Building Portuguese Language Resources for Natural Language Processing Tasks PT-Pump-Up

Rúben Almeida
INESC TEC
ruben.f.almeida@inesctec.pt

Abstract

The development of natural language processing solutions in mid-resourced languages like Portuguese is limited by the reduced number of corpora and models available. In recent years, with the advances in transformer-based models, it is possible to achieve state-of-the-art results with fewer documents that present high-quality features, leveraging fine-tuning techniques. In this extended abstract, we describe the work developed in our dissertation focused on extending the existing Portuguese NLP resources. We established a methodology that integrates dataset combination with data augmentation processes based on machine translation to train new off-the-shelf transformer-based models that can be reused using a few lines of Python code.

Our literature review surveyed the existing resources in 12 Portuguese NLP tasks. The results motivated us to explore further Portuguese named entity recognition and abstractive text summarization. The resources developed are available online in Huggingface[1] and establish the foundations for PT-Pump-Up[2], an initiative that aims to serve the Portuguese NLP community maintaining an updated indexation of the Portuguese NLP resources developed.

1 Introduction

The recent advances in Natural Language Processing (NLP) are not uniform across the different human languages (Magueresse et al., 2020). Many factors contribute to this inequality, like the economic strength of nations and the number of speakers of a language, to name just a few. In mid-resource languages like Portuguese (Suá rez et al., 2020), there is access to a moderate number of resources for NLP tasks resulting from decades of research like named entity recognition (NER) and part-of-speech tagging (POS-tagging). However, for other NLP tasks like question answering or common sense reasoning, resources are scarcer.

Not only are Portuguese resources less numerous, they are also less accessible: Firstly, they are dispersed across several platforms like Portulan Clarin[3], Linguateca[4], NILC[5], INESC-Tec[6] and across several GitHub repositories dedicated to Portuguese NLP; Secondly, many of these resources do not follow standardized formats, creating additional challenges for major platforms like HuggingFace[7] and Papers With Code[8] to index these resources; Thirdly, much of the research developed in NLP has a pure scientific goal and does not employ many of the modern software engineering sound practices that ensure maintainability and easy access in an off-the-shelf manner.

In this dissertation (de Almeida, 2023), we addressed these challenges by leveraging transformer-based models and data augmentation techniques based on machine translation (MT) to extend the existing resources in two NLP tasks: NER and abstractive text summarization (ATS). The literature review process surveys the existing resources in 12 NLP Portuguese tasks. The information collected establishes the foundations for PT-Pump-Up[9], a platform that aims to facilitate access to Portuguese NLP resources. PT-Pump-Up is the object of a demonstration submitted to PROPOR 2024. These research goals are summarised in the following four research questions:

RQ1: Does combining datasets with the same labelling scheme but different sources positively

[1] https://huggingface.co/arubenruben

[2] https://github.com/LIAAD/PT-Pump-Up

[3] https://portulanclarin.net/

[4] https://www.linguateca.pt/

[5] https://sites.google.com/view/nilc-usp/

[6] https://rdm.inesctec.pt/organization/inesc-tec

[7] https://huggingface.co/datasets?language=language:pt&sort=trending

[8] https://paperswithcode.com/

[9] https://github.com/arubenruben/PT-Pump-Up

impact the overall performance of models?

RQ2: What is the additional overhead to introduce silver labelled data in the training pipeline?

RQ3: Does European Portuguese data negatively impact Pretrained Brazilian Portuguese transformer models?

RQ4: Which are the best public framework to deploy an off-the-shelf NLP tool?

2 Literature Review: Survey of Existing Portuguese NLP Resources

For the literature review process we conducted a survey of existing resources in 12 NLP tasks with a special focus in NER and Abstractive Text Summarization (ATS).

2.1 Datasets

We identified 59 Portuguese language NLP datasets. Most of these resources are silver-labelled, whose annotation was performed using preexistent tools without human interference, are composed of Brazilian Portuguese corpora and do not follow the most recent annotation standards. In Figure 1, we present the datasets counting per NLP task. Further details and statistics about each corpus can be found in the disseration (de Almeida, 2023).

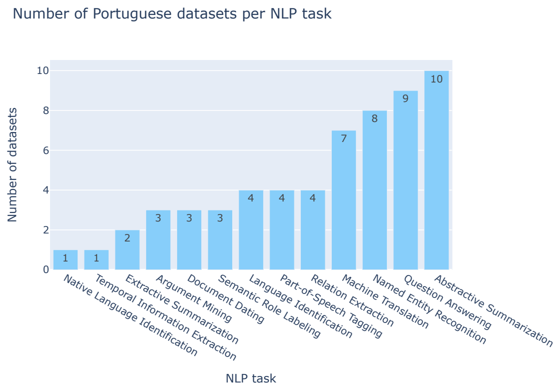

Figure 1: Counts of the Number of Portuguese Datasets Identified.

