Assessing the Effectiveness of Multilingual Transformer-based Text Embeddings for Named Entity Recognition in Portuguese

Diego Bernardes de Lima Santos[1], Frederico Giffoni de Carvalho Dutra[2][a], Fernando Silva Parreiras[3][b] and Wladmir Cardoso Brandão[1][c]

[1]*Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil*

[2]*Companhia Energética de Minas Gerais (CEMIG), Belo Horizonte, Brazil*

[3]*Laboratory for Advanced Information Systems, FUMEC University, Belo Horizonte, Brazil*

Keywords: Named Entity Recognition, Text Embedding, Neural Network, Transformer, Multilingual, Portuguese.

Abstract: Recent state of the art named entity recognition approaches are based on deep neural networks that use an attention mechanism to learn how to perform the extraction of named entities from relevant fragments of text. Usually, training models in a specific language leads to effective recognition, but it requires a lot of time and computational resources. However, fine-tuning a pre-trained multilingual model can be simpler and faster, but there is a question on how effective that recognition model can be. This article exploits multilingual models for named entity recognition by adapting and training tranformer-based architectures for Portuguese, a challenging complex language. Experimental results show that multilingual trasformer-based text embeddings approaches fine tuned with a large dataset outperforms state of the art trasformer-based models trained specifically for Portuguese. In particular, we build a comprehensive dataset from different versions of HAREM to train our multilingual transformer-based text embedding approach, which achieves 88.0% of precision and 87.8% in F1 in named entity recognition for Portuguese, with gains of up to 9.89% of precision and 11.60% in F1 compared to the state of the art single-lingual approach trained specifically for Portuguese.

1 INTRODUCTION

Natural Language Processing (NLP) is a computer science research field with several practical applications, such as automatic text reading and question answering, audio content interpretation, document classification, and predictive text analysis. Usually, NLP systems perform a set of basic preprocessing tasks on input text, such as parsing, tokenization, stop-words removal, stemming and tagging. Particularly, Named Entity Recognition (NER) is a NLP tagging task that extracts important information by marking up it on text, such as names of people, places and currency values (Borthwick, 1999). The extracted elements are relevant entities in the textual content that make sense within a context. For instance, the recognition of the entity "New York" as a location in a sentence can be important to detect where a particular event occurred or even to relate that location to other locations, dealing with similar entities or with entities with the same semantic value.

NER is strongly dependent on the context, i.e., words or expressions can be recognized as different types of entity in different contexts. For instance, in the sentence "Mary prays to Saint Paul for health", the expression "Saint Paul" refers to a person (religious entity), but in the sentence "We will move to Saint Paul next year", the expression "Saint Paul" refers to a place (location entity). Even if the spelling of a word or expression cited in different sentences is identical, the meaning can be distinct given different contexts. Additionally, sentences are formulated in distinct ways in different languages, and the languages differ from each other in structure, form and complexity, which impose even more challenging issues for NER.

Traditional NER approaches use hand-crafted linguistic grammar-based strategies or statistic models that requires a large amount of manually annotated training data to recognize entities in text (Marsh and Perzanowski, 1998). For years, Conditional Random Fields (CRF) has been the state of the art strategy for

[a] https://orcid.org/0000-0002-8666-0354
[b] https://orcid.org/0000-0002-9832-1501
[c] https://orcid.org/0000-0002-1523-1616

Santos, D., Dutra, F., Parreiras, F. and Brandão, W.
Assessing the Effectiveness of Multilingual Transformer-based Text Embeddings for Named Entity Recognition in Portuguese.
DOI: 10.5220/0010443204730483
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 473-483
ISBN: 978-989-758-509-8

NER, taking context into account in a learning model that support sequential dependencies between predictions (Lafferty et al., 2001). Recently, deep neural networks based approaches have achieved even more effective results than CRF for NER (Goldberg, 2016). They learn distributed text representations (text embeddings) from a huge amount of text to build a language model that can be effectively used in several NLP tasks, including NER.

Deep neural single-lingual models (training NLP models in a specific language) usually leads to effective entity recognition, requiring a lot of time and computational resources for training. In addition, such single-lingual approaches require a large amount of data in each specific language for training, sometimes not available or easily obtained for certain languages. However, fine-tuning a pre-trained multilingual model can be cheaper, simpler and faster, requiring no specific single-language training dataset and less time and computational resources for training. But how effective multilingual NER models can be compared to single-lingual models, particularly for complex languages, such as Portuguese?

