Extração de Relacionamentos Utilizando Redes Neurais Recorrentes

Bruno Dorscheidt Brandelli
Orientadora: Renata Vieira

Sumário

Introdução

- Motivação
  - Aumento da quantidade de dados disponíveis.
  - Necessidade de extrair informações relevantes dos dados.
  - Avanços na área de Extração de Informação

- Caracterização do Problema
  - Número reduzido de sistemas para língua portuguesa

Objetivos

- Gerais
  - Desenvolver um sistema capaz de realizar tarefas de Extração de Relacionamentos (ER)
- Específicos
  - Aprofundar conhecimento em Processamento de Linguagem Natural (PLN)
  - Estudar abordagens e estudos na área de PLN
  - Modelar e treinar um sistema capaz de realizar tarefa de ER
  - Analisar resultados obtidos

Fundamentação Teórica

- Aprendizado de Máquina
- Redes Neurais
- Processamento de Linguagem Natural
- Extração de Informação
- Reconhecimento de Entidades Nomeadas
- Extração de Relacionamentos

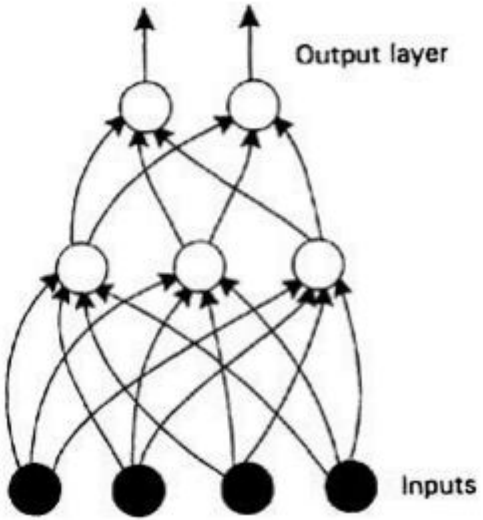
Fundamentação Teórica - Aprendizado de Máquina

Tornar uma máquina capaz de aprender de forma autônoma a resolver problemas específicos.

- Tipos de Aprendizado de Máquina
  - Supervisionado
  - Não-Supervisionado
  - Semi-Supervisionado

Fundamentação Teórica - Redes Neurais

Implementação que busca modelar o modo como o cérebro aprende.



Fundamentação Teórica - Redes Neurais

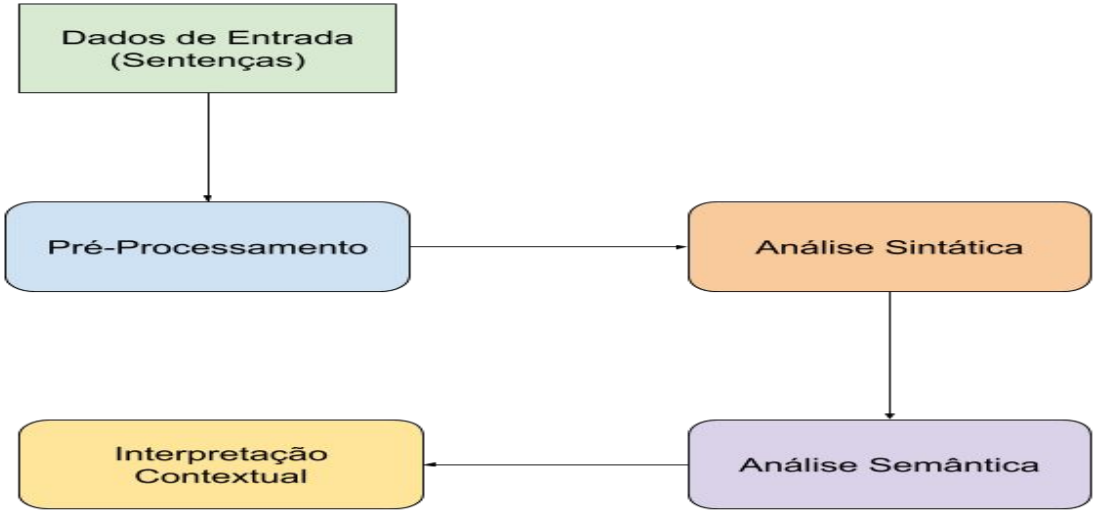
- Características
  - Estrutura Paralela Distribuída
  - Alta Capacidade de Aprendizado
  - Generalização