We observe that NER and ATS are two NLP tasks with more corpora available. In the case of NER, researchers tend to use the three variations of HAREM, first HAREM (Santos et al., 2006), second HAREM (Freitas et al., 2010) and mini HAREM (Santos et al., 2006), as benchmarking tools leveraging the human-annotation that characterize these datasets to extract unbiased performance metrics. In the case of ATS, we indexed ten datasets, most of which are silver-labelled resources composed of online news articles.

2.2 Models

Our literature review identified 33 models, revealing that datasets are more frequent than models. There are NLP tasks that have a corpus, but we were not able to find any literature about NLP models to address that specific task. Native Language Identification and Document Dating are two examples of this phenomenon. In Figure 2, we present the counting of models in each NLP task consider

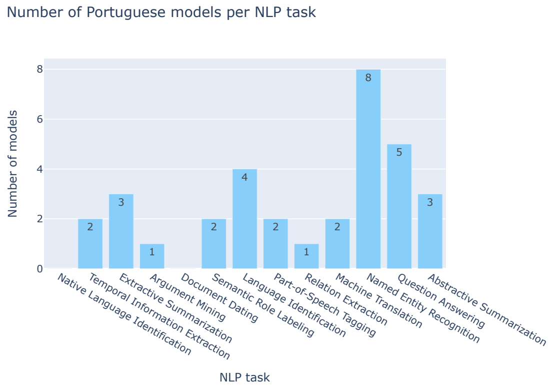

Figure 2: Counts of the Number of Portuguese Models Identified.

We observe that NER is the task we identify as containing more NLP models. It was also possible to identify previous research in ATS, simplifying the comparability of our methodology with the current state-of-the-art (SOTA). In NER, (Souza et al., 2020b) propose a BERT-CRF architecture founded on the usage of the Brazilian version of BERT transformer (Devlin et al., 2019), BERTimbau (Souza et al., 2020a), to score **83.2%** F1-Score in the mini-HAREM (Santos et al., 2006) dataset. For ATS, (Paiola et al., 2022) used the Portuguese version of T5 transformer (Raffel et al., 2020), PT-T5 (Lopes et al., 2020) scored **29.94** ROUGE-L to establish a new benchmark for this NLP task. Both works demonstrate the relevance of the fine-tuning process to address NLP challenges in mid-resource languages, serving as a foundation for the work developed in this dissertation.

2.3 Portuguese Language Varieties: A European Portuguese Dissertation

As stated in **RQ3**, assessing the impact of integrating different Portuguese varieties was paramount in our research. During our literature review process, we identified four works focused on the Portuguese language. These findings opened the door to a dissertation focused on European Portuguese resources.

However, these resources were already some years old, and the authors no longer own them. This fact made it very hard to produce a dissertation focused on European Portuguese. To overcome this limitation, we recently submitted to a major NLP conference an off-the-shelf Portuguese LID capable of discriminating between European and Brazilian Portuguese, operating independently of the textual domain. This tool is currently concealed in the anonymous review of that conference. This tool will unlock further research on the topic.

3 Methodology

We propose a two-stage approach to assess the challenge of developing NLP for mid-resource languages founded on machine learning. First, we leverage the findings of the survey conducted to compose different datasets identified in a single corpus, assessing if composing different datasets for the same NLP task can improve the current benchmark. The intuition behind this technique is to produce a bigger corpus, expecting better results. If the dataset combination process does not surpass the current SOTA, we use machine translation to extend the training corpus, leveraging the many resources available in the English language.

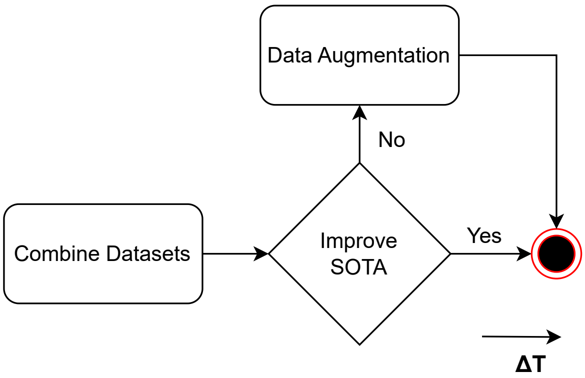

Figure 3: Graphical Representation of Methodology Proposed.

