPIC, COMPARISON) – as relações semânticas entre entidades presentes nos resumos de artigos científicos. Os artigos eram todos pertencentes ao domínio da Linguística Computacional. Segundo o artigo que descreve a tarefa e seus resultados, um de seus objetivos seria melhorar o acesso à literatura científica, atendendo a uma necessidade de informação que não estaria sendo suprida pelas ferramentas padrão de pesquisa, nem pelos humanos especialistas nos domínios específicos, que, muitas vezes, não dispõem do tempo necessário para se atualizar em relação aos avanços científicos nas suas áreas. Nesse caso, seriam de grande ajuda sistemas que estruturassem essas informações automaticamente.

A partir dessa tarefa principal, definiram-se três subtarefas: a classificação de relações em clean data (1); a classificação de relações em noisy data (2); e a identificação e a classificação de relações em material não previamente anotado (3), exceto pelas entidades, que estão anotadas em todos os experimentos. Ao todo, 32 times participaram de ao menos uma das subtarefas.

Na primeira subtarefa (clean data), as entidades cujas relações serão classificadas tinham sido previamente anotadas manualmente, enquanto na segunda (noisy data), isso havia sido feito de forma automática. Desse modo, a terceira subtarefa é aquela mais sujeita a dificuldades, já que as relações não só precisam ser categorizadas, mas também previamente identificadas pelos sistemas competidores, que não contam com as regras das duas primeiras subtarefas.

Para a anotação do corpus, foram recrutados anotadores especialistas dentre os membros da organização da tarefa, assim como estudantes de PLN. A concordância inter-anotadores relativa à rotulação das classes semânticas foi de 90.8%, tendo sido calculada entre dois anotadores a partir de uma amostra de 150 resumos provenientes da primeira subtarefa. É importante acentuar aqui o valor inegável de uma alta concordância entre anotadores, uma vez que um “golden” construído com base em divergências está muito mais sujeito a inconsistências.

Ao final da competição, concluiu-se que o maior desafio dentre as etapas do processamento era a identificação adequada das relações semânticas. Isso fica mais claro ao observarmos os melhores valores de F1 para cada subtarefa (Tabela 1), já que observamos uma queda na performance dos sistemas

Tabela I
RESUMO ENTRE OS RESULTADOS DAS COMPETIÇÕES DE EXTRACÃO DE INFORMAÇÃO.

Competição	Subtarefas	F1
SemEval 2017 Tarefa 10	A identificação dos termos-chave	43%
	B classificação dos termos-chave	64%
	C classificação das relações entre termos-chave	64%
SemEval 2018 Tarefa 7	1 classificação das relações em clean data	81.7%
	2 classificação das relações em noisy data	90.4%
	3 extração e classificação das relações	49.3%

na terceira subtarefa, em que a presença de relações entre entidades não era dada anteriormente.

Ademais, ao tratar dos tipos de relações semânticas, afirma-se que mais do que o significado dessas relações, a maior dificuldade é distribuição desigual das identificações de entidades nas subtarefas com clean e noisy data. As entidades anotadas manual e automaticamente foram de diferentes naturezas: os seres humanos eram orientados a anotar termos e expressões mais complexas, enquanto a máquina anotava termos menores e com um nível mais baixo de especificidade.

Destaca-se, por fim, a atualidade dos trabalhos objetivados pelo SemEval, que acompanham as pesquisas correntes e espelham os tópicos que se sobressaem em determinado período. Competições como essa não só estimulam seus integrantes a aprimorarem seus sistemas e a desenvolverem suas pesquisas, como também nos ajudam a entender como diferentes grupos estão pensando o mesmo assunto.

IV. LIÇÕES APRENDIDAS

Tendo em vista tudo o que foi exposto aqui, nota-se o grau de dificuldade das tarefas a que as máquinas estão submetidas. Se para nós, humanos, mesmo diante do nosso vasto conhecimento sobre língua e *expertise* nas temáticas específicas, definir os limites de palavras e expressões e se elas podem ser consideradas relevantes ou não em determinado contexto é tarefa árdua, quiçá para sistemas de processamento automático de texto, que, segundo os resultados das competições analisadas, não chegam a 80%.

No entanto, é preciso retomar aqui que em nenhuma das tarefas apresentadas foi necessário um pré-processamento detalhado dos textos analisados. Ou seja, nosso objetivo de aprender sobre os desafios do pré-processamento de documentos técnico-científicos com a experiência de outros trabalhos foi parcialmente alcançado, já que no SemEval de 2017 o corpus é composto de parágrafos, e no de 2018, de resumos. Com essa decisão, os organizadores evitaram todo o trabalho

anterior de preparação dos textos. Em nosso caso, por outro lado, iremos trabalhar com documentos completos, o que não permite essa abordagem imediata. A exceção cabe ao trabalho de [5], que se debruçaram sobre algumas das questões que também nos inquietam. No entanto, as autoras salientam a necessidade de algum acompanhamento humano (trata-se de um processamento semi-automático).

Um possível caminho a ser tomado é descartar a ideia de um processamento de texto integral e trabalhar apenas com fragmentos, como fizeram as tarefas que descrevemos. O quanto se ganha ao processar documentos completos? O quanto se perde processando apenas resumos ou parágrafos? Aliás, perdemos alguma coisa? É um interesse nosso avaliar a diferença na qualidade da informação extraída automaticamente em documentos completos ou em fragmentos especialmente relevantes, como resumos. No entanto, a única maneira de verificar isso é comparando as duas situações, o que envolve o processamento dos documentos na íntegra. Ao que parece, precisaremos desenvolver nossas próprias soluções para o pré-processamento de documentos técnico-científicos.

AGRADECIMENTOS

Este projeto foi financiado com o apoio da ANP - Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associado ao investimento de recursos oriundos das Cláusulas de P,D&I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio.

REFERÊNCIAS

- [1] J. R. Hobbs and E. Riloff, "Information extraction," in *Handbook of Natural Language Processing*, N. Indurkhy and F. J. Damerau, Eds. Chapman & Hall/CRC, 2010.
- [2] D. Palmer, "Text preprocessing," in *Handbook of Natural Language Processing*, N. Indurkhy and F. J. Damerau, Eds. Chapman & Hall/CRC, 2010.
- [3] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, Jun. 1999, pp. 3–10. [Online]. Available: <https://www.aclweb.org/anthology/P99-1001>
- [4] D. Santos and E. Bick, "Providing Internet access to Portuguese corpora: the AC/DC project," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: <http://www.lrec-conf.org/proceedings/lrec2000/pdf/85.pdf>
- [5] L. Lopes and R. Vieira, "Building domain specific parsed corpora in portuguese language," in *Proceedings of the 10th National Meeting on Artificial and Computational Intelligence (ENIAC)*, Fortaleza, Brasil, 2013.
- [6] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum, "SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 546–555. [Online]. Available: <https://www.aclweb.org/anthology/S17-2091>
- [7] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, ser. COLING '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 539–545. [Online]. Available: <https://doi.org/10.3115/992133.992154>

- [8] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, and T. Charnois, "SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 679–688. [Online]. Available: <https://www.aclweb.org/anthology/S18-1111>

Identificação Automática de Erros em Sumários Multidocumento

Henrique Papa A. Fonseca

Universidade Federal de Goiás
Unidade Acadêmica de Biotecnologia
Email: henrique_ahg@hotmail.com

Márcio de Souza Dias

Universidade Federal de Goiás
Unidade Acadêmica de Biotecnologia
Email: marcosouzadias@ufg.br

Nádia Félix Felipe da Silva

Universidade Federal de Goiás
Instituto de Informática
Email: nadia.felix@ufg.com

I. INTRODUÇÃO

Sem dúvidas, a capacidade de produzir, receber e assimilar informações é uma das características que contribuíram para a evolução da espécie humana. É possível perceber que desde as pinturas rupestres¹, realizadas no período Paleolítico, ou desde a invenção da escrita na antiga Mesopotâmia, as civilizações vêm buscando meios para transmitir conhecimentos uns aos outros. Sabe-se então que essa procura por novos meios de comunicação é constante, sendo um fator essencial para o desenvolvimento da humanidade.

Ao observar a evolução nos meios de comunicação, pode-se afirmar que a Internet revolucionou a maneira como as pessoas se relacionam. A Internet desenvolveu-se a partir da década de 1960 e trata-se de um sistema que interconecta diversos computadores em proporções mundiais, atingindo milhões de pessoas [1]. Devido às suas funcionalidades, essa ferramenta intensificou um fenômeno denominado globalização, caracterizado pela aproximação entre as diversas sociedades e nações, seja no âmbito econômico, social, cultural ou político [2].

Devido ao fenômeno supracitado e à grande quantidade de informações disponíveis na rede mundial de computadores, surgiu a necessidade de desenvolver tecnologias capazes de auxiliar o ser humano na assimilação de conteúdo. Assim, a partir desse contexto, a Sumarização Automática Multidocumento (SAM) surgiu e vem ganhando destaque na comunidade científica, pois esse processo é capaz de agrupar as informações contidas em diversos textos que tratam sobre o mesmo assunto em apenas um [3].

Contudo, apesar dos avanços alcançados, os summarizadores automáticos ainda não tratam de forma satisfatória os aspectos linguísticos que afetam a coesão e a coerência textual [4]. Os problemas linguísticos mais recorrentes estão relacionados à falta de pontuação, uso de sentenças muito longas e uso inadequado de parênteses ou outros elementos textuais [5]. Dias [6], por sua vez, identificou manualmente 12 tipos de erros linguísticos presentes em um corpus² de sumários automáticos multidocumento. Os erros identificados por Dias

(2016) estavam relacionados a menções de entidades, transgressões gramaticais e redundância.

Certamente, esses problemas linguísticos dificultam a compreensão do conteúdo e, por consequência, prejudicam o processo de comunicação. Entretanto, não há evidências científicas relacionadas à construção de identificadores automáticos de erros linguísticos para summarizadores. Por esse motivo, pesquisas relacionadas ao desenvolvimento de ferramentas capazes de constatar de forma automática esses problemas linguísticos tornam-se cada vez mais necessárias. Ademais, pode-se dizer que a identificação de erros é essencial para um futuro pós-processamento dos geradores textuais e summarizadores automáticos.

Assim, tendo em vista o impacto de erros relacionados à coesão e coerência sobre a qualidade dos sumários multidocumento, esta pesquisa, ainda em andamento, tem como objetivo produzir uma ferramenta capaz de identificar automaticamente cinco tipos de erros: menção de acrônimos sem explicação, menção subsequente com explicação, informações redundantes, informações contraditórias e pronomes sem referência. É válido mencionar que os autores deste trabalho conseguiram elaborar uma ferramenta capaz de identificar erros relacionados à citação de acrônimos sem definição. Assim, para facilitar o entendimento de todo o estudo, nas seções seguintes serão apresentados os trabalhos correlatos à área de pesquisa (Seção II), a metodologia escolhida para o desenvolvimento do trabalho (Seção III), os resultados parciais alcançados (Seção IV) e, por fim, algumas considerações finais (Seção V).