Fundamentação Teórica - Redes Neurais

- Principais tipos de Redes Neurais
  - Redes Neurais Recorrentes
    - Long Short-Term Memory (LSTM)
    - Bidirectional Long Short-Term Memory (BiLSTM)
  - Redes Neurais Convolucionais

Fundamentação Teórica - PLN

Processamento de Linguagem Natural (PLN) refere-se ao uso de métodos computacionais para processar linguagem escrita ou falada.



Fundamentação Teórica - Extração de Informação

Tarefa de alto nível de PLN, utilizada para extrair dados relevantes de entradas de texto.

- Principais tipos de informações procuradas
  - Entidades
  - Relacionamentos

Fundamentação Teórica - REN

Reconhecimento de Entidades Nomeadas (REN), busca identificar Entidades Nomeadas (EN) em entradas de texto.

Tipo	Tag	Exemplo
Pessoa	**PER**	**Turing** é o pai da computação
Organização	**ORG**	O **Grêmio** é o rei de copas
Local	**LOC**	**Porto Alegre** é demais
Veiculo	**VEH**	O **Uno** é um carro muito bom
Pessoa-Local	**PER-LOC**	**Bento Gonçalves** é muito grande

Fundamentação Teórica - ER

Extração de Relacionamentos (ER), tarefa de extrair ou classificar o relacionamento entre duas EN.

Frase	Relacionamento
André jogador do **Grêmio** está mal.	**Jogador do**
José é o filho mais novo de **João**	**Filho de**
Maria é a fundadora da fintech **Finmoney**	**Fundador de**

Trabalhos Relacionados

- Rede Neural Convolucional
- Rede Neural Recorrente
- NLPyPort
- RelP

Trabalhos Relacionados - Redes Neurais

- Rede Neural Convolucional
  - Sistema desenvolvido pelo MIT
  - Tarefa 10 do SemEval 2017

- Rede Neural Recorrente
  - Desenvolvido por Zhang e Wang
  - Tarefa 8 do SemEval 2010

Trabalhos Relacionados - IberLEF

- NLPyPort
  - Desenvolvido por Ferreira, Gonçalo e Rodrigues
  - Baseado em regras
  - Tarefa de ER do IberLEF 2019

- Relp
  - Desenvolvido por Colovini e Viera
  - Modelo Conditional Random Fields (CRF)

Recursos Utilizados

- Linguagem
  - Python
- Frameworks / Bibliotecas
  - Tensorflow
  - Keras
  - spaCy
  - NLTK
- Plataformas online
  - Google Colab
  - Kaggle

Recursos Utilizados

- Repositório NILC de Word Embeddings
  - 31 modelos treinados de Word Embeddings
  - 17 Corpus
  - 1.395.926.282 tokens
  - Algoritmos: Word2Vec, FastText, Wang2Vec, Glove
  - Dimensionalidades: 50, 100, 300, 600, 1000

IberLEF 2019 - O que é?

- Iberian Languages Evaluation Forum
- União de dois outros workshops de avaliação, TASS e IberEval.
- Encorajar a comunidade da área de pesquisas a organizar tarefas de PLN.