In this article, we exploit multilingual models for NER by adapting and training transformer-based text embeddings for named entity recognition in Portuguese. Particularly, we propose a NER approach by training and fine tuning a multilingual transformer-based NLP model using a comprehensive dataset we created by combining different versions of HAREM. Additionally, we evaluate our proposed approach by contrasting it with the state-of-the-art (SOTA) single-lingual approach for NER in Portuguese.

Experimental results show that our multilingual approach for NER in Portuguese outperforms the SOTA single-lingual approach with gains of up 9.89% of precision and 11.60% in F1, achieving 88.00% of precision and 87.80% in F1 in named entity recognition. The main contributions of this article are:

- We propose a comprehensive dataset to improve the training of NER models for Portuguese by combining different versions of the HAREM dataset.
- We propose a multilingual NER approach for Portuguese by adapting and training different transformer-based neural networks for multilingual NER in English.
- We provide a throughout evaluation of our proposed approach by contrasting them with the SOTA single-lingual approach for NER in Portuguese reported in literature.

The present article is organized as follows: Section 2 presents the theoretical background in named entity recognition, word embeddings and transformer-based architectures of neural networks. Section 3 presents related work reported in literature for NER, including the state-of-the-art approach for NER in Portuguese. Section 4 presents our multilingual NER approach for Portuguese, as well as the comprehensive dataset we create to improve the training of our approach. Section 5 presents the experimental setup and the results of the experiments we carry out to evaluate our proposed approach. Finally, Section 6 concludes this article, suggesting directions for future work.

2 BACKGROUND

Named Entity Recognition (NER) is a NLP task that identifies people, location, currency, and other relevant information within a text (Borthwick, 1999). While traditional NER approaches use hand-crafted linguistic grammar-based strategies or statistic models that require a large amount of manually annotated training data to recognize entities in text (Marsh and Perzanowski, 1998), recent NER approaches use deep neural networks to learn an effective recognition model (Goldberg, 2016). In particular, they learn text embeddings from a huge amount of text to build a language model that can be effectively used for NER.

2.1 Word Embeddings

Recently, different ways to represent text have emerged, allowing more accurate analyzes of textual information, e.g., the analysis of similarity between two words. A distributed text representation, or text embeddings, can be generated by deep neural network (NN) approaches that learn language models from a huge amount of natural language corpus. In particular, word embeddings take the form of a continuous vector representation describing the meaning of terms (Levy and Goldberg, 2014). Usually, this distributed representation is a not mutually exclusive continuous real-valued vector of fixed length learned by a NN, typically much smaller than the size of the vocabulary (Bengio et al., 2003).

The continuous vectors representation are capable of syntactically representing words, but also allow the learning of semantic values of terms, that is, word embeddings can capture similarity between words with similar meaning, even if their spelling is quite different among them (Mikolov et al., 2013b). Figure 1 presents groups of words with similar context measured by cosine similarity between word embeddings.

In recent years, different frameworks and algorithms for word embeddings generation have been

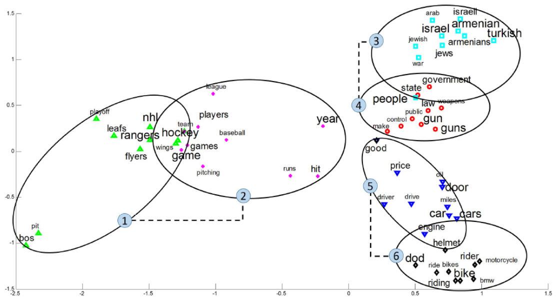

Figure 1: Correlation among words represented as continuous vectors measured by cosine similarity. Source: (Xun et al., 2017).

proposed, particularly WORD2VEC, GLOVE, and FASTTEXT. WORD2VEC (Mikolov et al., 2013a; Mikolov et al., 2013b) is a framework composed of the first efficient word embeddings models, particularly the continuous BoW (CBOW) and the continuous skip-gram (SKIP-GRAM), to learn distributed representations of words from large amount of unstructured text with billions of words. Training such models does not require dense matrix multiplications and can be done in one day on a hundred billion words dataset with a single machine. Particularly, the CBOW model is a simplification of the first practical neural language model approach proposed in literature (Bengio et al., 2003) that uses a fully connected feedforward neural network to learn simultaneously a distributed representation for words and the joint probability distribution function for these word representations from a huge corpus of natural language text with millions of words.

