Assessing Transfer Learning and automatically annotated data in the development of Named Entity Recognizers for new domains

Emanuel Matos[1], *Mário Rodrigues*[2], *António Teixeira*[1]

[1]IEETA, DETI, University of Aveiro, Aveiro, Portugal
[2]IEETA, ESTGA, University of Aveiro, Aveiro, Portugal
{easm,mjfr,ajst}@ua.pt

Abstract

With recent advances Deep Learning, pretrained models and Transfer Learning, the lack of labeled data has become the biggest bottleneck preventing use of Named Entity Recognition (NER) in more domains and languages. To relieve the pressure of costs and time in the creation of annotated data for new domains, we proposed recently automatic annotation by an ensemble of NERs to get data to train a Bidirectional Encoder Representations from Transformers (BERT) based NER for Portuguese and made a first evaluation. Results demonstrated the method has potential but were limited to one domain. Having as main objective a more in-depth assessment of the method capabilities, this paper presents: (1) evaluation of the method in other domains; (2) assessment of the generalization capabilities of the trained models, by applying them to new domains without retraining; (3) assessment of additional training with in-domain data, also automatically annotated. Evaluation, performed using the test part of MiniHAREM, Paramopama and LeNER Portuguese datasets, confirmed the potential of the approach and demonstrated the capability of models previously trained for tourism domain to recognize entities in new domains, with better performance for entities of types PERSON, LOCAL and ORGANIZATION.

Index Terms: Named Entities, Named Entity Recognition (NER), Transfer Learning, Automatic Annotation, Portuguese.

1. Introduction

The identification of entities is a key step in many natural language processing (NLP) tasks [1, 2] and consequently named entity recognition (NER) is a task important in several contexts, text genres, and languages. The approaches to NER include statistical models based on handcrafted templates that are instantiated on final rules using training data and, more recently, deep learning models [3, 4] that do not require designing the rules.

The training data for NER models are texts that have token-level labels indicating the boundaries of the entities and preferably their respective class. The creation of those corpora is time-consuming, thus costly, and prone to human error. The lack of labeled data is a relevant bottleneck that prevents NER being effectively used in some domains and tasks. Some techniques have been studied to mitigate this barrier. The most recent proposed methods based on BERT [5] to deal with the label scarcity problem using distant supervision. The idea is to avoid the traditional labeling procedure by locating concepts in knowledge bases such as Wikipedia and YAGO in the target corpus [6, 7].

Our approach avoids the use of annotated data by using as reference the output of an ensemble of 3 general purpose NER [8, 9] as well as an annotated dataset created by linguists [10, 11]. In this paper we evaluate the performance in a new domain. In this paper, the initial work for the Tourism domain is extended, including a new domain, cross-domain assessment of the trained models and tuning of the models to a new domain. We report the results obtained for 3 datasets, including one of legal documents, which are a difficult challenge due to the differences in the writing style when compared to most of other domain documents (e.g., profusion of capitalized words).

This document is organized as follows: after this Introduction follows the second section elaborating on the relevant Related Work. In the third section the proposed Method is explained followed by Results in section 4. The paper ends in section 5 with the Conclusion.

2. Related Work: NER for Portuguse

Named Entity Recognition, as many tasks in NLP, has been approached in two very different ways: systems based in rules and lists (the so-called Gazetteers) and data-driven systems base in machine learning. Machine Learning approaches to NER can be more flexible but they depend on the existence of adequate datasets for the target domain.

Through the years several machine learning methods have been applied to NER, being good results obtained with, for example, Support Vector Machines (SVM), Conditional Random Field (CRF) or Neural networks (NN) (e.g., [16]). Advances in Deep Learning, combined with larger datasets and increased computational power, resulted in major advances in NER systems, with systems based on Long Short-Term Memory (LSTM), Bidirectional LSTMs (Bi-LSTM) and Transformers (e.g., [6, 4]). These advances resulted, among others, from: LSTMs recursiveness providing better capabilities to deal with long range dependencies in data [17, 3]; Transformers [18], introduced in 2017, attention mechanism designed to handle sequential input data; and their potential for parallelization that enabled training using huge datasets. They created the conditions for the appearance of pretrained systems such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) [19].