II. TRABALHOS CORRELATOS

Conforme mencionado, o processo de sumarização automática é capaz de compilar as principais informações presentes em vários textos. Dessa forma, pode-se afirmar que os summarizadores automáticos são de grande importância para a área de PLN (Processamento de Linguagem Natural). Contudo, embora os sumários sejam importantes e informativos, ainda hoje, muitos summarizadores não prezam pela qualidade linguística dos resumos produzidos [6].

A qualidade linguística de um texto está relacionada à coesão e à coerência. A coesão refere-se ao bom uso de conectores entre as palavras, orações e frases, enquanto que a

¹Arte rupestre é o termo que denomina as representações artísticas pré-históricas realizadas em paredes

²Coleção de material escrito e/ou falado usado no estudo da língua (<https://www.dicio.com.br/corpus/>).

coerência refere-se ao ordenamento das ideias. Esses elementos, portanto, são os responsáveis por tornar os textos claros e compreensíveis para o leitor. Nesse contexto, Koch [7] afirma que a coerência e a coesão são essenciais para a transmissão da informação correta ao leitor.

Em relação aos sumários automáticos, Otterbacher e colaboradores [5] constataram que os principais problemas linguísticos estão relacionados à falta de pontuação, uso de sentenças muito longas e uso inadequado de parênteses e outros elementos textuais. Assim, a partir desse contexto, algumas pesquisas passaram a ser desenvolvidas com o propósito de melhorar a qualidade linguística dos sumários produzidos. Dentre esses trabalhos, destaca-se o projeto de Friedrich e colaboradores [8] que utilizou um corpus de sumários multidocumento denominado LQVSumm. Esses pesquisadores conseguiram identificar e corrigir manualmente erros relacionados à menção de entidades e problemas de gramática e redundância.

Dias [6], por sua vez, identificou 12 tipos de erros que afetam a qualidade dos resumos produzidos por summarizadores automáticos. Os erros mais frequentes encontrados pelo pesquisador estão relacionados a menções de acrônimos sem explicação, menções subsequentes de entidades com explicação, sintagma nominal definido sem referência e informações redundantes.

Outros trabalhos também foram desenvolvidos com o intuito de estabelecer modelos para a classificação de summarizadores automáticos. Dentre esses trabalhos, pode-se mencionar o classificador criado por Aguiar [9], denominado Cassiopeia. Esse classificador de sumários funciona de maneira automática e pode ser aplicado na avaliação de qualquer summarizador, independentemente da linguagem. Ademais, esse modelo não necessita de um sumário produzido manualmente no seu processo de avaliação.

III. METODOLOGIA

O processo de identificação de erros linguísticos não é uma tarefa fácil. Grande parte dessa dificuldade está na complexidade dos erros e nos recursos necessários para construir uma base de dados significativa. Por esses motivos, optou-se pelo uso de heurísticas para estabelecer um padrão e determinar a ocorrência de erros, uma vez que esse método funciona muito bem, mesmo com uma base de dados reduzida.

Para facilitar a compreensão do processo de elaboração do software, a metodologia deste trabalho foi dividida nas seguintes etapas: definição do corpus, definição dos tipos de erros, pré-processamento, elaboração de algoritmos para o tratamento do erro acrônimo sem explicação (único erro que já foi identificado pelo software), implementação da ferramenta para a identificação do erro supracitado e validação do software.

A. Corpus

Todo estudo linguístico relacionado à Sumarização Automática Multidocumento (SAM) utiliza uma amostra da língua, seja ela escrita ou falada, para estabelecer um estudo sobre o fenômeno linguístico que se queira analisar. Para este

trabalho, utilizou-se os sumários automáticos multidocumento do corpus CSTNews, o qual contém 50 coleções de textos jornalísticos extraídos de jornais importantes do país (“O Globo”, “Jornal do Brasil”, “Gazeta do Povo”, etc). Os textos extraídos desses jornais versam sobre os seguintes temas: Mundo, Política, Cotidiano, Ciência e Esporte e cada coleção possui de 2 a 3 textos de diferentes origens. Ao todo são 140 textos que contêm em média 334 palavras cada.

O CSTNews é um corpus rico em informações que foram anotadas por especialistas da área de linguística e linguística computacional. Nesse corpus, há sumários gerados por summarizadores monodocumento (sumário procedente de um único texto fonte) e multidocumento (sumário formado a partir de vários textos). Dentre as informações anotadas, destacam-se o alinhamento entre os textos fontes e seus respectivos sumários, expressões temporais, relações RST (Rhetorical Structure Theory) [10] e CST (Cross-document Theory) [11], além da segmentação sentencial dos textos fontes.

Ademais, os sumários multidocumento possuem anotações de erros linguísticos [6]. Para essa tarefa de anotação, foram utilizados 200 sumários gerados automaticamente por 4 summarizadores (GistSumm [12], RSumm [13], RC-4 [14] e MTRST-MCAD [15]). Assim, para cada uma das 50 coleções do corpus, um sumário de cada summarizador foi gerado.

B. Definição dos tipos de erros

A primeira etapa do processo de elaboração da ferramenta consistiu em selecionar os erros a serem identificados. Para isso, utilizou-se como base o trabalho de Dias [6] que identificou manualmente 12 erros em sumários automáticos procedentes do corpus CSTNews. Assim, a partir desse estudo e considerando o número de ocorrências e o impacto desses erros na compreensão do sumário, escolheu-se previamente cinco erros linguísticos para serem identificados pela ferramenta, a saber: menção de acrônimos sem explicação, menção subsequente com explicação, informações redundantes, informações contraditórias e pronomes sem referência. A seguir, há uma breve descrição dos erros supracitados.

Menção de acrônimos sem explicação

Esse erro ocorre quando um acrônimo (sigla) é mencionado em um texto sem que haja uma explicação referente ao seu significado. Dessa maneira, a compreensão do conteúdo pelo leitor pode ser prejudicada, uma vez que este pode não conhecer o significado do acrônimo citado. Na 1, pode-se perceber, por exemplo, que os acrônimos “PIS” e “Cofins” foram mencionados no texto sem qualquer explicação sobre seus significados. A anotação para esse tipo de erro é “TYPE=ARC-EXP”.

[S6] Antecipar o crédito de <e TYPE=ACR-EXP>PIS</e>-<e TYPE=ACR-EXP>Cofins</e> incidente sobre as exportações de 24 meses para 18 meses; e reduzir pela metade o prazo da depreciação dos bens de capital e imóveis, que é um abatimento que as empresas fazem na CSLL (Contribuição Social sobre o Lucro Líquido) sobre investimentos feitos.

Figura 1. Menção de acrônimos sem explicação [6]

Menção subsequente com explicação

Esse erro ocorre quando um sumário menciona o significado de uma entidade duas ou mais vezes. Na Figura 2, por exemplo, a explanação referente ao cargo ocupado por Leonardo Quintanilha foi mencionada duas vezes no mesmo texto. Esse tipo de erro é identificado no sumário pela anotação “TYPE=nM+EXP”.

[S2] O presidente do Conselho de Ética do Senado, Leomar Quintanilha PMDB ...
[S3] <e TYPE=RED SENT=S2><e TYPE=nM+EXP SENT=S2,>O presidente do conselho, Leomar Quintanilha (PMDB-TO)</e>, disse que é contra ...

Figura 2. Menções subsequentes com explicação [6]

Informações Redundantes

Em um texto, o erro “informações redundantes” ocorre quando a ideia contida em uma sentença já fora mencionada anteriormente. Dessa forma, o texto contém informações desnecessárias que não acrescentam em nada ao leitor. A Figura 3 ilustra um sumário contendo informações redundantes. No trecho abaixo, as anotações “TYPE = RED” e “SENT=S1” indicam que a informação sobre a intensificação da fiscalização pela Receita Federal foi repetida tanto na sentença S1 quanto na sentença S2.

[S1] A Receita Federal intensificou a fiscalização sobre as declarações das pessoas físicas neste ano.
[S2] <e TYPE=RED SENT=S1>A Receita Federal intensificou a fiscalização e o resultado foi um aumento do número de contribuintes que caíram na malha fina.</e>

Figura 3. Informação Redundante [6]

Informações contraditórias

Um mesmo sumário pode conter sentenças com informações que se contradizem. Dessa forma, o leitor torna-se incapaz de distinguir qual informação de fato é a verdadeira. A partir do trecho ilustrado pela Figura 4, pode-se notar, por exemplo, que as anotações “TYPE=CONTR” e “SENT=S1” indicam que há divergência entre o número de mortos e feridos nas sentenças S1 e S2.

[S1] O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas....

[S2] <e TYPE=RED SENT=S1><e TYPE=CONTR SENT=S1>Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas ...

Figura 4. Informação Contraditória [6]

Pronomes sem referência

Esse erro ocorre quando um pronome é utilizado em uma sentença sem que o sujeito ao qual ele se refere tenha sido apresentado no texto. Quando isso ocorre, o leitor fica impossibilitado de saber a que o pronome se refere. A Figura 5 apresenta um exemplo sobre esse tipo de erro, destacado no corpus pela anotação ”TYPE=PRO-REF”.

[S1] Internado em um hospital em Buenos Aires, <e TYPE=PRO-REF>ele</e> teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.

Figura 5. Pronomes sem referência [6]

C. Pré-processamento

Após a definição do corpus e os tipos de erros a serem identificados, a próxima etapa consistiu em realizar um pré-processamento dos textos. Durante esse processo, as seguintes tarefas foram realizadas: divisão do corpus, limpeza das anotações, análise sintática e segmentação textual.

Divisão do corpus

Conforme mencionado, para a elaboração da ferramenta utilizou-se 200 sumários. Contudo, como alguns resumos não possuem todos os cinco tipos de erros que serão identificados pelo software, optou-se, inicialmente, por dividir o corpus em cinco grupos, conforme a ocorrência do erro: menção de acrônimos sem explicação, menção subsequente com explicação, informações redundantes, informações contraditórias ou pronomes sem referência.

Em seguida, cada grupo foi dividido em dois subgrupos. Isso foi feito para que uma parte dos sumários seja utilizada para o desenvolvimento do software (fase de treino), e a outra parte sirva para a análise dos resultados (fase de teste). Essa divisão foi feita de forma randômica, de modo que 60% dos textos sejam utilizados na fase de treino e os 40% restantes sejam utilizados para validar a ferramenta

Limpeza das anotações

O corpus de sumários utilizado neste trabalho possui anotações específicas para cada tipo de erro, conforme as ilustrações Figuras 1, 2, 3, 4 e 5. Essas anotações, sem dúvidas, serão úteis para o desenvolvimento e validação da ferramenta.