In the CBOW model the non-linear hidden layer is removed, the projection layer is shared for all words, and word context is captured by a log-linear classifier trained to predict a target word given its two previous and two next neighboring words. The SKIP-GRAM model is similar to CBOW but the log-linear classifier is trained to predict the two previous and two next neighboring words given a target word (Mikolov et al., 2013a). Additionally, the models perform subsampling of frequent words, resulting in faster training and improved representations of uncommon words. Moreover, they use two replacement training methods for full softmax resulting in speedup and accurate distributed representations especially for frequent words (Mikolov et al., 2013b). The replacement training methods are hierarchical softmax (Morin and Bengio, 2005) and negative sampling, a simplified NCE (Gutmann and Hyvärinen, 2012).

An interesting property of the word embeddings learned by WORD2VEC models is that simple vector operations can often produce meaningful results. For instance, the sum operation between the vector(*usa*) and the vector(*capital*) results in a vector close to the vector(*washington*). Additionally, word embeddings can be combined using simple operations to represent longer pieces of text, such as sentences, paragraphs and documents. For instance, the vector(*boston*) and the vector(*globe*) can be combined to get the vector(*boston globe*). However, the resulting word embeddings is often unable to represent idiomatic sentences that are not compositions of the individual words, such as *boston globe*. Moreover, the word embeddings learned by WORD2VEC models exhibit linear structure that makes precise analogical reasoning possible. For instance, the vector(*queen*) being the nearest representation of the vector(*king*) minus the vector(*man*) plus the vector(*woman*) provide a way to test the analogy pair *man*:*king*::*woman*:*queen*.

GLOVE incorporates global statistics of words occurrences typically captured by count-based language models in a log-bilinear model for unsupervised learning of word embeddings (Pennington et al., 2014). The intuition is that shallow window models, such as SKIP-GRAM, poorly utilize statistics of the corpus since they train on local context window instead of on global co-occurrence counts. Therefore, training a NN model simultaneously on local context and on global word-word co-occurrence counts, making efficient use of statistics, produces word embeddings with meaningful substructure.

FASTTEXT is another simple and unsupervised approach that learns distributed representations by considering subword units and representing words by a sum of their character n-grams (Bojanowski et al., 2017). It is an extension of the continuous skip-gram model (Kiros et al., 2015) that incorporates n-grams, taking into account the internal structure of words, which is important for morphologically rich languages where many word formations follow rules. For instance, in Latin languages most verbs have more than dozens different inflected forms. These languages contain many word forms that occur rarely (or not at all) in the training corpus, making it difficult to learn good word representations. Additionally, FASTTEXT is capable of building word vectors for words that do not appear in the training set. Experimental results on word similarity and word analogies tasks show that FASTTEXT outperforms WORD2VEC models that do not take into account subword information, as well as methods relying on morphological analysis in different languages (Joulin et al., 2016).

2.2 Transformers

Transformers are sequence transduction models based exclusively on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention, consequently allowing more parallelization (Vaswani et al., 2017). In particular, it follows an encoder-decoder structure using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, where the encoder maps an input sequence of symbol representations to a sequence of continuous representations feeding the decoder that generates an output sequence of symbols one element at a time. At each step the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next. Experimental results on machine translation and English constituency parsing show that Transformers outperform baseline discriminative models at a fraction of the training cost.