Following the general tendency, for Portuguese, the last years main developments on NER were systems featuring machine learning. A representative selection of works is presented in Table 1.

NER systems for PT have recently explored CRF, LSTMs and, since 2020, BERT. The initial use of BERT in NER for Portuguese, in 2020 [15], adopted a BERT-CRF architecture, combining transfer capabilities of BERT with the CRF forecasts. BERT was pre-trained using the brWac corpus (2.68 billion tokens), and the NER model trained with the First HAREM and tested with MiniHAREM. This system advanced state-of-the-art performance despite being trained with much less data.

Table 1: *Recent related work on NER for Portuguese.*

System	Ref	Year	Models	Datasets	Manual annot.	Performance
LeNER Br	[12]	2018	LSTM CRF	Paramopama (train) LeNER (test)	YES	F1 = 97.0 % (legislation) F1 = 88.0 % (judicial)
Pirovani et al.	[13]	2019	CRF + Grammar	NER task IberLEF 2019	YES	F1 = 56.5 % (overall)
Lopes et al.	[14]	2019	Bi LSTM	Clinical data	YES	F1 $>$ 80 %
Souza et al.	[15]	2020	BERT	HAREM	YES	F1 = 83.3 % (selective)
Matos et al.	[9]	2022	BERT	Tourism (Wikivoyage) HAREM	NO	F1 = 64.9 % (word based) F1 = 47.1 % (entity based)

To the best of our knowledge all these machine learning systems were trained with data annotated by humans, at least in a revision stage. This is a major limitation in the development of NER for new domains, which are in large demand by the expansion of potential application areas. To address this limitation, the authors, in [9], developed NER systems for PT based on BERT using automatically annotated data. They adopted Transfer Learning, finetuning pretrained BERT models with an automatically annotated dataset for the Tourism domain, based in Wikivoyage texts. Best F1 obtained was 64.9 %.

3. Method

3.1. Overview

The process is summarized in Fig. 1 and the main blocks are described in the following subsections.

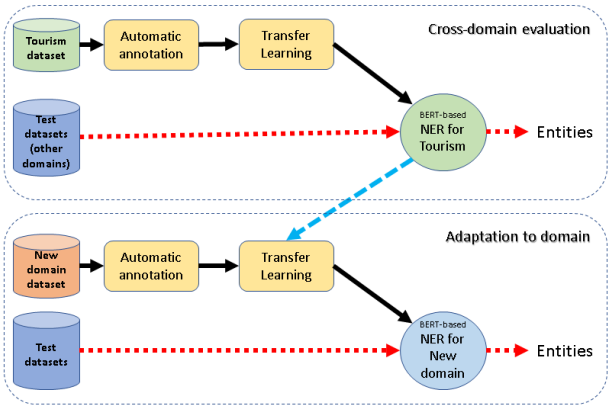

Figure 1: *Overview of the process adopted to explore the cross-domain potential of BERT-based NER models with and without tuning with in-domain additional training.*

The process consists of two parts:

1. Cross-domain application of a BERT-based NER model previously trained with automatically annotated data (top part of the Figure);
2. Fine-tuning of this base model to a new domain resorting to automatically annotated data for that domain.

For this initial exploration of cross-domain BERT-base NERs, the previously explored domain of Tourism was selected (see [9, 8]) and complemented with the Legal domain.

3.2. Datasets

For this work, Portuguese (European and Brazilian variants) datasets regarding 3 domains were used:

Wikivoyage Tourism dataset – resulting from automatic annotation of texts from Wikivoyage[1][20] a tourism dataset, in development by the authors and previoulsy used for the initial proof-of-concept of the NER training with automatically annotated data[8].

LeNER Legal texts dataset – Consisting of 60 texts from LeNER dataset of legal documents in Brazilian Portuguese [12, 21]. An excerpt of one of the texts is presented in Table 2.

The texts of this dataset were extracted and processed to automatically derive NER tags. The method we proposed in [8] was adopted. It consists in the application of 3 existing NERs and combination of the outputs.