Todavia, os sumários que serão utilizados durante a fase de teste precisam estar limpos de qualquer marcação para que a ferramenta possa identificar os possíveis erros linguísticos

sem nenhum tipo de viés. Assim, para apagar todas essas marcações nos resumos da fase teste, utilizou-se um *script*.

Análise Sintática e segmentação textual

As informações sintáticas de cada sentença dos sumários podem ser úteis na construção das heurísticas de identificação dos erros. Por isso, optou-se por utilizar um analisador morfossintático. Os pesquisadores decidiram utilizar o analisador Palavras [16], devido à sua precisão e fácil obtenção de informações morfossintáticas. Dessa forma, para cada palavra que compõe uma sentença, o Palavras é capaz de determinar sua classe morfológica e o seu papel sintático no texto.

Por fim, a última etapa do pré-processamento consistiu em dividir cada texto em sentenças e cada sentença em palavras. Tal divisão será necessária para facilitar o processo de análise que será realizado pela ferramenta.

D. Elaboração de algoritmos para o tratamento de todos os erros

O processo de identificação automática de erros linguísticos, ao contrário do que se pensa, é uma tarefa delicada. Esta complexidade deve-se à riqueza de detalhes, como também, à dificuldade de elaborar heurísticas capazes de compreender o contexto expresso no texto. Dessa forma, para cada erro será feito um estudo minucioso sobre o seu padrão de ocorrência. Esse estudo será necessário para a elaboração de heurísticas específicas para cada erro. Até o momento, foram definidas apenas as heurísticas para identificar o erro menção de acrônimos sem explicação.

E. Implementação da ferramenta de identificação de erros

Para a implementação do protótipo, decidiu-se utilizar a linguagem de programação Python, devido à sua facilidade de manipular textos, à sua sintaxe simples e à sua grande versatilidade.

F. Avaliação da ferramenta e análise dos resultados

Após a implementação do protótipo, serão realizados testes para determinar a precisão da ferramenta. O cálculo dessa precisão é realizado comparando-se o número de anotações manuais presentes no corpus de teste (segmentado no tópico de pré-processamento) e o número de ocorrências identificadas automaticamente pelo protótipo. Contudo, apenas essa medida de precisão não é capaz de traduzir de forma justa a eficiência do protótipo produzido, uma vez que este cálculo pode indicar a presença de erros onde na verdade não há. Por esse motivo, os autores deste trabalho também irão mensurar a acurácia da ferramenta. A Figura 6 ilustra a fórmula que será utilizada nesse processo. É válido mencionar que esta pesquisa já tem os resultados de precisão e acurácia da ferramenta construída para identificar o erro menção de acrônimo sem explicação.

$$\frac{Qi}{Qa} = \text{Precisão}$$

$$\frac{Qi - Qe}{Qa} = \text{Acurácia}$$

Qi = Quantidade de erros indetectados pelo protótipo

Qa = Quantidade de erros anotados manualmente

Qe = Quantidade de erros identificados erroneamente

Figura 6. Fórmula da Precisão e acurácia adaptada com base do trabalho de Marçal et al. [17]

IV. RESULTADOS PARCIAIS

Esta pesquisa ainda está em andamento, entretanto, os pesquisadores já alcançaram alguns resultados. A versão atual do protótipo já possui toda a etapa de pré-processamento implementada. Dessa maneira, o usuário necessita apenas ajustar o diretório de arquivos em sua máquina local e fornecer uma lista contendo os textos que devem ser processados. Feito isso, a ferramenta fará automaticamente todas as etapas mencionadas no pré-processamento.

A identificação dos erros vem sendo implementada com base nos mesmos critérios que foram utilizados para definir os tipos de erros a serem identificados. Desse modo, devido à grande quantidade de ocorrências, decidiu-se, inicialmente, identificar de forma automática o erro acrônimo sem explicação. Pode-se afirmar que o protótipo construído é um sucesso, uma vez que a versão atual da ferramenta alcançou os valores de acurácia e precisão igual a 98.7%, conseguindo identificar corretamente 76 dos 77 erros presentes no corpus de teste. Atualmente, os pesquisadores estão se esforçando para construir heurísticas capazes de identificar de forma automática o erro relacionado a menções subsequentes com explicação.

V. CONSIDERAÇÕES FINAIS

Em suma, pode-se dizer que esta pesquisa visa aprimorar a qualidade linguística dos sumários provenientes da técnica de Sumarização Automática Multidocumento. Até então, os autores deste trabalho conseguiram implementar de forma satisfatória todo o processo de pré-processamento, necessário para a identificação de erros. Ademais, as heurísticas para a identificação do erro acrônimo sem explicação já foram implementadas e avaliadas como pertinentes. Em continuidade ao trabalho, os autores pretendem identificar os quatro erros remanescentes: menção subsequente com explicação, informações redundantes, informações contraditórias e pronome sem referência. Espera-se, dessa maneira, trazer contribuições científicas e incentivar o desenvolvimento de pesquisas voltadas à correção de erros linguísticos.

REFERÊNCIAS

- [1] L. Monteiro, “A internet como meio de comunicação: possibilidades e limitações,” in *Congresso Brasileiro de Comunicação*, vol. 24, 2001.
- [2] I. N. C. Teobaldo, “A cidade espetáculo: efeito da globalização,” *Sociologia: Revista da Faculdade de Letras da Universidade do Porto*, vol. 20, 2017.
- [3] I. Mani, *Automatic summarization*. John Benjamins Publishing, 2001, vol. 3.
- [4] A. Nenkova, K. McKeown *et al.*, “Automatic summarization,” *Foundations and Trends® in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2011.
- [5] J. C. Otterbacher, D. R. Radev, and A. Luo, “Revisions that improve cohesion in multi-document summaries: A preliminary study,” in *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, ser. AS '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 27–36.
- [6] M. S. Dias, “Investigação de modelos de coerência local para sumários multidocumento,” Ph.D. dissertation, Instituto de Ciências -USP, 2016.
- [7] I. G. V. Koch, *A coesão textual – Mecanismos de Constituição Textual, A organização do Texto, Fenômenos de Linguagem*, 10th ed. Linguística Contexto – Repensando a Língua Portuguesa, 1998.
- [8] A. Friedrich, M. Valeeva, and A. Palmer, “Lqvsumm: A corpus of linguistic quality violations in multi-document summarization,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014.
- [9] L. H. G. d. Aguiar, “Modelo cassiopeia como avaliador de sumários automáticos: aplicação em um corpus educacional,” 2017.
- [10] W. C. Mann and S. A. Thompson, *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute, 1987.
- [11] D. R. Radev, “A common theory of information fusion from multiple text sources step one: cross-document structure,” in *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-VOLUME 10*. Association for Computational Linguistics, 2000, pp. 74–83.
- [12] P. P. B. Filho, T. A. S. Pardo, and M. das Graças Volpe Nunes, “Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm,” NILC - ICMC-USP. 23 p., Tech. Rep., 2007.
- [13] R. Ribaldo, “Investigação de mapas de relacionamento para sumarização multidocumento,” 2013, monografia de Conclusão de Curso, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Novembro, 61p.
- [14] P. Cardoso, M. Castro Jorge, and T. Pardo, “Exploring the rhetorical structure theory for multi-document summarization,” in *Proceedings of the 5th Workshop RST and Discourse Studies*, 2015, pp. 1 – 10.
- [15] M. L. R. Castro Jorge, “Modelagem gerativa para sumarização automática multidocumento,” Ph.D. dissertation, Instituto de Ciências Matemáticas e de Computação - ICMC/USP, 2015.
- [16] E. Bick, *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus Universitetsforlag, 2000.
- [17] E. K. Marçal *et al.*, “Auditoria da qualidade de um software de contabilidade,” *Gestão & Regionalidade*, vol. 23, no. 66, 2009.

Chatbot para auxiliar os discentes nos procedimentos administrativos de uma universidade

Wesley Benício dos Santos Silva
Departamento de Ciência da Computação (DCC)
Universidade Federal de Goiás – Regional Catalão (UFG-RC)
Catalão, Goiás
wesbdss@discente.ufg.br

Márcio de Souza Dias
Departamento de Ciência da Computação(DCC)
Universidade Federal de Goiás – Regional Catalão (UFG-RC)
Catalão, Goias
marciosouzadias@ufg.br

Nádia F. F. da Silva
Instituto de Informática (INF)
Universidade Federal de Goiás – Regional Goiânia (UFG – RG)
Goiânia, Goiás
nadia@ufg.br

I. INTRODUÇÃO

A capacidade de comunicação fluente das máquinas, por meio de linguagem natural, era vista apenas em filmes de ficção científica. Em 1950, Alan Turing colaborou com seus conceitos e fundamentos computacionais, desenvolvendo as bases para a construção da inteligência computacional. Alan Turing já imaginava a possibilidade de haver máquinas inteligentes no futuro e, desenvolveu um método que pudesse referendar os seus pensamentos. Tal método ficou conhecido como Teste de Turing [7].

A definição de um sistema de diálogo que se comporta como uma pessoa, feita por [19] diz que: um chatbot é um programa de “Inteligência Artificial” que tenta simular uma conversa digitada, com o objetivo de enganar, pelo menos temporariamente, um humano a pensar que estava conversando com outra pessoa.

O Teste de Turing consiste em verificar se as pessoas conseguem identificar com quem elas estão interagindo, por meio de um terminal. Caso as pessoas não consigam distinguir no diálogo quem é a máquina e quem não é, a máquina era considerada “inteligente”. Tal teste deu início aos chatbots e como seria posto em prova a sua inteligência. [7].

Um dos chatbots mais conhecidos e que conseguiu passar no Teste de Turing foi a Eliza. Eliza é um chatbot que simula uma psicóloga virtual, programada com cartões perfurados e desenvolvida por Joseph Weizenbaum [11]. Em 1972, um chatbot que simulava um paciente com paranoíia para a Eliza foi projetado pelo psiquiatra Kenneth Colby, tal sistema foi chamado de Parry [1].

Com os avanços nos campos de pesquisas voltadas para PLN (Processamento de Linguagem Natural), o surgimento de chatbots complexos, com a aplicação de algoritmos de inteligência artificial para processamento de texto, podem manter diálogos mais interessantes e atrativos. Pensando na quantidade de informações que um chatbot pode distribuir automaticamente, a seguinte ideia é proposta: o desenvolvimento de uma ferramenta automática para o auxílio e esclarecimento sobre informações sobre a universidade. Esse problema é observado em grande parte dos ingressantes das universidades que, geralmente, não possuem conhecimento de todas as estruturas que formam a universidade, como informações sobre bolsas, participações

acadêmicas, procedimentos de trancamento de matrículas, etc. Por muitas vezes, a obtenção dessas informações é através da exploração do site da instituição que o estudante faz parte ou a que ele deseja fazer parte.