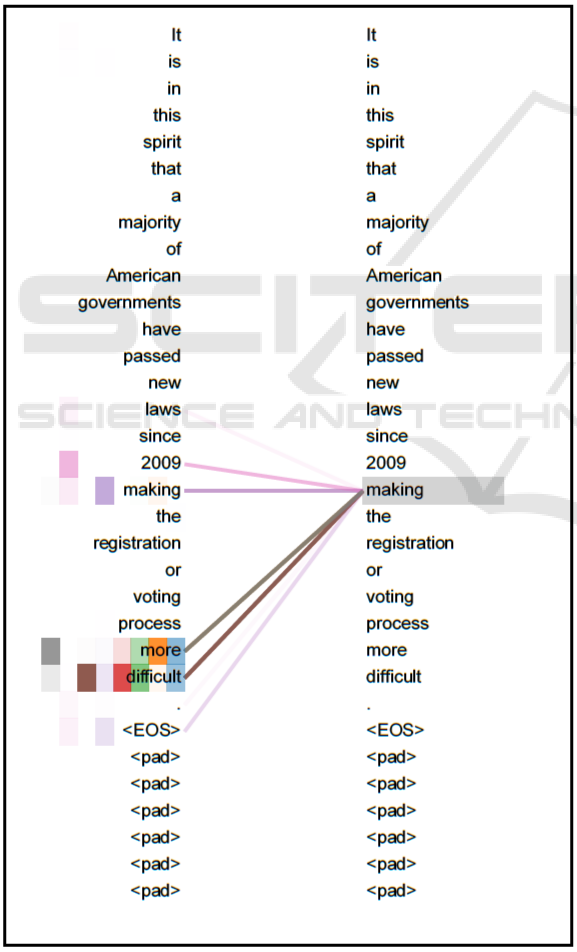

Figure 2: The attention mechanism's mapping. Source: (Vaswani et al., 2017).

The attention mechanism is a strong differentiation between Transformers and other NN architectures, allowing the estimation of the correlation between elements in a bidirectional way. Typically, there are two attention mechanism:

- Self-attention: intra-analysis of a sentence embeddings vectors, performing the similarity calculation between different words within the same sentence. In this analysis the mechanism extracts the correlation between words in the sentence. The sense of the vectors represents whether the words have similar or distinct semantic values.
- Multi-head-attention: divides the sentences into smaller parts to perform the similarity calculation between the matrices. It is similar to the self-attention mechanism, but between different portions of the sentences, identifying the relationship between words using text segments (sub-spaces).

Figure 2 presents the attention mechanism that estimates the correlation between words with similar semantic values, in a bidirectional way. From Figure 2 we can observe that the word "making" has a close relationship with the words "2009" and "laws" for instance, i.e., the word "making" appears in the same expressions than "2009" and "laws". This relationship allows the prediction of the next terms in sentences with words with similar meanings.

2.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language representation approach designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2018). In particular, a deep bidirectional TRANSFORMER is pre-trained in a masked language model and next sentence prediction objectives, enabling the representation to fuse the left and the right context, thus reducing the need for many heavily-engineered task-specific architectures. BERT is the first fine-tuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks, outperforming many task-specific architectures.

Figure 3 presents the NN layers of the BERT architecture. Particularly, we can observe the pre-training and the fine-tuning steps. During pre-training the input data set is used without labels, thus performing unsupervised training of the data. There are two main tasks during this stage:

- Token masking: randomly selecting a percentage of about 15 % of the tokens of input and applying a mask to them so that the training makes the prediction of these tokens.

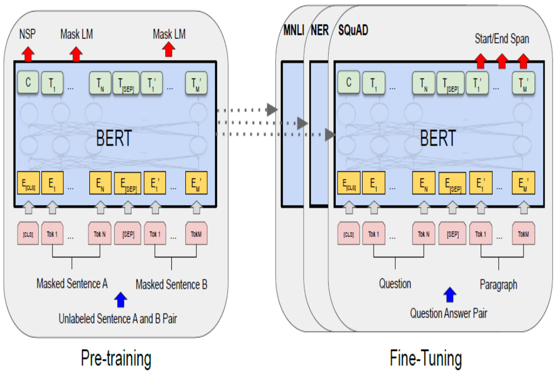

Figure 3: The BERT layers. Source: (Devlin et al., 2018).

- Next sentence prediction (NSP): training for question answering, predicting which sentences are subsequent to previous sentences.

After the pre-training step, the output data can be used as input to another NLP tasks. From Figure 3 we can observe that the pre-training output data is used for natural language inference (MNLI), named entity recognition (NER) and question answering (SQuAD).

2.2.2 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is a BERT-based framework for language model pre-training that extends BERT by training the model with bigger batches, over more data, and on longer sequences, also removing the next sentence prediction objective and dynamically changing the masking pattern applied to the training data (Liu et al., 2019). Experimental results on downstream tasks using the GLUE, RACE and SQuAD benchmarks show that RoBERTa achieves state-of-the-art results outperforming BERT and XLNet, an autoregressive learning approach (Yang et al., 2019). Table 1 presents the experimental parameters and results (performance measured by precision) comparing RoBERTa and BERT in three different tasks: question answering (SQuAD), natural language inference (MNLI), and sentence classification (SST).

Table 1: BERT/RoBERTa parameters and performance. Source: (Liu et al., 2019).