3.1 Evaluation

The evaluation step is paramount in every machine learning problem. To simplify the comparability with the current SOTA, we evaluated our work using the same metrics used on those works: F1-score for NER and ROUGE-L for ATS. Additionally, these metrics were extracted from the same test sets found in the literature. In the case of NER, mini-HAREM (Santos et al., 2006) is widely used; in ATS, the usage of XL-Sum (Hasan et al., 2021) test set is increasing in popularity. Our evaluation process also considered silver-labelled test sets, but only when data augmentation (DA) is applied. The intuition is to obtain insights about the deviation of the silver-labelled case when compared with the gold-labelled corpora. Additionally, in the case of NER, we develop a framework to extract baselines from the widely used NLP library spaCy[10] to establish a straightforward comparison to the results obtained.

3.2 Hybrid Architecture: Combining HuggingFace with Pytorch

In our quest to deliver off-the-shelf solutions, we selected HuggingFace[11] as the most suitable platform to support this goal. HuggingFace offers a user-friendly and efficient environment at the expense of limiting low-level operations. To address this constraint, we devised a hybrid approach that integrates PyTorch with HuggingFace when needed. This strategy not only simplifies model deployment but also enhances compatibility with a diverse range of datasets and models available on the HuggingFace platform.

4 Portuguese Named Entity Recognition

4.1 BERT-CRF

The good results obtained by (Souza et al., 2020b) operating a BERT-CRF architecture used exclusively first HAREM (Santos et al., 2006) as a training set, and the outcome produced was not an off-the-shelf architecture, presenting a great overfit to operations based on the HAREM ecosystem. We reconstructed this promising architecture, adding off-the-shelf capabilities and testing it with different data sources according to the methodology proposed (3). This architecture uses BERTimbau (Souza et al., 2020a) as an encoder to obtain a contextualised embedding as output, delegating the token classification task to a decoder based on Conditional Random Fields (CRF)[12].

[10] https://spacy.io/

[11] https://huggingface.co/

[12] https://pytorch-crf.readthedocs.io/en/stable/

Test Set	Training Set	F1-Score
Mini-HAREM Selective	F.HAREM	0.832
	Portuguese MAPA	0.149
	F.HAREM + S.HAREM	0.789
	S.HAREM	0.730
	F.HAREM + Ontonotes PT	0.790
	Ontonotes 5.0-PT HAREM-Selective	0.558
Mini-HAREM Default	F.HAREM	0.767
	F.HAREM + S.HAREM	0.724
	S.HAREM	0.693
Mini-HAREM CoNLL-2003	F.HAREM CoNLL-2003	0.781
	F.HAREM + S.HAREM CoNLL-2003	0.756
	Wikineural	0.535
Wikineural	Wikineural	0.951
Ontonotes 5.0-PT	Ontonotes 5.0-PT	0.587

Table 1: F1-scores obtained in NER experiments using different test and training datasets

4.2 Corpus Definition

We identify five datasets of interest to follow our methodology: a) First HAREM, this dataset was already used in the original proposal of BERT-CRF (Souza et al., 2020b); b) Second HAREM, an extension to First HAREM performed by the same annotation team; c) Mini HAREM, an extension to First HAREM traditionally used in literature as test set; d) MAPA (Arranz et al., 2022), a recent European dataset to support anonymisation that includes NER annotations; e) Wikineural (Tedeschi et al., 2021), a silver-labeled dataset of Wikipedia documents. These five datasets were also complemented with Ontonotes 5.0 (Weischedel et al., 2013), a SOTA English NER dataset to support the DA process based on MT. These datasets follow different labelling schemes, forcing manual labelling conversion and normalisation. We summarise this information in Table 2.

Dataset	Format	# Labels
F./S./M. HAREM	Selective	5
F./S./M. HAREM	Default	10
MAPA	Coarse Grained	7
MAPA	Fine Grained	23
Wikineural	CoNLL-2003	5

Table 2: Label Distribution of the Different Labelling Schemes Considered.

4.3 Results

We summarize some of the results obtained for Portuguese NER in Table 1. Our experiments cover three types of labelling schemes according to the specificities of the datasets considered: a) HAREM Selective; b) HAREM Default; c) CoNLL-2003. A complete analysis of the results obtained can be found in Sections 5.3 and 5.4 of the dissertation.

The results reveal that the technique of dataset combination was not capable of outperforming training scenarios using a single dataset. A particular case of F1-Score downgrading is *F.HAREM + S.HAREM*, where the combination of First and Second HAREM, two datasets produced by the same annotators with a year of difference in between, negatively impacts the F1-Scores obtained. The results reach the current SOTA in the HAREM Selective labelling scheme and establish a new benchmark in the ConLL-2003 format for the Wikineural test set. Additionally, it is possible to observe that a training process founded exclusively on silver-labelled data was not feasible.