IberLEF 2019 - Tarefas

- Tarefas Disponíveis
  - Análise de Humor (Espanhol)
  - Análise de Sentimento (Espanhol)
  - Detecção de Ironia (Espanhol)
  - Reconhecimento de Entidades Nomeadas (Português e Espanhol)
  - Extração de Relacionamentos (Português e Espanhol)
- Tarefa Selecionada
  - Extração de Relacionamentos (Português)

IberLEF 2019 - Dataset

- Sentenças
  - Treino: **90**
  - Teste: **149**
- Exemplo de Sentença, Relacionamento e Tripla

Sentença	A **Marfinite** fica em o **Brasil**
Relacionamento	fica em
Tripla	(Marfinite, fica em, Brasil)

IberLEF 2019 - Dataset

- Categorias de EN e distribuição nos datasets

Categoria	Sigla	Dataset Treino	Dataset Teste
Organização	**ORG**	119	196
Pessoa	**PER**	31	56
Lugar	**PLC**	30	46

IberLEF 2019 - Dataset

- Número de relacionamentos entre pares de EN

PAR	Dataset Treino	Dataset Teste
ORG-ORG	29	47
ORG-PLC	30	46
ORG-PER	31	56
Total	90	149

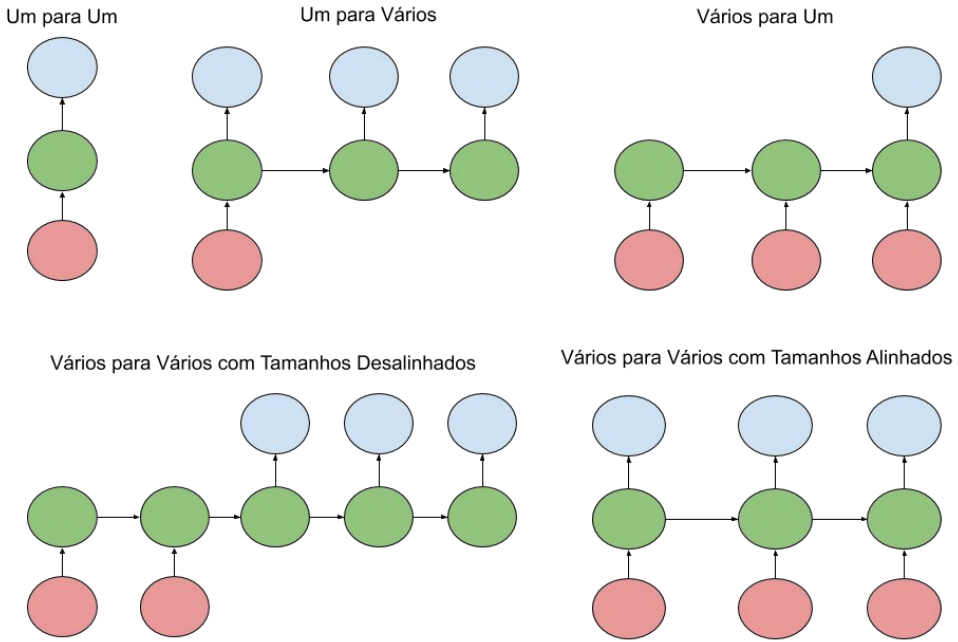
Pré-Processamento

- Padronização de Tamanho
- Identificador de Palavra
- Indicador de EN
- Identificador de POS Tag
- Saída do Sistema

Pré-Processamento

Dado	Representação
Sentença Original	Romildo é o presidente do Grêmio
Tripla	(Romildo, presidente do, Grêmio)
Padronização de Tamanho	Romildo é o presidente do Grêmio <PAD> <PAD>
Identificador de Palavras	6 3 7 45 2 30 0 0
Indicador de EN	1 0 0 0 0 1 0 0
POS Tag	PROPN CONJ DET NOUN CONJ PROPN PAD PAD
Identificador de POS Tag	1 2 3 4 2 1 0 0
Saída do Modelo	0 0 0 1 1 0 0 0