Due to the different approaches adopted by the 3 NERs for tokenization, the outputs of the 3 NERs need non-trivial alignment. For that, an alignment process was developed. Briefly, it consists of: (1) selecting one of the NERs as reference, preferably the one producing more tokens; (2) aligning each of the other two NERs with the reference; (3) using the alignment information for the 2 pairs of previous step, align the outputs of the 3 NERs. Alignment of a pair of NERs was performed using a Python implementation of the Needleman-Wunsch algorithm [22, 23].

Paramopama datasets – Extends the PtBR version of WikiNER corpus, revising incorrect assigned tags in order to improve corpus quality, also extend the corpus size and provide proper evaluation[24]. This dataset has a total of 240,755 words tagged as part of an entity, considering 4 types (PERSON, LOCATION, ORGANIZATION and TIME).

HAREM datasets – Two HAREM [10, 11] datasets were used, the First HAREM and the MiniHAREM, both having manually annotated entities. Both text and BIO annotated files were obtaining using the processing and XML made available by authors of [15][2]. First HAREM was used in the training of the model for Tourism and MinHAREM, that provides 21.901 tagged words, in the evaluations reported in this paper.

3.3. BERT-base NER for Tourism domain

As basis for the work presented, the NERs obtained in recent work of the authors for the tourism domain were selected. They are based in BERT [25] implementation by Tobias Sterbak [26] using the Transformers package by Huggingface [27], Keras and TensorFlow. For the sake of completeness, the models were

[1] https://pt.wikivoyage.org/

[2] https://github.com/neuralmind-ai/portuguese-bert/tree/master/ner_evaluation

Table 2: *Small excerpt of one of LeNER dataset texts, showing the profusion of capitalized words.*

irregularidades nas obras de construção do prédio da 1ª Circunscrição Judiciária Militar no Rio de Janeiro.
Sumário RECURSO DE REVISÃO INTERPOSTO PELO MINISTÉRIO PÚBLICO JUNTO AO TCU CONTRA ACÓRDÃO QUE JULGOU REGULARES COM RESSALVA AS CONTAS DO STM DE 1999.
SUPERVENIENTE CONSTATAÇÃO, EM PROCESSO DE TOMADA DE CONTAS ESPECIAL, DE IRREGULARIDADES OCORRIDAS NO MESMO EXERCÍCIO.
LIQUIDAÇÃO IRREGULAR DE DESPESA E DANO AO ERÁRIO DECORRENTE DE PAGAMENTOS ANTECIPADOS PARA EXECUÇÃO DAS OBRAS DO EDIFÍCIO SEDE DA 1ª CIRCUNSCRIÇÃO JUDICIÁRIA MILITAR NA ILHA DO GOVERNADOR RJ.
CONHECIMENTO DO RECURSO E PROVIMENTO.

obtained as follows: Fine-tuning of a pretrained BERT models with the results of our recent work on automatic annotation [8] as data to train the models.

Briefly, a corpus based on Wikivoyage [20] - was annotated by 3 NERs (Linguakit NER [28, 29], Alen NLP [30, 31] and a DBPedia-based NER developed by the authors [8]). The output tags of these 3 NERs were combined to tag a word as part of an ENTITY or not, without including classification of the entity. More details can be found in [9, 8].

3.4. Cross-domain evaluation

To assess the potential of the pretrained models in a new domain, they were used without any additional training to annotate all but the dataset for the tourism domain.

3.5. Fine-tuning NER model to a new domain

Additional training was applied to the BERT model pretrained with Tourism data, using part of the LeNER legal dataset. Due to problems with 3 of the 60 texts of this dataset, 47 texts were used for training and 10 for test. To make possible evaluation of the effect of the amount of additional training data, models were obtained with 20%, 40%, 60%, 80% and all of the tokens of the training data. As a result 5 new models were obtained. They were assessed following a similar process to the one used for the initial model for Tourism domain.

As for the Tourism model, training was performed using a GPU NVIDIA GeForce RTX 2060 and adopting AdamW optimizer with the default parameters of [26], `lr`=3×10^{-5} and `eps`=10^{-8}. Also, the training datasets were split into training and validation, with 90% for training and 10% for validation. The stop criteria adopted were a maximum of 10 epochs or the increase of loss in the validation set.

4. Results

The main results obtained are summarized in this section. The performance was evaluated using the standard metrics (Precision, Recall and F1) considering the individual words of the entity.