A característica dos chatbots como comunicadores virtuais são bem vistas na internet, entretanto tornar chatbots bons auxiliares automáticos com o objetivo de responder dúvidas dos usuários pode se tornar um desafio. Algumas dificuldades para criação do chatbot são: o tratamento e reconhecimento da linguagem natural do usuário, o processo de tratamento dos dados e estruturação, e retorno de uma resposta correspondente e coerente sobre a dúvida do usuário.

Pensando na melhor forma de comunicação automatizada, o objetivo dessa proposta é desenvolver um chatbot que seja capaz de responder aos questionamentos sobre os procedimentos de uma universidade de forma mais precisa e com um diálogo comprehensível a todos os discentes, futuros discentes e comunidade em geral.

Esta aplicação auxiliará os usuários a obterem respostas rápidas sobre algum procedimento e funcionamento da universidade, onde informações que ajude o usuário a encontrar o que procura serão repassadas. Em diversas universidades não possuem esse sistema de auxílio automático, mesmo sendo uma ferramenta que seria de grande ajuda a estudantes.

II. FUNDAMENTAÇÃO TEÓRICA

Diversos chatbots utilizaram a linguagem de marcação AIML, porque ela mantém as respostas em um arquivo organizado por *tags*.

Com uma organização mais elaborada dos arquivos AIML é possível simular diversas respostas e contextos diferentes durante o diálogo com o chatbot. Porém, o AIML não é suficiente para cobrir todas as possíveis entradas sobre temas diferenciados que o usuário pode solicitar e responder corretamente sem que o chatbot redirecione para resposta padrão do arquivo AIML.

Os arquivos AIML são estruturados por *tags*. Na Figura 1 é mostrado as principais *tags*:

```

<AIML>
...
<category>
<pattern>OI</pattern>
<template>Olá</template>
</category>
...
</AIML>

```

Fig 1. Exemplos de estruturação das tags AIML

As tags `<AIML>` `</AIML>` definem o começo e o fim do arquivo, todas as outras tags são inseridas em seu interior.

As tags `<category>` `</category>` separam cada bloco de diálogo e as `<pattern>``</pattern>` definem o local que deve ser comparada com a entrada do usuário. Por fim, `<template>``</template>` define a resposta para a respectiva entrada inserida na tag `<pattern>`.

Outras tags que estão disponíveis no AIML: `<star>`, `<srai>`, `<random>`, `<set>`, `<get>`, `<that>`, `<topic>`, `<think>` e `<condition>`, que podem tornar as respostas mais interessantes e diversificadas abrangendo algumas funcionalidade a serem executado pelo kernel de AIML.

Como observado no AIML, para cada entrada o chatbot deve produzir uma saída correspondente, essas características podem ser vistas nas tags `<pattern>` e `<template>`, respectivamente. Para isso, uma forma de analisar a entrada e devolver uma saída será obtida. Para saber qual a intenção do usuário, deve-se classificar a entrada dele em alguma categoria, para isso deve-se extrair a intenção proposta na entrada.

Para o desenvolvimento do chatbot proposto, algumas ferramentas se sobressaíram entre artigos de pesquisadores e algumas foram selecionadas para serem usadas no processo.

Para a estruturação do chatbot, a linguagem Python será utilizada, por ser uma linguagem rápida e com bastante apoio da comunidade nas áreas de *data mining* e algoritmos de inteligência artificial. Para a distribuição de acesso do chatbot, pode ser usado um servidor desenvolvido com o framework Django, em Python, para disponibilização do chatbot na rede interna da universidade.

Para a extração de intenções da entrada há a possibilidade de replicar o experimento realizado no trabalho [5]. Neste trabalho os autores conseguiram representar a entrada do usuário em representação vetorial e depois usaram algoritmo de Redes Neurais Convolucionais (CNN)[22] pré-treinada para classificação desses vetores, obtendo bons resultados em trabalhar com a vetorização das sentenças.

Outro fator importante no chatbot é a representação dos estados de diálogos, para que o chatbot possa manter o estado de diálogo, podendo utilizar uma máquina de estados para manter o controle do chatbot do andamento da conversa e realizar as funções disponíveis no contexto do

estado, por exemplo: se o usuário estiver no estado para marcar realizar um função de enviar uma sugestão para o professor e o bot estiver esperando um número de matrícula como entrada e mesmo assim o estudante enviar outra coisa, o cancelamento daquela função deve ser realizado, e a mudança de estado deve voltar para o início da conversa, reiniciando-a. Os estados de diálogos definem qual resposta é retornada dependendo de qual estado ela se localiza, esses estados podem variar a resposta dada para uma mesma entrada.

III. ESTADO DA ARTE

Diversos chatbots foram criados com o intuito de auxiliar em alguma tarefa, como a retirada de dúvidas. O trabalho [6] propôs a criação de um chatbot com a utilização da linguagem de marcação AIML (*Artificial Intelligence Modelling Language*), introduzida a primeira vez no chatbot ALICE (*Artificial Linguistic Internet Computer Entity*) [16] e LSA (*Latent Semantic Analysis*) [4]. O AIML foi utilizado como um gerenciador de respostas, devido a sua estrutura e o LSA utilizado para o gerenciamento das perguntas. O LSA assume que palavras com significados semelhantes aparecem em textos bases semelhantes, assim podendo padronizar a entrada do usuário para encontrar respostas no arquivo AIML. A aplicação dessa ferramenta era para satisfazer dúvidas sobre a universidade, o problema observado é que a falta de informações vindas do usuário pode não resultar em uma resposta adequada.

Os autores de [8] produziram um chatbot utilizando a ferramenta Wmatrix4 [14], para transcrever os diálogos expostos no corpus BNC (*British National Corpus*) [15] e FAQ's (*Frequently Asked Questions*) sobre assuntos diversificados para a linguagem AIML. Como resultado, foi obtido diálogos mais animados devido ao corpus em que os diálogos foram transcritos, herdando a forma de fala de personagens do corpus, resultado em chatbots que conseguem responder perguntas e retornar textos com características correspondentes do personagem em que o diálogo foi transscrito do corpus.

Em [3] é demonstrado o uso do AIML junto com o processamento de voz para o desenvolvimento de um chatbot com objetivo de ser aplicado em sistemas bancários mobiles. As dúvidas são sanadas com o auxílio de um banco de dados de respostas adequadas e depois são geradas estruturas de AIML que alimentará as respostas do chatbot. Os autores citam o desenvolvimento de uma interface com um personagem humanoide, representando uma assistente real com movimentações de fala ao se comunicar, deixando a conversa mais interativa.

No trabalho de [12] foi desenvolvido um chatbot com a aplicabilidade em recomendação de literatura acadêmica, utilizando análises de bases de dados em uma linguagem XML (*Extensible Markup Language*) chamada *Pattern Matching Technique* e utilizaram o algoritmo *Graphmaster* [17] como máquina de inferência para descrever as regras de decisões que irão compor o banco de conhecimento. O chatbot foi utilizado para auxiliar os estudos de estudantes do curso de Ciências da Computação. Nomeado como “Ubibot”, o bot foi aplicado em uma plataforma educacional adaptada localmente, alguns estudantes voluntários foram subdivididos em dois grupos, estudantes que iriam estudar com auxílio do bot e os que não plataforma e outros não, para realização da comparação do desempenho dos estudos

entre os estudantes. Como conclusão, o melhor desempenho ficou com os estudantes que foram auxiliados pelo bot, tendo um desempenho maior, mostrando a diferença que um simples bot aplicado a educação pode melhorar o desempenho dos estudantes.

Em [13] foi introduzido um conceito de árvore discursiva em um chatbot, possibilitando classificar algumas perguntas e respostas como coesas ou não. Utilizando a ferramenta OpenNLP [18], que realiza processamentos de texto para interfaces Web e uma API (*Application Programming Interface*) de buscas na internet. O objetivo proposto era buscar perguntas e respostas para serem classificadas como coesas e se tornarem parte de um banco de dados, segundo os autores o uso da árvore discursiva é uma maneira eficiente e produtiva de acesso a informação, pois ao serem aplicados em chatbots, conseguiram simular uma atividade intelectual humana.

IV. NOSSA PROPOSTA

O objetivo do chatbot proposto é conseguir responder um escopo grande e diferenciado de perguntas com maior precisão possível. Para alcançar esse objetivo, pretende-se fazer uso de algoritmos de Aprendizado de Máquina.

A estrutura do chatbot proposto pode ser dividida em duas partes: ambiente de treino e o ambiente de comunicação.

Na Figura 2, a estrutura do ambiente de treino é mostrada. O ambiente de treino é o local onde as ferramentas de construção do chatbot serão colocadas, onde pode-se fazer modificações nos componentes disponibilizados e treinamento supervisionado. Este ambiente, pode ser subdividido em: geração de corpus inicial, pré-processamento, treinamento do chatbot, atualização dos estados, geração de respostas.

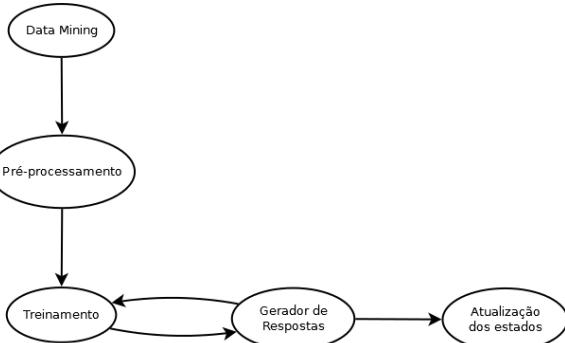


Fig 2. Descrição dos componentes do ambiente de treino

A geração do corpus inicial é realizada pelo algoritmo de *data mining*, que será responsável pela extração de dados para gerar os primeiros dados do corpus. Ele tem a função de percorrer o site da instituição em busca de dados relevantes para geração das primeiras perguntas e respostas que chatbot deverá responder, como perguntas mais frequentes disponibilizadas pela instituição. Esse algoritmo não será usado com frequência, entretanto caso o site da instituição estiver em constantes mudanças, o algoritmo deve atualizar os dados já existentes.

O pré-processamento será a etapa em que toda a entrada é tratada para que seja utilizada corretamente e uniforme

por algum algoritmo de aprendizagem de máquina. A estruturação os textos e a retirada de palavras que não contribuem com o treinamento do chatbot ou que podem causar conflitos na obtenção das respostas, consequentemente diminuindo a acurácia do treinamento.

O treinamento do bot é a etapa que os dados já passaram pré-processamento são aplicados no algoritmo de aprendizagem de máquina. As abordagens com algoritmos de aprendizagem de máquina frequentemente utilizam redes neurais profundas para o reconhecimento de intenções dos usuários mostrado nos artigos [2, 10] e o uso do CNN demonstrado em [5]. Essas técnicas citadas podem ser replicadas nesta etapa, pois demonstram bons resultados.