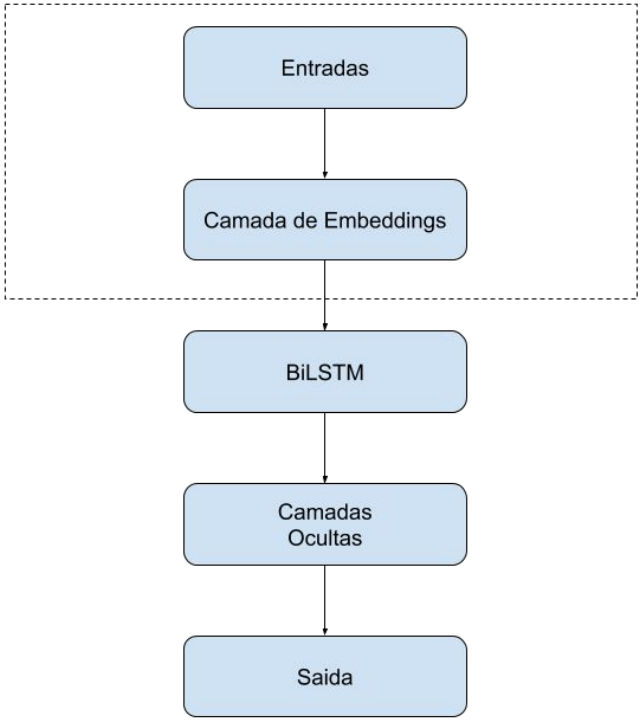
Modelos Desenvolvidos - Arquiteturas



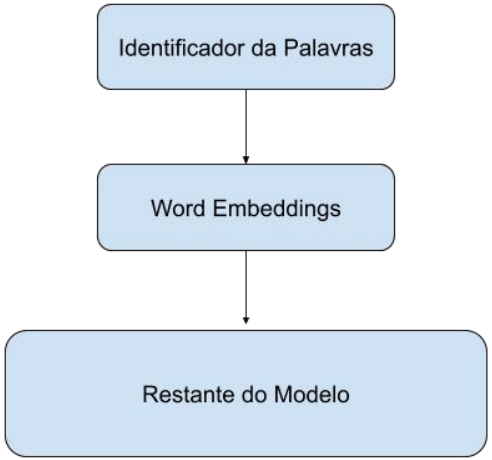
Modelos Desenvolvidos - Especificação

Nome	Entradas
Modelo Simples	Identificador de Palavras
Modelo Intermediário	Modelo Simples + Indicador de EN
Modelo Completo	Modelo Intermediário + Identificador de POS Tag