	BERT-Large	RoBERTa
Data	13GB	16GB
Batches	256	8K
Steps	1M	100K
SQuAD v1.1	90,9	93,6
SQuAD v2.0	81,8	87,3
MNLI-m	86,6	89.0
SST-2	93,7	95.3

From Table 1 we can observe that there are significant differences in training, with changes in the size of batches and in the number of steps in training. While RoBERTa uses a larger dataset than BERT to carry out its training, vigorously larger batches for its processing, however the processing occurs in a smaller number of steps. RoBERTa is a robust approach, however, as can be seen in theSQuAD, MNLI and SST tasks, RoBERTa presents similar and even better results than in the BERT approach.

2.2.3 DistilBERT

DistilBERT (Distilled BERT) is a general-purpose smaller and faster pre-trained version of BERT, that retains almost the same language understanding capabilities (Sanh et al., 2019). In particular it uses language models pre-trained with knowledge distillation, a compression technique in which a compact model is trained to reproduce the behaviour of a larger model or an ensemble of models, resulting in models that are lighter and faster at inference time, while also requiring smaller computational training. Particularly, it keeps 97% of language comprehension in its model with approximately 60% reduction in the model size, running 60% faster. DistilBERT can be fine-tuned on several downstream tasks, keeping the flexibility of larger models while it is small enough to run on the edge, e.g. on mobile devices.

The distillation technique (Hinton et al., 2015) consists of training a distilled (student) model to reproduce the behavior of a larger (teacher) model. Thus, DistilBERT is a leaner model based on the behavior of the original BERT model. Table 2 presents a comparison of precision performance in different NLP tasks among BERT, DistilBERT, and ELMo, a deep contextualized word representation approach that models complex syntactic and semantic characteristics of word uses and how these uses vary across different linguistic contexts (polysemy) (Peters et al., 2018).

Table 2: BERT, DistilBERT and ELMo performance. Source: (Sanh et al., 2019).

	Score	CoLA	MNLI	QNLI
ELMo	68.7	44.1	68.6	76.6
BERT-Base	79.5	56.3	86.7	88.6
DistilBERT	77.0	51.3	82.2	87.5

From Table 2 we observe that DistilBERT performance is close to BERT, even providing a reduced model.

2.2.4 ALBERT

ALBERT is another BERT based efficient architecture with significantly fewer parameters than a traditional BERT architecture (Lan et al., 2019). In particular, ALBERT incorporates parameter reduction techniques that lift the major obstacles in scaling pre-trained models, also acting as a form of regularization that stabilizes the training and helps with generalization. First, it incorporates factorized embedding parametrization, i.e., decompose the large vocabulary embedding matrix into two small matrices, thus separating the size of the hidden layers from the size of vocabulary embedding, making it easier to grow the hidden size without significantly increasing the parameter size of the vocabulary embeddings. Second it incorporates cross-layer parameter sharing, preventing the parameter from growing with the depth of the network. Additionally, ALBERT replaces the next sentence prediction proposed in the original BERT by a self-supervised loss for sentence-order prediction. Experiments with GLUE, RACE and SQuAD benchmarks show that ALBERT achieves state-of-the-art performance on natural language understanding tasks outperforming BERT, XLNET and ROBERTA.

In particular, ALBERT address the scalability problem of BERT derived from memory consumption issues. The growth in the number of parameters of BERT has become an important challenge due to the high memory consumption. Few works reported in literature address this problem, by using parallelism (Shazeer et al., 2018) or effectively managing memory consumption through a cleaning mechanism to minimize performance impact (Gomez et al., 2017). However, the obstacle created by the communication overhead of the BERT architecture is not addressed by these reported works.

Thus, BERT was extended by ALBERT in order to reduce around 89% of the number of parameters, improving performance in NLP tasks. Table 3 shows a comparison between the hyperparameters of BERT and ALBERT. Even using less hyperparameters than BERT, ALBERT provide improved results in different NLP tasks, such as SQuAD v1.1 (+1.9%), SQuAD v2.0 (+3.1%), MNLI (+1.4%), SST-2 (+2.2%), and RACE (+8.4%) using relatively less resources and with a faster training phase (Lan et al., 2019).

3 RELATED WORK

The emergence of approaches that use Transformers to improve performance in NLP tasks has grown in recent years. Particularly for NER in complex languages, a recent work reported in literature (Arkhipov et al., 2019) uses Transformers for named entity recognition in Slavic languages, achieving up to 93% of performance in F1 measure when applied to the Czech language.