A atualização dos estados é a etapa de progressão do diálogo do chatbot, para que ele não dê uma resposta fora de contexto e consiga continuar um diálogo fluente.

A geração de respostas é para quando não houver resposta para perguntas proposta por usuários, este algoritmo deve conseguir procurar por respostas correspondente a entrada do usuário, posteriormente validada manualmente, se o contexto é compatível. Para isso a geração de respostas deve ser fluente, legível, adequado e variado como citado em [9]. Esse algoritmo será produzido para trabalhar em pequenos documentos sobre a instituição, trabalhando com técnicas de *data mining* para extração de informações em documentos não estruturados.

O ambiente de comunicação é o local em que o chatbot será disponibilizado para interagir com o usuário. O ambiente em que os dados gerados pela aprendizagem de máquina serão lidos e utilizados para entender a entrada do usuário.

V. RESULTADOS ESPERADOS

O resultado esperado é o desenvolvimento do chatbot que possa compreender os questionamentos do usuário e retornar respostas de forma eficiente sobre a instituição de ensino.

A aplicação do chatbot em sistemas locais para que possa ser realizado testes com usuários reais retornando um feedback de melhorias para o chatbot, pois a partir das informações dos usuários sobre o uso, nível de satisfação ou dúvidas sobre a aplicação, podem ser utilizadas para avaliar o desempenho do chatbot, podendo expandir o seu escopo para outras áreas de aplicação.

VI. CONSIDERAÇÕES FINAIS

O desenvolvimento do chatbot ainda está em sua fase inicial, neste período foram realizados alguns testes na linguagem de marcação AIML e observa-se que o escopo de entradas e saídas do AIML é trabalhoso de ser automatizada, pois o algoritmo deveria gerar todas as possíveis entradas no arquivo AIML, para que ele possa selecionar e retornar a resposta para o usuário. Esse procedimento não é viável para a automatização do chatbot.

Para a próxima etapa do trabalho será a criação do primeiro módulo para testes das ferramentas para a

identificação de intenções, se o método for promissor passaremos para o desenvolvimento dos módulos seguintes.

VII. REFERÊNCIAS

- [1] Colby, K. M. (1974). Ten criticisms of parry. *ACM SIGART Bulletin*, (48), 5-9.
- [2] Deng, L., Tur, G., He, X., & Hakkani-Tur, D. (2012, December). Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *2012 IEEE Spoken Language Technology Workshop (SLT)* (pp. 210-215). IEEE.
- [3] Dole, A., Sansare, H., Harekar, R., & Athalye, S. (2015). Intelligent Chat Bot for Banking System. *International Journal of Emerging Trends & Technology in Computer Science*, 4(5), 49-51.
- [4] Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1), 188-230.
- [5] Hashemi, H. B., Asiaee, A., & Kraft, R. (2016). Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining. Workshop on Query Understanding*.
- [6] Ranoliya, B. R., Raghuwanshi, N., & Singh, S. (2017, September). Chatbot for university related FAQs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1525-1530). IEEE.
- [7] Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and machines*, 10(4), 463-518.
- [8] Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics*, 10(4), 489-516.
- [9] Stent, A., Marge, M., & Singhai, M. (2005, February). Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 341-351). Springer, Berlin, Heidelberg.
- [10] Tur, G., Deng, L., Hakkani-Tür, D., & He, X. (2012, March). Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5045-5048). IEEE.
- [11] Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.
- [12] Paschoal, L. N., de Oliveira, M. M., & Chicon, P. M. M. (2018). A Chatterbot Sensitive to Student's Context to help on Software Engineering Education. In *XLIV Conferencia Latinoamericana de Informática*.
- [13] Galitsky, B., & Ilvovsky, D. (2017, April). Chatbot with a discourse structure-driven dialogue management. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 87-90).
- [14] WMATRIX CORPUS ANALYSIS AND COMPARISON TOOL. Disponível em:<<http://ucrel.lancs.ac.uk/wmatrix/>>. Acesso em:06 abril 2019
- [15] SEARCH THE BRITISH NATIONAL CORPUS ONLINE. Disponível em:<<http://www.natcorp.ox.ac.uk/>>. Acesso em:06 abril 2019.
- [16] Schumaker, R. P., Liu, Y., Ginsburg, M., & Chen, H. (2006). Evaluating mass knowledge acquisition using the alice chatterbot: The az-alice dialog system. *International journal of human-computer studies*, 64(11), 1132-1140.
- [17] B. A. Shawar and E. Atwell. Chatbots: can they serve as natural language interfaces to qa corpus? In 6th International Conference Advances in Computer Science and Engineering, pages 183–188, 2010.
- [18] OPENNLP. Disponivel em: <<https://opennlp.apache.org/>>. Acesso em:06 abril 2019.
- [19] LAVEN, S. "The Simon Laven page." Disponível em: < <http://www.simonlaven.com/>> Acesso em: 02 junho 2019.
- [20] Asimov, I. (1980). *In joy still felt: the autobiography of Isaac Asimov, 1954-1978*. Doubleday Books.
- [21] J. C. K. Cheung and X. Li. Sequence clustering and labeling for unsupervised query intent discovery. In Proceedings of the fifth ACM international conference on Web search and data mining, pages 383–392. ACM, 2012.
- [22] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

A Brief Survey of Deep Learning based methods against OpenNLP NameFinder for Named Entity Recognition on Portuguese Literary Texts

1st Vinicius Amaro Sampaio
Ceará State University
Fortaleza, Brazil
vinicius.sampaio@aluno.uece.br

2nd Mardônio J. C. França
Casa Magalhães
Fortaleza, Brazil
mardfranca@gmail.com

3rd Paulo Bruno Lopes da Silva
University of São Paulo
Fortaleza, Brazil
paulobruno.ls.fr@usp.br

4th Gustavo Augusto Lima de Campos
Ceará State University
Fortaleza, Brazil
gustavo@larces.uece.br

5th Lara Domingos Hissa
Federal University of Ceará
Fortaleza, Brazil
larahissa@fisica.ufc.br

I. INTRODUCTION

Named Entity Recognition is the task of identifying in a text words that have relevance in a given context. For example, in a literary text is useful to identify persons, places and dates. Or in a store being able to have a system that can automatically identify brand and product can improve a database system or even the overall store organization. Many methods are used to build NER models, including rule-based approaches, unsupervised learning approaches, feature-based supervised learning approaches and deep-learning approaches [1].

However, despite these improvements, recognizing proper nouns is still a problem studied by many researchers. For Portuguese language there still a lack of gold standard corpora for NER approaches [2].

In this context, this paper is an comparative study of widely used NER techniques [3] [4] [5] [6] [7] against Deep Learning based methods for the later development of a tool capable of NER in Portuguese literary texts.

Table I [8] summarizes information for some conferences in which the main focus was NER. All conferences have Person, Organization and Location as an entity class in common. English is the most used language when dealing with NER, being HAREM the only conference which focused on the Portuguese language.

TABLE I
NER CONFERENCES

Conference	Relevant years	Language
MUC	1996, 1998	English
CoNLL	2002, 2003	Spanish, Dutch, English, German
ACE	2003	English, Arabic, Chinese
HAREM	2005, 2008	Portuguese

II. NER SYSTEMS

A. Apache OpenNLP

For our baseline we used OpenNLP, a free open-source tool for various NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, language detection and coreference resolution [9].

The NameFinderME function can detect named entities using maximum entropy [10], a classification method that generalizes logistic regression to multiclass problems. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables [11].

B. Conditional Random Fields

Conditional Random Fields (CRF) are a type of discriminative undirected probabilistic graphical model. They are used to encode known relationships between observations and construct consistent interpretations and are often used for labeling or parsing of sequential data [12], in our case natural language.

C. Deep Learning methods

A neural network is composed of a number of nodes, or units, connected by links. Each link has a numeric weight associated with it. Learning usually takes place by updating the weights. Some of the units are connected to the external environment, and can be designated as input or output units. The weights are modified so as to try to bring the network's input/output behavior more in line with that of the environment providing the inputs [13].

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [14].

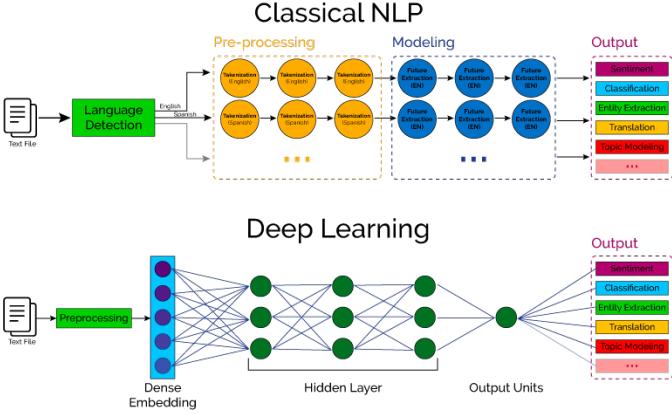


Fig. 1. A typical NLP pipeline

In [15] Deep Learning is used to specific entity terms such as diseases, tests, symptoms, and genes in Electronic Medical Record (EMR) that can be extracted by Named Entity Recognition (NER). In [1], authors present a survey on deep learning for NER.

1) *TensorFlow*: TensorFlow is a interface that express and implements machine learning algorithms. It can be used with little to no change for a wide variety of heterogeneous systems, from mobile devices to large-scale distributed systems [16].

On TensorFlow we used the linear classifier function of the estimator class. It achieves a classification decision based on the value of a linear combination of the characteristics. The characteristics are also know as feature values and usually represented in a vector called a feature vector.

If the input feature vector to the classifier is a real vector \vec{x} , then the output score is:

$$y = f(\vec{w} \cdot \vec{x}) = f\left(\sum_j w_j x_j\right) \quad (1)$$

where \vec{w} is a real vector of weights and f is a function that converts the dot product of the two vectors into the desired output. Often f is a threshold function, which maps all values of $\vec{w} \cdot \vec{x}$ above a certain threshold to the first class and all other values to the second class. [17]

2) *H₂O Framework* : H₂O is a open source machine learning platform, like TensorFlow, H₂O supports the most widely used machine learning algorithms. It uses algorithms based on the feed-forward neural network. [18]

Layered feed-forward networks were first studied in the late 1950s under the name perceptrons. Today the name perceptron is used as a synonym for a single-layer, feed-forward network [13]. On the other hand a multi-layer feed-forward network or a Multi-layer perceptron (MLP) consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function, that is, a function that maps the weighted inputs to the output of each neuron [19]. Learning occurs by making small adjustments in the weights to reduce the difference between the observed and predicted values.