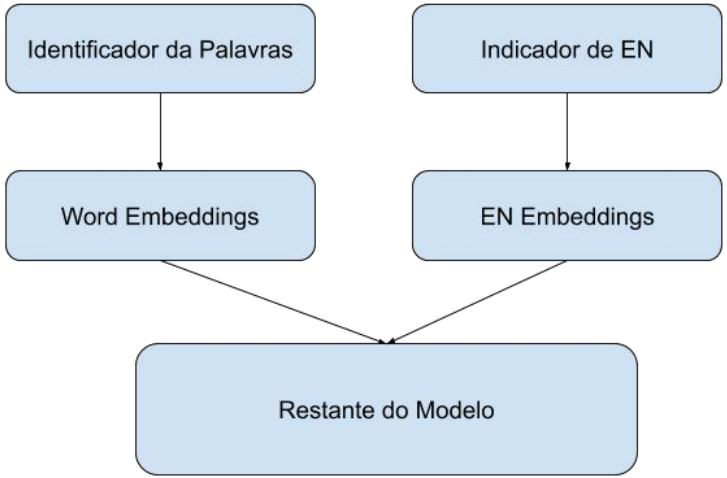
Modelos Desenvolvidos - Modelo Genérico



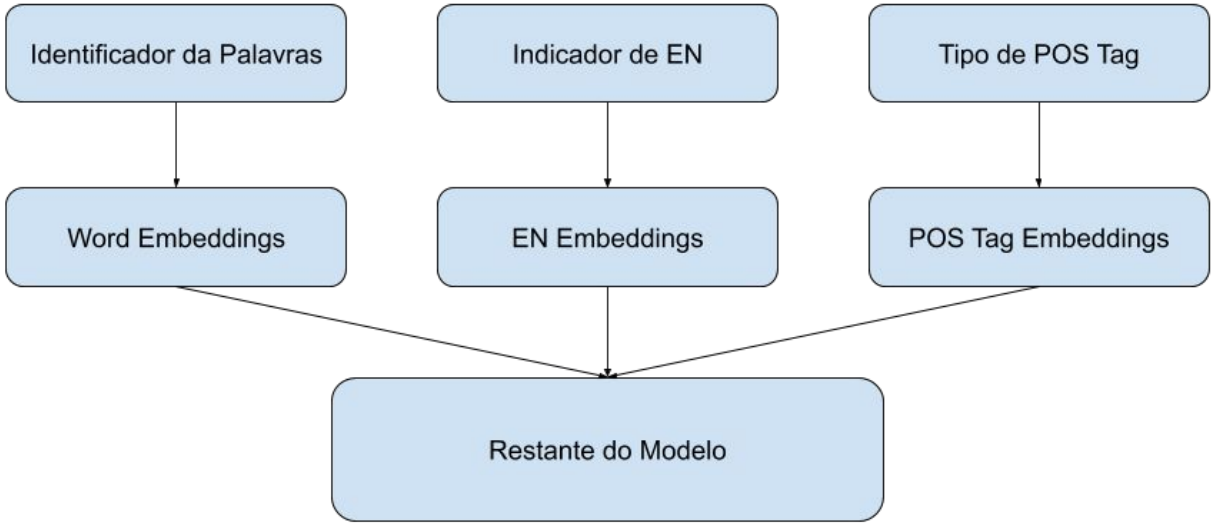
Modelos Desenvolvidos - Simples



Modelos Desenvolvidos - Intermediário



Modelos Desenvolvidos - Completo

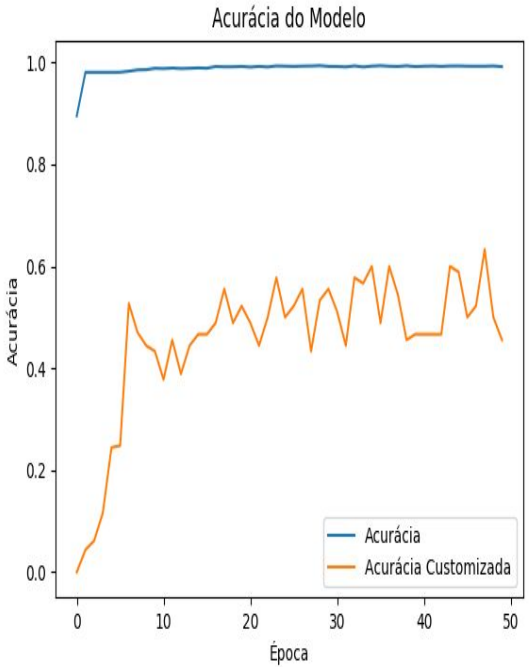


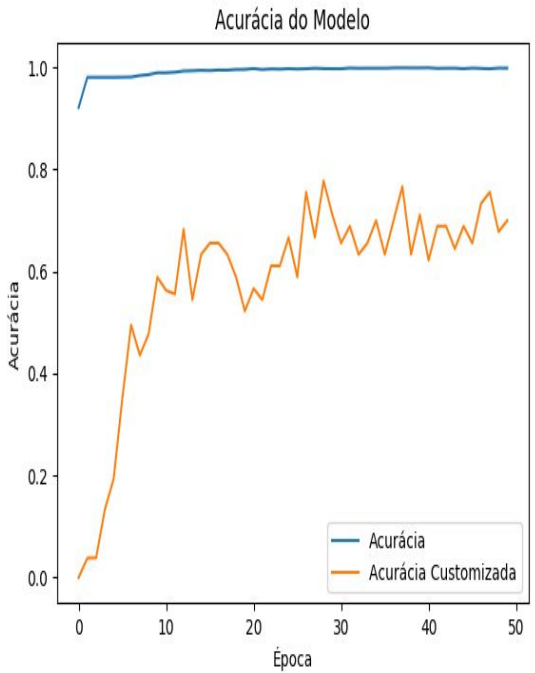
Modelos Desenvolvidos - Hiper-Parâmetros