4.1. Global results

The results for the selected metrics as function of the NER model and test-set are presented in Table 3. For easier understanding of the results, part of the information in the Table is presented graphically in Figure(s) 2 and 3.

Table 3 and plots show that:

- The results obtained for LeNER legal dataset are worst than the ones obtained for the two other test sets, being particularly worst in terms of recall;
- Results for Paramopama are close to the one obtained for MiniHAREM, confirming for a new dataset the results reported in [9].

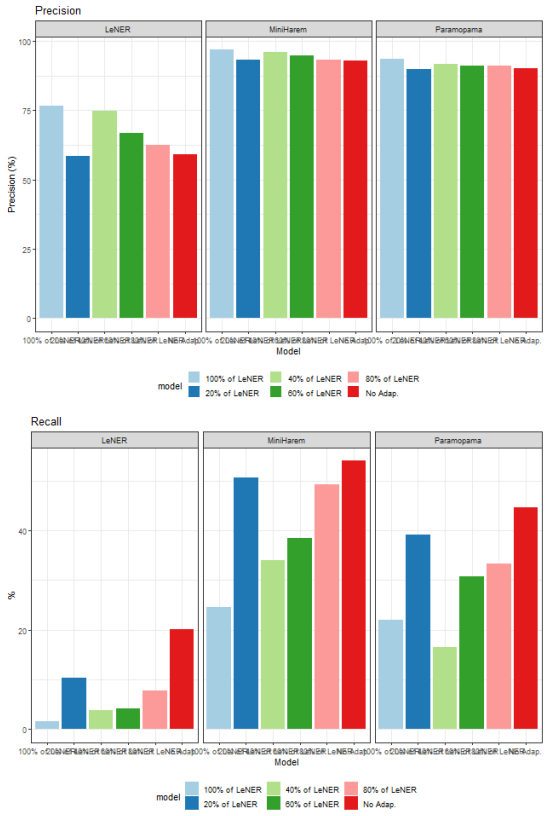

Figure 2: *Evaluation results (Precision and Recall) by model and test set.*

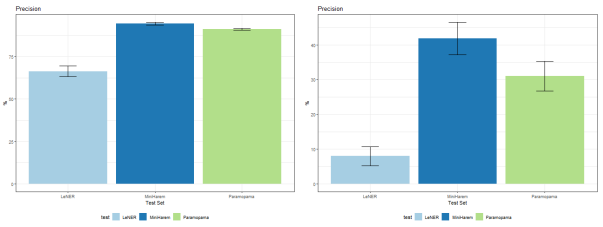

Figure 3: *Evaluation results as function of test set: Precision (at left) and Recall.*

- Despite showing some reduced contribute to improve Precision, the additional training (fine-tuning) of the model with LeNER train dataset resulted in degradation of Recall, for all 3 test sets. The best Recall is obtained without additional training for the 3 datasets, being the best overall score obtained with MiniHAREM.

Table 3: *Evaluation results for the 6 models with the 3 datasets. Values in percentage.*

Domain Adaptation	**Paramopama**			**LeNER (test set)**			**MiniHAREM**		
	Prec.	Rec.	F1	Prec	Rec	F1	Prec.	Rec.	F1
No Adaptation	90.33	44.53	59.65	59.04	20.17	30.07	92.83	53.98	68.26
20 % of LeNER	89.75	39.10	54.47	58.52	10.38	17.63	93.25	50.58	65.59
40 % of LeNER	91.67	16.51	27.98	74.87	3.85	7.32	96.12	33.92	50.14
60 % of LeNER	91.25	30.72	45.97	66.67	4.20	7.90	94.93	38.44	54.72
80 % of LeNER	91.00	33.36	48.82	62.56	7.74	13.78	93.17	49.24	64.43
100 % of LeNER	93.68	22.06	35.71	76.79	1.59	3.12	96.87	24.49	39.10

4.2. Performance by entity type

As the datasets contain several types of entities, some not annotated in the datasets used to train the initial model, it is worth looking in detail at the performance by class. The Precision and Recall by entity class is presented graphically in Fig. 4. F1 showed similar results to Recall.