Typically, the updating process is divided into epochs. Each epoch involves updating all the weights for all the examples [13].

With that in mind we chose the Apache OpenNLP [9], TensorFlow [16], H₂O framework [18] and a Python-based implementation of CRF as the systems to be evaluated.

III. FEATURE ENGINEERING

For our corpus to be able to serve as a input for machine learning models it needs to organized in to a feature vector. Each column of that vector represents a characteristic that needs to be evaluated.

For our work we used 1-gram approach, meaning that when trying to classify a token our model looks for the previous and next token. That gives some textual context to our predictions [20].

The following features were used:

- token: the word itself.
- tag: the PoS-Tag of the token.
- i: the position of the token in the sentence.
- token_n_1: the next token.
- tag_n_1: the PoS-Tag of the next token.
- token_p_1: the previous token.
- tag_p_1: the PoS-Tag of the previous token.
- class: the class of the token, being 0 no class, 1 person, 2 place.

Figure 2 offers a example of our feature vector.

As can be seen on our feature vector we also consider the grammatical weight of each token. This process is called Part-of-Speech tagging [21] and was done in a automatic fashion by OpenNLP's PoS-tagger. The annotations generated by this process were used in all methods evaluated in this paper.

	token	tag	i	token_n_1	tag_n_1	token_p_1	tag_p_1	class
0	tomada	v-pcp	0	de	prp	0	0	0
1	de	prp	1	temor	n	tomada	v-pcp	0
2	temor	n	2	.	punc	de	prp	0
3	.	punc	3	helena	prop	temor	n	0
4	helena	prop	4	puxa	v-fin	,	punc	1

Fig. 2. Feature Vector

IV. RESULTS AND DISCUSSION

A. Pipeline

Before training and evaluating our models the raw texts needed to be pre-processed and organized in a way that the evaluated systems can work with.

For that, first we cleared the txt files of unnecessary information such as: page numbers, chapter heads, editorial information, etc. Then the texts were split into sentences and then into tokens using OpenNLP's Sentence Detector and Tokenizer. After that the texts were PoS-tagged using OpenNLP's PoS-Tagger.

A CSV file was then generated for each text with the adequate features for each system with a blank column for the class which was manually populated.

B. Corpora

For the evaluation of the models, we selected 4 literary works: 2 novels written by Jayson Aguiar [22], [23], a contemporary Brazilian author and 2 novels written by Machado de Assis [24], a classic author of Brazilian literature. Machado de Assis was chosen because his works are of great importance for Portuguese literature. Jayson Aguiar works are more recent than Machado's, that offers us a linguistic contrast and a wider range for our corpora. I addition both authors are available in the public domain.

For this analysis, we extracted the first 3 chapters of each book, manually tagging the entities Person and Place, generating an initial dataset with the following properties presented in Table II.

TABLE II
CORPORA

Source	Rows	Tokens	NE	%NE/Tokens
Vermelho - J. Aguiar	89	1792	18	1.0%
Carmim - J. Aguiar	250	3492	129	4.0%
Memorias. - Mac. Assis	101	2444	54	2.2%
Dom Casm. - Mac. Assis	129	2001	46	2.3%

C. Evaluation Metrics

To calculate the performance of the systems we use the following scheme described in [1]:

- **True Positive (tp):** number of labels of a class that are predicted correctly.
- **False Positive (fp):** number of predictions of a class that are wrongly predicted.
- **False Negative (fn):** number of predictions that predict a class but are not labeled as belonging to the class

Then precision is defined as:

$$precision = \frac{tp}{tp + fp} \quad (2)$$

And recall is defined as:

$$recall = \frac{tp}{tp + fn} \quad (3)$$

Finally the F1-score is the harmonic mean of precision and recall

$$f1 = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

D. Results

We compared the results of OpenNLP, TensorFlow, H2O framework and CRF. The results are show in Tables III to VI.

As we can see, Deep Learning approaches had better results when applied to Jayson Aguiar's books. However with

TABLE III
RESULTS - OPENNLP

Source	Precision	Recall	F1-Score
Vermelho - J. Aguiar	0.75	0.34	0.46
Carmim - J. Aguiar	0.94	0.57	0.71
Memorias. - Mac. Assis	0.45	0.34	0.38
Dom Casm. - Mac. Assis	0.50	0.46	0.48

TABLE IV
RESULTS - TENSORFLOW

Source	Precision	Recall	F1-Score
Vermelho - J. Aguiar	0.83	0.71	0.77
Carmim - J. Aguiar	0.92	1.00	0.96
Memorias. - Mac. Assis	0.64	0.70	0.67
Dom Casm. - Mac. Assis	0.94	0.70	0.80

TABLE V
RESULTS - H2O FRAMEWORK

Source	Precision	Recall	F1-Score
Vermelho - J. Aguiar	0.73	0.77	0.75
Carmim - J. Aguiar	0.98	1.00	0.99
Memorias. - Mac. Assis	1.00	0.27	0.46
Dom Casm. - Mac. Assis	1.00	0.23	0.36

TABLE VI
RESULTS - CRF

Source	Precision	Recall	F1-Score
Vermelho - J. Aguiar	0.64	0.54	0.58
Carmim - J. Aguiar	0.61	0.66	0.63
Memorias. - Mac. Assis	0.60	0.46	0.51
Dom Casm. - Mac. Assis	0.90	0.63	0.71

Machado de Assis novels TensorFlow had the highest F1-scores and H2O had the lowest ones. CRF was stable across all the analyzed corpus.

Even if H2O Framework is a Deep Learning technique, its results aren't so different when compared to OpenNLP with Machado de Assis' novels. Despite this, H2O Framework got better results with Aguiar's texts.

Some systems reach a very high F1-score, this can be a symptom of overfitting. We hope that a future extension of the dataset can lead to more reliable results, but it shows the data dependency of those methods.

V. CONCLUSION

The results of our research are still preliminary, but indicate the potential that Deep Learning based approaches can bring improvements to the Named Entities Recognition process. For future works we plan of expanding the corpus and looking at other techniques and frameworks such as Pytorch and Spacy, other Deep Learning methods like LSTM and a combination of CRF+LSTM.

REFERENCES

- [1] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," 2018.
- [2] H. P. D. A. L, E. D. C. T, R. R. D. O. R, S. M, C. S, and D. S. Bermejo, "Lener-br: A dataset for named entity recognition in brazilian legal text," 2018.

- [3] R. Al-Rfou, V. Kulkarni, and . S. S. Perozzi, B., “Polyglot-ner: Massive multilingual named entity recognition,” *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015.
- [4] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications.” *Journal of the American Medical Informatics Association*, 17(5), 507–513., 2015.
- [5] Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” *arXiv:1508.01991*, 2015.
- [6] K. M and K. M., “Crf-based czech named entity recognizer and consolidation of czech ner research.” *Lecture Notes in Computer Science*, 153–160., 2015.
- [7] X. Liu and M. Zhou, “Two-stage ner for tweets with clustering.” *Information Processing Management*, 49(1), 264–273., 2013.
- [8] A. R. O. Pires, “Named entity extraction from portuguese web text,” 2017.
- [9] Apache, “Opennlp,” accessed: 2019-06-01. [Online]. Available: <https://opennlp.apache.org/>
- [10] O. N. Class, accessed: 2019-11-08. [Online]. Available: <https://opennlp.apache.org/docs/1.9.1/apidocs/opennlp-tools/opennlp/tools/namefind/NameFinderME.html>
- [11] W. H. Greene, *Econometric Analysis*. Pearson Education., 2012.
- [12] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann. pp. 282–289., 2001.
- [13] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall, 1994.
- [14] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, 521(7553), 436–444., 2015.
- [15] D. X., C. S. Q. L., L. X., G. Y., Y. J., and Yu, “Deep learning for named entity recognition on chinese electronic medical records: Combining deep transfer learning with multitask bi-directional lstm rnn.” 2019.
- [16] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org Accessed: 2019-06-01. [Online]. Available: <https://www.tensorflow.org/>
- [17] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, “Recent advances of large-scale linear classification,” 2012.
- [18] H. Framework, accessed: 2019-06-01. [Online]. Available: <https://www.h2o.ai/>
- [19] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, 61, 85–117, 2015.
- [20] Z. Wei, D. Miao, J.-H. Chauchat, R. Zhao, and W. Li, “N-grams based feature selection and text representation for chinese text classification,” *International Journal of Computational Intelligence Systems*, vol. 2, no. 4, pp. 365–374, 2009.
- [21] A. Ratnaparkhi, “A maximum entropy model for part-of-speech tagging,” in *Conference on Empirical Methods in Natural Language Processing*, 1996.
- [22] J. Aguiar, *O vermelho do céu*. Corsário, 2011.
- [23] ——, *Carmim*. Armazém, 2019.
- [24] M. de Assis, *Obras completas*, 2019.

Compilação de um Banco Multilíngue de Acolhimento a Pessoas Refugiadas

Furtado, Anna B. D.

Graduanda em Línguas Estrangeiras Aplicadas – LEA - MSI

Universidade de Brasília - UnB

Brasília, Brasil

abdimas@gmail.com

Teixeira, Elisa D.

Professora no Departamento de Línguas Estrangeiras e Tradução

Universidade de Brasília - UnB

Brasília, Brasil

elisadut.unb@gmail.com

I. INTRODUÇÃO

O fenômeno da migração não é novo no mundo e tampouco no Brasil. De acordo com o *Atlas da Migração* [1], havia cerca de 200 milhões de migrantes no mundo em 2009. Conforme o relatório de 2016 da Agência da ONU para Refugiados (ACNUR), apenas durante a primeira metade de 2016, 3,2 milhões de pessoas foram forçadas a deixar seus locais de residência em razão de conflitos ou perseguições. Dessas, 1,5 milhão são solicitantes de refúgio ou formalmente consideradas refugiadas.

Segundo dados da Polícia Federal, somente em 2016 o Brasil recebeu 10.308 solicitações de refúgio [2]. É nesse contexto que surgiu o grupo de pesquisa Mobilidades e Línguas em Contato (MOBILANG), que busca investigar contatos de língua em situações de mobilidade. Esse grupo propôs um projeto de extensão intitulado “Migrações e Fronteiras no Distrito Federal: a Integração Linguística como Garantia dos Direitos Humanos”, cujo principal objetivo é implementar soluções linguísticas para imigrantes e refugiados; ou seja, prover acessibilidade linguística e cultural a essa população que chega ao Brasil.

Conforme pesquisas iniciais do MOBILANG [3] [4] [5] [6], um dos problemas frequentes nos contatos iniciais dos migrantes em solo brasileiro é a dificuldade de comunicação, pelo fato de que alguns deles não falam fluentemente o português. Uma das soluções propostas pelo Projeto para facilitar essa comunicação, especialmente com os órgãos públicos, é a criação de uma base de dados terminológico multilíngue a partir do levantamento de termos e expressões frequentes em textos informativos, normativos e outros documentos utilizados no acolhimento e na inserção dessas pessoas na sociedade brasileira.