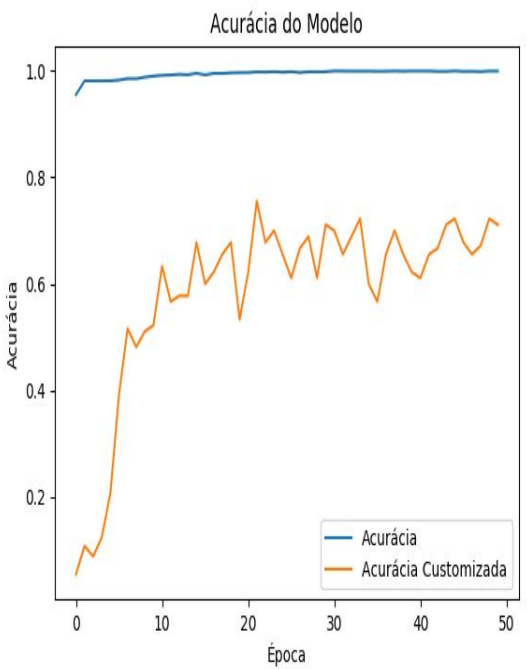
Hiper-Parâmetro	Valor
Épocas	15
Tamanho das Entradas	100 Palavras
Batch de Treino	3 Sentenças
Dropout	50%
Taxa de Aprendizado	1%
Word Embedding	50 Dimensões
EN Embedding	5 Dimensões
POS Tag Embedding	5 Dimensões
Algoritmo de Word Embedding	GloVe