Recently, different NN architectures were proposed to perform NER in Portuguese (Souza et al., 2019). In addition to the comparative analysis between the architectures, the authors proposed an effective approach for both word embbedings generation and named entity recognition in Portuguese. The proposed approach uses BERT to first generate the word embeddings for Portuguese and finally use this word embeddings for NER. The authors also evaluate different NN architectures, such as LSTM (Long-Short Term Memory) and BiLSTM (Bidirectional LSTM) for named entity recognition in Portuguese. They also combine these different architectures with CRF (Conditional Random Fields) (Lafferty et al., 2001) to improve performance. Table 4 summarizes the experimental results of the proposed architectures for multilingual (ML) and Portuguese (PT) in two scenarios: a full scenario using all the HAREM dataset with 10 classes, and a selective scenario using a subset of 5 classes of HAREM where the proposed approach performs better.

From Table 4 we observe that the BERT-LARGE approach outperforms BERT-BASE. Additionally, the LSTM architecture does not provide any gain, however combining CRF brings outstanding performance. Moreover, single-lingual models outperforms multi-lingual models for NER in Portuguese. Thus, the best results were obtained with a single-lingual model trained specifically for Portuguese. Although the single-lingual approach performs better, the computational cost of training the model in Portuguese is much higher than using a pre-trained multilingual model.

Although the authors provide a single-lingual SOTA approach for NER in Portuguese, a question remains: is it possible that multilingual NER models can outperform single-lingual models, particularly for complex languages, such as Portuguese?

4 PROPOSED APPROACH

In this section we present our multilingual transformer-based text embeddings approach for NER in Portuguese. First, we present a comprehensive dataset we propose to improve the training of NER models for Portuguese. Second, we present the architecture of our proposed approach.

Table 3: BERT and ALBERT hyperparameters. Source: (Lan et al., 2019).

Model	Parameters		Layer		Embedding
	#	Sharing	#	Hidden	Size
BERT-Base	108M	No	12	768	768
BERT-Large	334M	No	24	1024	1024
ALBERT-Base	12M	Yes	12	128	768
ALBERT-Large	18M	Yes	24	128	1024
ALBERT-XLarge	60M	Yes	24	128	2048
ALBERT-XXLarge	235M	Yes	12	128	4096

Table 4: Performance in precision, recall and F1 of the SOTA single-lingual approach for NER trained specifically for Portuguese in two experimental scenarios. Source: (Souza et al., 2019).

Approach	Full Scenario			Selective Scenario		
	Precision	Recall	F1	Precision	Recall	F1
CharWNN	67.16	63.74	65.41	73.98	68.68	71.23
LSTM-CRF	72.78	68.03	70.33	78.26	74.39	76.27
BiLSTM-CRF+FlairBBP	74.91	74.37	74.64	83.38	81.17	82.26
ML-BERT-Base	2.97	73.78	73.37	77.35	79.16	78.25
ML-BERT-Base-CRF	74.82	73.49	74.15	80.10	78.78	79.44
ML-BERT-Base-LSTM	69.68	69.51	69.59	75.59	77.13	76.35
ML-BERT-Base-LSTM-CRF	74.70	69.74	72.14	80.66	75.06	77.76
PT-BERT-Base	78.36	**77.62**	77.98	83.22	82.85	83.03
PT-BERT-Base-CRF	78.60	76.89	77.73	83.89	81.50	82.68
PT-BERT-Base-LSTM	75.00	73.61	74.30	79.88	80.29	80.09
PT-BERT-Base-LSTM-CRF	78.33	73.23	75.69	84.58	78.72	81.66
PT-BERT-Large	78.45	77.40	77.92	83.45	**83.15**	**83.30**
PT-BERT-Large-CRF	**80.08**	77.31	**78.67**	**84.82**	81.72	83.24
PT-BERT-Large-LSTM	72.96	72.05	72.50	78.13	78.93	78.53
PT-BERT-Large-LSTM-CRF	77.45	72.43	74.86	83.08	77.83	80.37

4.1 Training Dataset

To improve the training of multilingual NER models for Portuguese, we build a comprehensive dataset from HAREM (Santos and Cardoso, 2007). HAREM[1] is a manually annotated dataset used to assess the performance of information systems for named entity recognition in Portuguese. HAREM is widely used by several NLP approaches reported in literature (Souza et al., 2019; de Castro et al., 2018; Gonçalo Oliveira and Cardoso, 2009; Fernandes et al., 2018; Consoli and Vieira, 2019; Pires, 2017). In particular, the HAREM dataset has the following divisions:

- "CD Primeiro HAREM": 129 documents and 80,060 words.
- "CD Segundo HAREM": 129 documents and 147,991 words.
- "Mini-HAREM CD": 128 documents and 54,074 words.