Uma etapa inicial e necessária para a elaboração de um banco de dados como esse consiste na compilação de *corpora* eletrônicos que possam servir de base para o levantamento de termos e combinatórias recorrentes de palavras presentes nos tipos de textos e situações comunicativas com os quais os refugiados e imigrantes têm contato. Nesse sentido, o presente trabalho tem como objetivo sistematizar os primeiros passos dados na coleta de um *corpus* multilíngue sobre migração e refúgio, e na extração de padrões linguísticos relevantes (tais como

combinatórias recorrentes de palavras) presentes nesse *corpus* usando metodologias propostas pela Linguística de *Corpus* (doravante LC). O projeto piloto que ora apresentamos descreve a coleta e o preparo (limpeza e conversão para formato legível por computador) de duas cartilhas e quatro formulários disponibilizados aos solicitantes de refúgio e refugiados pelo Ministério da Justiça, por meio do CONARE (Comitê Nacional para os Refugiados) [7] e do ACNUR [8], nos formatos eletrônico e impresso. Os arquivos resultantes foram compilados e explorados usando a ferramenta online *Sketch Engine* [9]. Os dados levantados foram, então, utilizados para a criação de um pequeno glossário multilíngue, cujas equivalências tradutorias, advindas do *corpus* paralelo, comentamos brevemente. Por fim, discutimos os acertos, erros e futuros passos a serem dados para a compilação do referido banco de dados terminológico utilizando *corpora* multilíngues paralelos e comparáveis.

II. LINGUÍSTICA DE CORPUS PARA EXTRAÇÃO DE TERMINOLOGIA

A Linguística de *Corpus* pode ser usada como abordagem ou metodologia no estudo da língua em uso [10], seja a língua geral, seja a língua utilizada em âmbitos especializados e específicos da comunicação humana. É importante diferenciar os estudos direcionados por *corpus*, que usam a LC como abordagem (*corpus driven*, em inglês), dos que a utilizam como ferramenta metodológica apenas, (*corpus based*, em inglês), caso em que o *corpus* serve para confirmar ou refutar teorias previamente estabelecidas [11] [12] [13]. No caso dos estudos direcionados por *corpus*, como é o caso desta pesquisa, os textos são a fonte primária de informações e permitem o estudo de regularidades linguísticas e sua probabilidade de ocorrência nos âmbitos da comunicação representados pelo *corpus*.

De acordo com diversos autores da área (p. ex., Bowker e Pearson [10], Teixeira [12], Berber Sardinha [14], Biber e Conrad [15], McEnery e Wilson [16]), um *corpus* é um apanhado de textos autênticos que foram coletados eletronicamente com base em critérios específicos. Para esses autores, para que um *corpus* possa ser utilizado eficazmente é necessário que atenda a critérios como autenticidade, naturalidade, representatividade e propósito, entre outros.

Adicionalmente, pontuam que, ao se planejar a compilação, é importante determinar algumas características específicas do *corpus*: a) **Língua**: podem ser coletados textos em uma única língua (monolíngues), ou em duas ou mais línguas (bi- e multilíngues); b) **Data de publicação**: um *corpus sincrônico* apresenta um recorte do uso da língua em uma determinada janela temporal, ao passo que um *corpus diacrônico* pode estudar como uma língua evoluiu ao longo de um período de tempo. Além disso, pode ser também **contemporâneo**, apresentando textos atuais, ou **histórico**, com textos de períodos anteriores. Por fim, pode ser **fechado** (não há acréscimo de textos depois de compilado) ou **aberto** (há a possibilidade de acréscimo posteriormente); c) **Modo**: o *corpus* pode ser composto de textos escritos e/ou **falados** (consiste em transcrições de material falado); d) **Conteúdo**: os textos podem representar a língua geral (textos de tipos e gêneros variados) ou um recorte da língua (textos especializados, de cunho regional, etc.); e) **Propósito**: pode ser de **referência**, geralmente contendo uma grande quantidade de material e representativo de uma língua ou variedade linguística, e usado como termo de comparação, ou pode ser de **estudo**, que é aquele contendo os textos que se pretende analisar; f) **Autoria**: os textos podem ter sido escritos por nativos, não-nativos (tradutores, estudantes, etc.); podem ter sido escritos coletivamente, individualmente e/ou institucionalmente; g) **Tamanho**: um *corpus* pode ser pequeno (até 80 mil palavras), pequeno-médio (80 a 250 mil), médio (250 mil a 1 milhão), médio-grande (1 milhão a 10 milhões) e grande (acima de 10 milhões de palavras) [14].

Os *corpora* multilíngues podem ser diferenciados, ainda, com relação à sua organização interna. Podem ser comparáveis (textos com características semelhantes, mas produzidos em línguas diferentes) ou paralelos (textos originais e suas respectivas traduções, alinhados ou não).

Quanto à análise dos *corpora*, segundo Berber Sardinha [14], em LC elas devem observar três fenômenos básicos: a) **Ocorrência**: só se pode utilizar dados que ocorrem nos resultados; b) **Recorrência**: para que se possa afirmar algo sobre um fenômeno linguístico, é necessário que ele ocorra várias vezes; c) **Coocorrência**: uma ocorrência deve ser associada a outros fenômenos linguísticos para que generalizações possam ser feitas.

Os resultados devem, também, ser analisados à luz da padronização da linguagem, que se refere à recorrência sistemática e simultânea de unidades em diversos níveis de análise linguística (lexical, gramatical, sintático, entre outros) [14]. Essa padronização pode ser resumida em três conceitos básicos: a) **Colocação**: relação entre itens lexicais ou entre o léxico e campos semânticos; b) **Coligação**: relação entre itens lexicais e gramaticais; c) **Prosódia semântica**: relação entre itens lexicais e conotação (que pode ser negativa, positiva ou neutra), ou um grupo específico de significados.

Na maioria dos estudos que utilizam LC, os *corpora* são explorados por ferramentas computadorizadas, desenhadas exclusivamente para este fim. São elas: **listadores de palavras** (geram listas de palavras ordenadas por frequência de ocorrência, ordem alfabética, etc.); **concordanciadores** (geram listas que exibem o contexto no qual a palavra ou frase ocorre); **listadores de palavras-chave** (geram as palavras-chave de um *corpus* – aquelas cuja frequência

relativa no *corpus* de estudo é significativa, na comparação com um *corpus* de referência. Geralmente, são ordenadas por ordem decrescente de chavicez; entretanto, é possível ordená-las alfabeticamente, por frequência, entre outros).

III. COLETA DAS CARTILHAS E FORMULÁRIOS

Como dito anteriormente, o *corpus* de estudo aqui utilizado é composto de duas cartilhas e quatro formulários fornecidos nos formatos eletrônico e impresso aos refugiados ou solicitantes de refúgio após sua chegada ao Brasil. Todo o material foi produzido coletiva e institucionalmente pelo ACNUR, em parceria com o CONARE e/ou o Ministério da Justiça do Brasil. Escrito inicialmente em português, todo o material foi traduzido para o espanhol, o inglês, o francês e o árabe. No que concerne à autoria, não há quaisquer informações sobre os tradutores nas cartilhas ou nos endereços nos quais são disponibilizadas. O mesmo ocorre com os formulários: foram produzidos pelo próprio CONARE e disponibilizados na página [17] da Polícia Federal em português, espanhol, francês e inglês. Não há qualquer informação sobre a procedência ou autoria da tradução de nenhum deles, o que é problemático no caso específico deste trabalho pois pode interferir na confiabilidade das equivalências tradutórias que comporão o banco de dados terminológico.

Com base nos princípios já mencionados, os seguintes parâmetros de coleta foram utilizados na compilação do *corpus* deste estudo: a) **tamanho**: inicialmente, 25 mil palavras, a ser ampliado posteriormente; b) **textos**: materiais a que os solicitantes de refúgio e refugiados têm acesso em sua chegada à fronteira brasileira; c) **modo do texto**: escrito; d) **delimitação temática**: imigração / refúgio; e) **línguas**: português, espanhol, francês e inglês; f) **data de publicação**: recente, não mais que 5 anos; g) **tipo de corpus**: paralelo e alinhado.

Todo o material foi coletado *online* e em formato .pdf. As cartilhas foram transformadas em texto sem formatação (.txt), codificados em UTF-8, com uma ferramenta do programa *Adobe Reader Premium* [18]. Os formulários também foram coletados em .pdf no site da Polícia Federal e transformados em texto sem formatação (.txt) em UTF-8 com o *Abbyy FineReader* [19], um programa de gestão de .pdfs e reconhecimento óptico de caracteres (OCR).

Os textos necessitaram de um tratamento manual, pois ao serem transformados em texto sem formatação, muitas palavras foram corrompidas e ocorreram pequenos erros de conversão; por exemplo, o programa reconheceu as caixas de marcação de opção como uma letra “O”. Além disso, algumas setas foram reconhecidas como a letra “E”. Todos esses fenômenos ocorreram de forma aleatória, o que impossibilitou a criação de uma rotina de tratamento do texto em Python, por exemplo.

Para exploração de *corpora* em LC é necessário que os textos estejam disponíveis em formato eletrônico legível, mais especificamente em arquivo de texto sem formatação, extensão .txt. Depois da limpeza, os textos foram anotados, isto é, cada arquivo recebe um cabeçalho em HTML contendo metadados do texto (como autoria, data de publicação, endereço de coleta, nome do arquivo, título do texto, domínio e subdomínio do texto ao qual o texto pertence, data de acesso, entre outros).

O processamento do *corpus* desta pesquisa foi feito com o programa *Sketch Engine* [9], por sua capacidade de processar grandes quantidades de texto nos formatos multilíngue e paralelo. Depois que o texto é inserido no programa, é possível acessar dados sobre as características gerais do *corpus*, tais como a quantidade de *types* (palavras distintas) e *tokens* (palavras ocorrencias) do *corpus*.

Outra função que determinou o uso do programa foi a opção de se obter multi-palavras, também chamadas de “candidatos a termo” [20], expressões mais frequentes no *corpus* de estudo se obtidas em comparação a um *corpus* de referência, além de corresponderem a estruturas terminológicas típicas da língua.

IV. RESULTADOS

O *corpus* piloto apresentou os seguintes resultados: a) **língua portuguesa:** 15.001 *types* e 17.605 *tokes*; b) **língua espanhola:** 15.724 *types* e 18.425 *tokes*; c) **língua francesa:** 16.859 *types* e 19.604 *tokes*; d) **língua inglesa:** 15.145 *types* e 17.870 *tokes*.