Resultados - Acurácia Treino

Modelo Simples	Modelo Intermediário	Modelo Completo

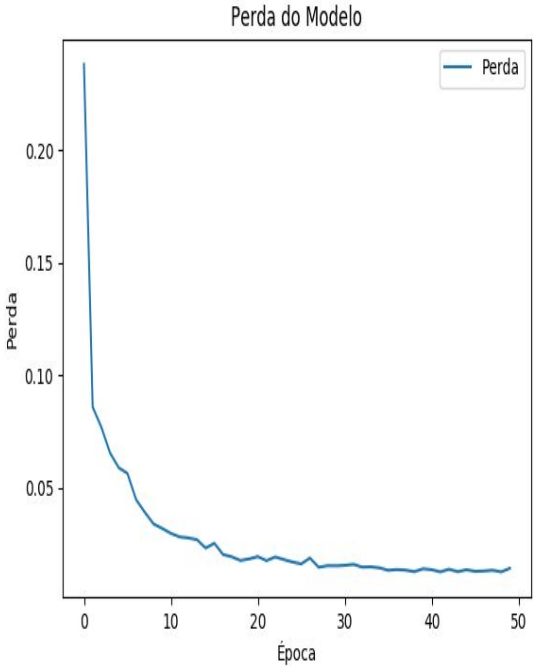


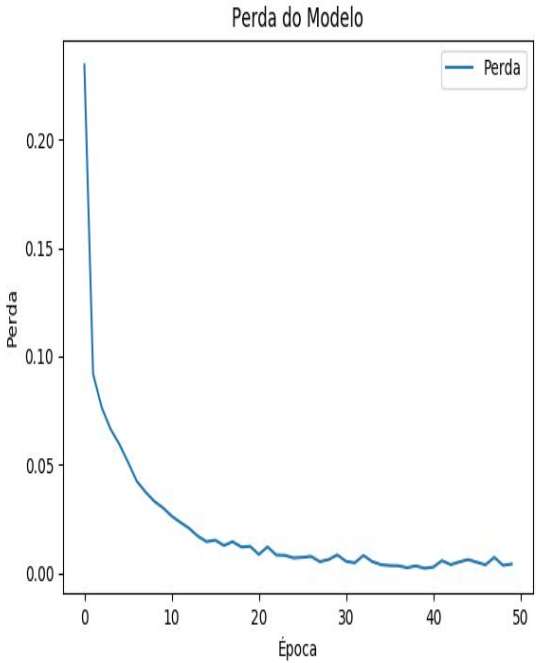


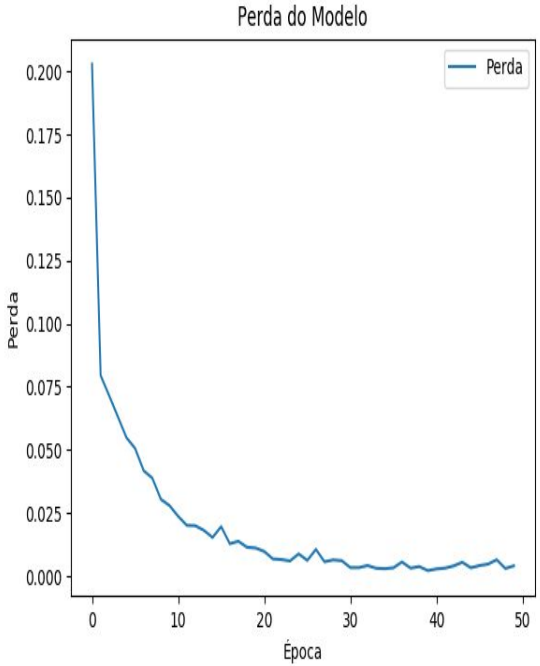


Resultados - Perda Treino

Modelo Simples	Modelo Intermediário	Modelo Completo







Resultados - Métricas IberLEF

- Relacionamentos Completamente Corretos (RCC)
- Relacionamentos Parcialmente Corretos (RPC)
- Relacionamentos Parcialmente Corretos Absolutos (RPCA)
- Total de Relacionamentos no Dataset (TR)
- Relacionamentos Identificados (RI)

Resultados - Métricas IberLEF

- Precisão
  - Proporção de respostas corretas, com a proporção de respostas dadas pelo sistema.
- Recall
  - Proporção de respostas corretas, com a proporção de respostas esperadas no dataset.
- F-Measure
  - Combinação das duas métricas anteriores.

Resultados - Comparação de Sistemas

	Simples	Intermediário	Completo	RelP	NLPyPort
RI	133	141	132	74	144
RCC	28	75	83	46	106
RPCA	61	45	30	21	-
Precisão Ex.	0.210	0.531	0.628	0.621	0.736
Recall Ex.	0.187	0.503	0.577	0.307	0.711
F-Measure Ex.	0.198	0.517	0.590	0.412	0.723
Precisão P.	0.326	0.588	0.684	0.685	0.766
Recall P.	0.288	0.557	0.606	0.340	0.748
F-Measure P.	0.305	0.572	0.642	0.454	0.757

Conclusão

- Conclusão Geral
  - Objetivos atingidos
  - Resultados Satisfatórios
- Trabalhos Futuros
  - Datasets diferentes
  - Customização do Desenvolvimento do Modelo
  - Modelo BiLSTM-CRF

Refência Bibliográfica

1. “Colaboratory - frequently asked questions”. Capturado em: https://research.google.com/colaboratory/faq.html, Junho 2019
2. “General python faq”. Capturado em:https://docs.python.org/3/faq/general.html#what-is-python, Abril 2019
3. “Iberlef 2019 portuguese named entity recognition and relation extraction tasks”.Capturado em: http://www.inf.pucrs.br/linatural/wordpress/iberlef-2019/, Março 2019
4. Amaral, D.; Fonseca, E.; Lopes, L.; Vieira, R. “Comparative analysis of portuguese named entities recognition tools”. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), 2014, pp. 2554–2558