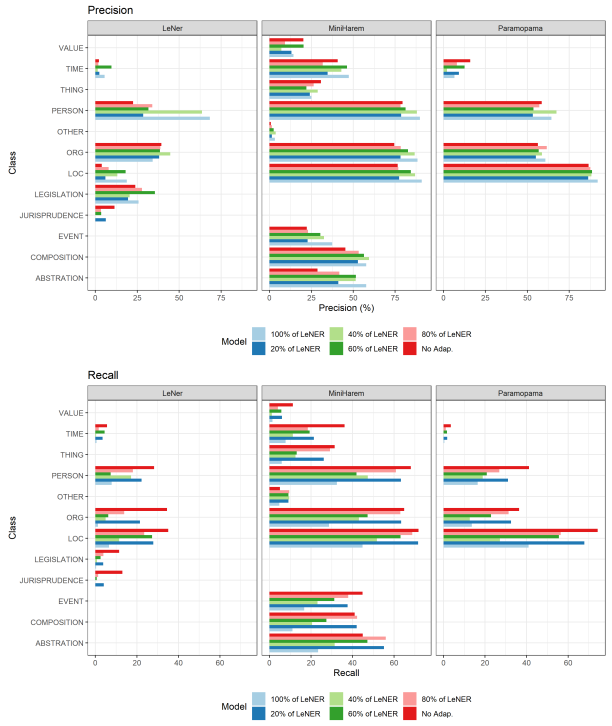

Figure 4: *Precision (top) and Recall (bottom) by entity class and dataset.*

As in our previous experiments, the best results were obtained for PERSON, ORGanization and LOCation, being lower for LeNER. The best Precision in LeNER was obtained for PERSON, with values similar to Paramopama and close to 70%.

Recall, in general lower than Precision, presents high variation across entity classes, and is more affected than Precision by the additional training.

Entity classes specific of a dataset (not part of the training for the initial model) present some recall. Despite with lower precision and recall when comparing with MiniHAREM EVENT, COMPOSITION and ABSTRACTION, some interesting results were obtained for LeNER even without any in-domain training, namely: precision around 25% for LEGISLATION; recall of more than 10% for both LEGISLATION and JURISPRUDENCE.

5. Conclusion

Addressing the problem of development of NERs for new domains without annotated datasets, this paper explores the cross-domain potential of application of BERT-based NER models with and without tuning with in-domain additional training.

Good results were obtained without any additional training with the Paramopama dataset, corroborating the previous results pointing to the capability of generalization of the BERT-based models. The results for the legal domain (LeNER dataset) were lower. Nevertheless, without any in-domain training, models maintained a Recall above 30% for classes PERSON, ORGANIZATION and LOCATION, that we can qualify as interesting considering the specificities of this dataset, particularly the exaggerated use of capitalized words (see example in Table 2, a major challenge as capitalization is a common feature for entity detection.

As expected, models had a much higher difficulty in recognizing entities of classes not present in the training dataset(s). Nevertheless, the Recall of approx. 10% for both LEGISLATION and JURISPRUDENCE in LeNER without any in-domain training is an interesting result, with potential to be explored in bootstrap methods.

Additional training with a small amount of in-domain data (automatically annotated) was not very useful. Despite some improvement in Precision (e.g., for PERSON in LeNER), it contributed to degradation of Recall.

5.1. Future work

The results point to the potential of the approach but many challenges an limitations remain. Future work will include: (1) improvement of the quality of the dataset obtained for the new domain, by post-processing of the annotations to solve problems detected (e.g., first word of a sentence considered an entity); (2) development of new models specialized in only one entity class (e.g., PERSON) instead the multi-class models explored so far; (3) exploration of the potential of bootstrapping methods in the creation of NER for new classes introduced by a new domain (as the LEGISLATION in LeNER dataset); (4) exploration of recent alternatives to BERT, such as GPT-3 or FLAN [32, 33];

6. Acknowledgement

This research was supported by IEETA Research Unit, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020.

7. References

[1] Z. Zhu, Y. Zhou, X. Deng, and X. Wang, "A graph-oriented model for hierarchical user interest in precision social marketing," *Electronic Commerce Research and Applications*, vol. 35, p. 100845, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1567422319300225

[2] M. Ketsmur, A. Teixeira, N. Almeida, S. Silva, and M. Rodrigues, "Conversational assistant for an accessible smart home: Proof-of-concept for portuguese," in *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, ser. DSAI 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 55–62.