Utilizando-nos das listas de candidatos a termo levantadas pelo programa e das linhas de concordâncias geradas a partir destas, propusemos um pequeno glossário de padrões recorrentes e respectivas traduções obtidas exclusivamente a partir dos textos paralelos do *corpus* (vide Tabela 1).

Tabela 1: Amostra de glossário dos candidatos a termo obtidos.

	Português	Inglês	Francês	Espanhol
1	Policia Federal	Federal Police	Police Fédérale	Policía Federal
2	país de origem	country of origin	pays d'origine	país de origen
3	condição de refugiado	refugee status	condition de refugié	condición de refugiado
4	residência habitual	habitual residence	résidence habituelle	residencia habitual
5	formulário de solicitação de refúgio	asylum application Form	formulaire de demande d'Asile	formulario de solicitud de asilo
6	solicitação de refúgio	asylum claim [24]	demande d'asile	solicitud de refugio [5]
		asylum application [12]		solicitud de asilo [32]

V. DISCUSSÃO

Na Tabela 1, apresentamos alguns dos padrões linguísticos levantados utilizando o *Sketch Engine* e os equivalentes escolhidos pelos tradutores de cada língua. O padrão da linha 6, por exemplo, possui mais de uma tradução no *corpus* e, por isso, entre colchetes fornecemos o número de ocorrências de cada uma na referida língua.

No que tange à escolha dos termos que comporiam o banco de dados terminológico, é importante saber que o programa sugere ao usuário candidatos a termos e palavras-chave com base em cálculos estatísticos e de probabilidade [20], mas cabe ao pesquisador determinar quais combinatórias selecionará para incluir em seu trabalho –

etapa ainda não realizada desta pesquisa. Quando se usar um *corpus* paralelo para levantamento de equivalentes tradutórios, como fizemos aqui, a qualidade do glossário resultante dependerá diretamente da qualidade das traduções dos materiais incluídos no *corpus*. Uma outra maneira de levantar equivalentes seria usar *corpora* comparáveis, contendo textos produzidos naturalmente em cada língua. Neste caso, não há garantia de que todos os padrões linguísticos recorrentes no *corpus* de uma das línguas terão um equivalente facilmente identificável nos *corpora* das outras línguas.

Por se tratar de um *corpus* paralelo, uma análise superficial permite apontar algumas incongruências tradutórias. Por exemplo, na linha 1 da Tabela 1, uma das estratégias utilizadas para traduzir o termo “Policía Federal” foi a tradução literal (também conhecido como “(de)calque”). Um dos problemas de se traduzir literalmente, especialmente quando se trata de nomes próprios de organizações, é a falta de correspondência entre a entidade da língua de partida e seu correspondente na língua de chegada. “Federal Police”, “Policie Fédérale” e “Policía Federal” não correspondem à “Polícia migratória” nas línguas inglesa, francesa e espanhola, respectivamente. Uma maneira de oferecer uma maior compreensão para o migrante seria manter o nome do órgão em português, para que o leitor se familiarize com a grafia (já que possivelmente entrará em contato com ela) e oferecer uma explicação do significado na língua de chegada.

Outro exemplo que nos chamou a atenção foi “solicitação de refúgio”, na linha 6, que não apresenta consenso na escolha do equivalente em inglês e nem em espanhol. Isso pode ser problemático, do ponto de vista da compreensão do migrante, pois pode levá-lo a crer que se trata de duas ações distintas: “claim asylum” e “asylum application”, no caso do inglês, e “solicitud de refugio” e “solicitud de asilo”, no espanhol. Vale lembrar que os formulários e cartilhas são documentos oficiais fornecidos a estrangeiros que, em sua maioria, não são falantes do inglês, francês e/ou do espanhol, que podem ser sua segunda ou terceira língua. Portanto, o uso de expressões sinônimas, como estas, poderia ser evitado, em prol de uma melhor compreensão do texto. O fenômeno foi observado também em outros padrões linguísticos levantados no *corpus*.

VI. CONSIDERAÇÕES FINAIS

Ao longo deste resumo, descrevemos o processo de compilação de um pequeno *corpus* de estudo multilíngue sobre migração e refúgio e apresentamos os primeiros passos dados na extração de combinatórias utilizando a ferramenta *Sketch Engine*, objetivo que atingimos com sucesso, uma vez que nossos resultados já serviram de base para outros trabalhos do grupo de pesquisa [21].

Para tanto, apresentamos a abordagem e metodologia que escolhemos, a Linguística de *Corpus*, devido a possibilidade de se trabalhar com textos autênticos que representam a área que se está estudando.

Em seguida, extraímos as combinatórias com as ferramentas do programa *Sketch Engine*, cuja escolha foi guiada pela possibilidade de se manejar grandes quantidades de *corpora* paralelos e multilíngues, além de sua ferramenta multi-palavras, que permite a extração de candidatos a termo.

Por fim, propusemos um pequeno glossário de combinatórias e suas respectivas traduções obtidas exclusivamente por meio do *corpus*. Ao fazer uma análise superficial da tradução feita, concluímos, então, que seria necessário a compilação de um *corpus* que fosse comparável e multilíngue para que se possa obter combinatórias e traduções que tenham maior grau de confiabilidade.

No que concerne as limitações deste projeto, pode-se dizer que a primeira encontrada ao explorar as cartilhas foi a linguística: o *corpus* em árabe não pôde ser extraído, uma vez que, para o alinhamento no nível da sentença, seria necessário que o pesquisador tivesse conhecimentos da língua para corrigir pequenas falhas que possam aparecer (palavras quebradas, falta de pontuação, etc.), não obstante a dificuldade de processar um *corpus* alinhado com uma língua que não usa o alfabeto latino.

Um outro fator limitante foi a quantidade total de texto que compõe este *corpus* de estudo; quanto menor o número de palavras, menos observações e generalizações podem ser feitas. No caso deste estudo piloto, o tamanho é satisfatório para fins de planejamento e teste dos parâmetros de coleta, além de ter permitido a observação e estudo de alguns candidatos a termos e palavras-chave. Mas, para que pudéssemos fazer generalizações mais abrangentes, bem como para a compilação do banco de dados terminológico a que o Projeto MOBILNG se propõe, será necessário que o *corpus* seja muito maior.

Finalmente, o passo seguinte foi coletar uma maior quantidade de textos para aumentar a representatividade do *corpus* – o objetivo é obter cerca de 500 mil palavras por língua – e, consequentemente, gerar um banco de dados multilíngue mais confiável e abrangente ao tema. Com um banco de dados consolidado, será possível criar glossários e dicionários para facilitar a tradução/interpretação de materiais para os solicitantes de refúgio. Futuramente, uma outra possibilidade é considerar a inclusão do árabe no *corpus*, uma vez que é a língua nativa de muitos dos refugiados.

REFERÊNCIAS

- [1] Wihtol, Catherine. *Atlas des migrations dans le monde, Réfugiés ou migrants volontaires, Alternatives Economiques*. Editora Autrement, 2009.
- [2] United Nations High Commissioner for Refugees. (2016) “Mid-year Trends 2016”. [Online] Disponível em: www.unhcr.org/statistics/unhcrstats/58aa8f247/mid-year-trends-june-2016.html. Acesso em: 09/07/2019.
- [3] Miranda, Jessica, “Presas estrangeiras no Brasil: barreiras linguísticas”. [Online] <http://bdm.unb.br/handle/10483/14888>. Em 09/07/2017. Acesso em: 09/07/2019.
- [4] Molina Cabreira, Martha Ingrith. “Migrações e impasses no acesso à saúde: traduzir-se é preciso”. [Online] Disponível em: [Cabrera, http://repositorio.unb.br/handle/10482/31561](http://repositorio.unb.br/handle/10482/31561). Acesso em: 09/07/2019.
- [5] Militão, Cinthia Duarte. “O processo de pedido de refúgio e a integração acadêmica de refugiados na Universidade de Brasília” [Online] Disponível em: <http://bdm.unb.br/handle/10483/19457>. Acesso em: 09/07/2019.
- [6] Urbina, Daniela Bandeira. “Aplicativo integra: protótipo para a promoção do acesso à informação de imigrantes e refugiados de Brasília”. [Online] Disponível em: <http://bdm.unb.br/handle/10483/19459>. Acesso em: 09/07/2019.
- [7] Comitê Nacional para os Refugiados – CONARE. [Online] Disponível em: <https://www.justica.gov.br/seus-direitos/refugio>. Acesso em: 15.07.2019.
- [8] Alto Comissariado das Nações Unidas para Refugiados – ACNUR. [Online] Disponível em: <https://www.acnur.org/portugues/>. Acesso em: 09/07/2019.
- [9] Kilgariff, Adam, et al. “The Sketch Engine”. (2004) [Online] Disponível em: <https://www.sketchengine.eu>. Acesso em: 09/07/2019.
- [10] Bowker, Lynne e Pearson, Jennifer. “Working with Specialized Language”. Routledge. 2002.
- [11] Teixeira, Elisa Duarte. “A lingüística de *corpus* a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual”. [Online] doi:10.11606/T.8.2008.tde-16022009-141747. Acesso em: 09/07/2019.
- [12] Barros, Lídia Almeida. “Curso Básico de Terminologia”. Editora Universidade de São Paulo. 2004.
- [13] McEnery, Tony e Hardie, Andrew. “*Corpus Linguistics: Method, Theory and Practice*”. Cambridge: Cambridge University Press. 2012.
- [14] Berber Sardinha, Tony. “Linguística de *Corpus*”. Editora Manole, 2004.
- [15] Biber, D.; Conrad, S.; Reppen, R. “*Corpus Linguistics: investigating language structure and use*”. Cambridge: Cambridge University Press, 1998.
- [16] McEnery, Tony e Wilson, Andrew. “*Corpus Linguistics: An Introduction*”. Edinburgh University Press, 2001.
- [17] Polícia Federal. “Solicitação de Refúgio” [Online] Disponível em: <http://www.pf.gov.br/servicos-pf/imigracao/refugio>. Acesso em: 09/07/2019.
- [18] Adobe Reader. [Online] Disponível em: <https://acrobat.adobe.com/br/pt/acrobat.html?promoid=C12Y324S&mv=other>. Acesso em: 09/07/2019.
- [19] Abbyy FineReader. [Online] Disponível em: <https://www.abbyy.com/pt-br/>. Acesso em: 09/07/2019.
- [20] Sketch Engine. Keywords and term extraction. [Online] Disponível em: <https://www.sketchengine.eu/guide/keywords-and-term-extraction/#toggle-id-2>. Acesso em: 27 jun 2019.
- [21] Furtado, A. B. D. “Glossário Multilíngue Online sobre Migração e Refúgio: Uma Proposta para Tradutores e Intérpretes”. [115] f. il. Trabalho de Conclusão de Curso (Bacharelado em Línguas Estrangeiras Aplicadas) — Universidade de Brasília, Brasília, 2019.