Refência Bibliográfica

5. Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; McCallum, A. “SemEval 2017 task10: ScienceIE - extracting key phrases and relations from scientific publications”. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 546–555
6. Burkov, A. “The Hundred-page Machine Learning Book”. Andriy Burkov, 2019
7. Cimiano, P. “Ontology Learning and Population from Text: Algorithms, Evaluation and Applications”. Berlin, Heidelberg: Springer-Verlag, 2006
8. de Abreu, S. C.; Vieira, R. “Relp: Portuguese open relation extraction”,Knowledge Organization, vol. 44–3, 2017, pp. 163–177
9. Gers, F. “Long short-term memory in recurrent neural networks”, 2001

Refência Bibliográfica

10. Ferreira, J.; Gonçalo Oliveira, H.; Rodrigues, R. “Nlpyport: Named entity recognition with crf and rule-based relation extraction”. In: Iberian Languages Evaluation Forum(IberLEF 2019), 2019
11. Goodfellow, I.; Bengio, Y.; Courville, A. “Deep Learning”. MIT Press, 2016, http://www.deeplearningbook.org.
12. Hartmann, N.; Fonseca, E. R.; Shulby, C.; Treviso, M. V.; Rodrigues, J.; Aluísio, S. M.“Portuguese word embeddings: Evaluating on word analogies and natural language tasks”,CoRR, vol. abs/1708.06025, 2017, 1708.06025
13. Haykin, S. S. “Neural networks and learning machines”. Upper Saddle River, NJ:Pearson Education, 2009, third ed
14. Hochreiter, S.; Schmidhuber, J. “Long short-term memory”,Neural Comput., vol. 9–8,Nov 1997, pp. 1735–1780

Refência Bibliográfica

15. Hearst, M. A. “Automatic acquisition of hyponyms from large text corpora”. In:Proceedings of the 14th Conference on Computational Linguistics - Volume 2, 1992,pp. 539–545.
16. Hendrickx, I.; Kim, S. N.; Kozareva, Z.; Nakov, P.; Ó Séaghdha, D.; Padó, S.;Pennacchiotti, M.; Romano, L.; Szpakowicz, S. “Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals”. In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions,2009, pp. 94–99
17. Jurafsky, D.; Martin, J. H. “Speech and Language Processing (2nd Edition)”. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009
18. Kriesel, D. “A Brief Introduction to Neural Networks”. 2007.

Refência Bibliográfica

19. Lee, J. Y.; Dernoncourt, F.; Szolovits, P. “MIT at SemEval-2017 task 10: Relation extraction with convolutional neural networks”. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 978–984.
20. Li, Y.; Xu, L.; Tian, F.; Jiang, L.; Zhong, X.; Chen, E. “Word embedding revisited: A new representation learning and explicit matrix factorization perspective”. In: Proceedings of the 24th International Conference on Artificial Intelligence, 2015, pp. 3650–3656.
21. Santos, D.; Freitas, C.; Gonçalo Oliveira, H.; Carvalho, P. “Second harem: New challenges and old wisdom”, 2008, pp. 212–215.
22. Sarawagi, S. “Information extraction”,Found. Trends databases, vol. 1–3, Mar 2008,pp. 261–377

Refência Bibliográfica

23. Singh, S. “Natural language processing for information extraction”,CoRR, vol.abs/1807.02383, 2018, 1807.02383
24. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. “Relation classification via convolutional deep neural network”,the 25th International Conference on Computational Linguistics:Technical Papers, 01 2014, pp. 2335–2344
25. Zhang, D.; Wang, D. “Relation classification via recurrent neural network”,CoRR, vol.abs/1508.01006, 2015.