[3] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[4] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1064–1074.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang, "BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1054–1064.

[7] F. Mahdisoltani, J. Biega, and F. M. Suchanek, "YAGO3: A Knowledge Base from Multilingual Wikipedias," in *CIDR*, Asilomar, United States, Jan. 2013. [Online]. Available: https://hal-imt.archives-ouvertes.fr/hal-01699874

[8] E. Matos, M. Rodrigues, P. Miguel, and A. Teixeira, "Towards Automatic Creation of Annotations to Foster Development of Named Entity Recognizers," in *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*, ser. Open Access Series in Informatics (OASIcs), R. Queirós, M. Pinto, A. Simões, F. Portela, and M. J. a. Pereira, Eds., vol. 94. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, pp. 11:1–11:14. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2021/14428

[9] ——, "Named Entity Extractors for New Domains by Transfer Learning with Automatically Annotated Data," in *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, ser. LNCS, V. Pinheiro, P. Gamallo, R. Amaro, C. Scarton, F. Batista, D. Silva, C. Magro, and H. Pinto, Eds. Springer, 2022, pp. 288–298, https://link.springer.com/chapter/10.1007/978-3-030-98305-5_27.

[10] D. Santos, N. Seco, N. Cardoso, and R. Vilela, "HAREM: An advanced NER evaluation contest for Portuguese," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, and D. Tapias, Eds., 2006.

[11] C. Mota and D. Santos, Eds., *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, iSBN: 978-989-20-1656-6. [Online]. Available: http://www.linguateca.pt/LivroSegundoHAREM

[12] P. H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "LeNER-Br: a dataset for named entity recognition in Brazilian legal text," in *PROPOR*, ser. LNCS. Springer, 2018.

[13] J. P. Pirovani, J. Alves, M. Spalenza, W. Silva, C. da Silveira Colombo, and E. Oliveira, "Adapting NER (CRF+ LG) for many textual genres." in *IberLEF@ SEPLN*, 2019, pp. 421–433.

[14] F. Lopes, C. Teixeira, and H. G. Oliveira, "Contributions to clinical named entity recognition in portuguese," in *Proc. 18th BioNLP Workshop and Shared Task*, 2019.

[15] F. Souza, R. Nogueira, and R. Lotufo, "Portuguese Named Entity Recognition using BERT-CRF," 2020.

[16] A. Goyal, V. Gupta, and M. Kumar, in *International Conference on Advanced Informatics for Computing Research*. Springer, 2019, pp. 184–195.

[17] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[19] A. Patel and A. Arasanipalai, *Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand*. O'Reilly Media, Incorporated, 2021.

[20] "Wikivoyage." [Online]. Available: https://pt.wikivoyage.org/

[21] P. H. L. de Araujo, T. E. de Campos, R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "Lener-br: a dataset for named entity recognition in brazilian legal text," in *PROPOR*. Springer, 2018.

[22] J. Lekberg, "Solving the sequence alignment problem in python," October 25, 2020, https://johnlekberg.com/blog/2020-10-25-seq-align.html.

[23] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," *Machine learning*, vol. 34, no. 1, pp. 211–231, 1999.

[24] C. M. Júnior, H. Macedo, T. Bispo, F. Santos, N. Silva, and L. Barbosa, "Paramopama: a brazilian-portuguese corpus for named entity recognition," *Encontro Nac. de Int. Artificial e Computacional*, 2015.

[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[26] T. Sterbak, "Named entity recognition with BERT," 2018, last updated: 2020-04-24. Accessed 24 october 2021. [Online]. Available: https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[28] "Linguakit full," Jun 2020. [Online]. Available: https://linguakit.com/en/full-analysis

[29] P. Gamallo and M. Garcia, "Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação," *Linguamática*, vol. 9, no. 1, pp. 19–28, 2017.

[30] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A deep semantic natural language processing platform," 2018.

[31] "Allen NLP - An Apache 2.0 NLP research library, built on PyTorch, for developing state-of-the-art deep learning models on a wide variety of linguistic tasks." [Online]. Available: https://github.com/allenai/allennlp

[32] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[33] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.