5 Portuguese Abstractive Text Summarization

The limitations faced while exploring MT, specifically for NER, motivated us to leverage ATS to further explore the impact of silver-labelled data in the training process of Portuguese NLP resources.

5.1 T5 Transformer

Abstractive Text Summarization is a challenging task because it requires advanced text generation capabilities (El-Kassas et al., 2021). This feature saw relevant improvements recently, unlocking further research in ATS. The most recent research

Train Set	Test Set	ROUGE-L
Portuguese XL-Sum	Portuguese XL-Sum	26.73
PT-CNN-Dailymail Azure PT-PT (10k)	PT-CNN-Dailymail Azure PT-PT (10k)	25.23
PT-CNN-Dailymail Google (10k)	PT-CNN-Dailymail Google (10k)	25.56

Table 3: ROUGE-L Scores for Different ATS Datasets

in Portuguese ATS, (Paiola et al., 2022), uses the Portuguese variation of the T5 transformer (Raffel et al., 2020), PT-T5 (Lopes et al., 2020) to obtain SOTA results in this NLP task. This architecture is a complete encoder-decoder suitable for tasks that require a generative step.

5.2 Corpus Definition

The process of extending the existent resources in Portuguese ATS focused mainly on Data Augmentation (DA). Unlike the NER task, where we dedicated ourselves to covering different pre-existing datasets, in ATS, we focused on a single Portuguese dataset, XL-Sum. To support the DA process, we translated the CNN-Dailymail corpus (Nallapati et al., 2016).

5.3 Assessing the Impact of Different MT Tool in the Training Process

We leveraged ATS to test different commercial Machine Translation (MT) tools: a) Amazon Web Services (AWS), b) Microsoft Azure, and c) Google Translator, both in their European and Brazilian varieties (when available). However, the elevated cost of these commercial services limited the experiment to 10k documents translated in the European Portuguese variety using Microsoft Azure and the Brazilian variety provided by Google Translator.

5.4 Results

The results obtained in the experiments dedicated to ATS are summarized in Table 3. It is possible to observe that similar performances were obtained using the XL-Sum dataset and the translated corpora. This shows that a training process founded on MT is possible.

The results presented were heavily limited by the cost of these commercial MT systems, leading us to translate only a subset of CNN-Dailymail and by hardware constraints that advised the use of the small variation of the T5 transformer[13].

6 Conclusion & Future Work

The recent advances in NLP using Transformer architectures have raised the bar for NLP tasks, but these advancements rely on significant amounts of data and hardware resources, making them less practical for mid-resource languages like Portuguese. To address this limitation, we proposed a two-step methodology that leverages pre-trained transformer-based models to extend the number of Portuguese NLP resources. First, we index the few existing corpora and combine them to create bigger and more diverse datasets, and when necessary, we use MT to augment the training data, leveraging the vast number of resources existent for the English language. We validated this methodology in NER and ATS, where the research answered questions regarding model performance. Although some limitations were encountered, it was possible to establish a new benchmark for Portuguese NER in the CoNLL-2003 labelling scheme and deliver several NER and ATS resources on Huggingface that operate in an off-the-shelf manner using just a few lines of Python code.

We identified several future work topics for this dissertation, two of which are already under review or development: a) The development of an off-the-shelf Portuguese language identifier; b) The development of PT-Pump-Up. Shortly, we intend to overcome the most significant limitation during the dissertation, the absence of a reliable token alignment framework between English and Portuguese, by creating a neural approach to the problem and an upstream performance metric for this type of problem.

Relevant Links

- CV
- Dissertation
- PT-Pump-Up Github

[13] https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab

References

Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum. 2022. MAPA project: Ready-to-go open-source datasets and deep learning technology to remove identifying information from text documents. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 64–72, Marseille, France. European Language Resources Association.

Rúben Filipe Seabra de Almeida. 2023. Building portuguese language resources for natural language processing tasks.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Cláudia Freitas, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota, and Diana Santos. 2010. Second harem: advancing the state of the art of named entity recognition in portuguese. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association*. European Language Resources Association.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. 2020. Lite training strategies for Portuguese-English and English-Portuguese translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 833–840, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Pedro H. Paiola, Gustavo H. de Rosa, and João P. Papa. 2022. Deep learning-based abstractive summarization for brazilian portuguese texts. In *Intelligent Systems*, pages 479–493, Cham. Springer International Publishing.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. Harem: An advanced ner evaluation contest for portuguese. In *Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020a. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020b. Portuguese named entity recognition using bert-crf.

Pedro Javier Ortiz Suá rez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.