[1]Available at http://www.linguateca.pt

All HAREM divisions were joined into a single unified training dataset. Originally, some expressions in HAREM are ambiguous, i.e., some of them have two entity labels with different meanings. To build the unified training dataset we choose the first classification described in the HAREM dataset, discarding the second one. Thus, all expressions were classified in a single entity label. Additionally, the paragraph structure was converted into smaller sentences so that the BERT-based algorithm can receive input data in an appropriate format. Paragraphs of up to 256 tokens were automatically converted to sentences and the paragraphs were divided with entity labels also been incorporated into the unified training dataset.

4.2 Architecture

The proposed multilingual approach for NER in Portuguese can use multiple transformer-based text embeddings. In particular, we implement and evaluate BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019). Figure 4 presents the architecture of our proposed ap-

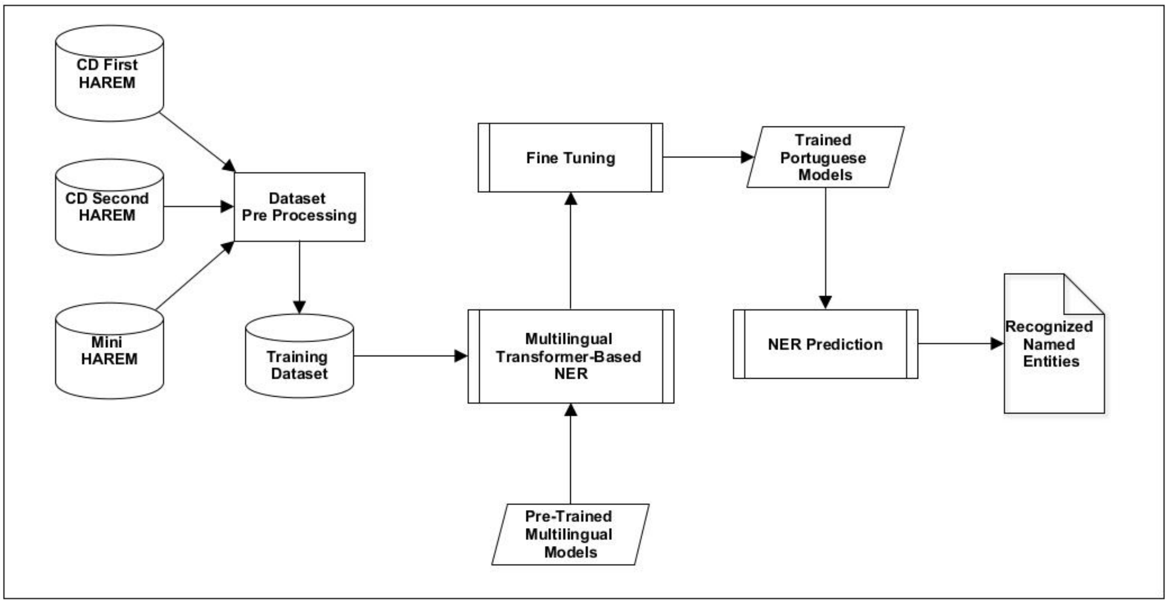

Figure 4: The architecture of the proposed multilingual transformer-based text embeddings approach for NER in Portuguese.

proach. Particularly, there are four processing steps: i) Dataset preprocessing; ii) Multilingual transformer-based NER; iii) Fine-tuning; iv) NER prediction.

In the dataset preprocessing step our approach builds the training dataset as described in Section 4.1, removing the original ambiguities in HAREM, standardizing the data in sentences within the BERT standard and consolidation in a single data file. In the second step our approach selects the multilingual transformers-based model for NER, instantiating them in the processing engine and loading the pre-trained multilingual models for the generation of the text embeddings. In the fine tuning step our approach sets the model hyperparameters for the NER task, generating the final NER model by training the model using Portuguese training data. Finally, in the prediction step our approach loads the trained model, receives all the sentences to be evaluated and generates a final output with the named entities recognized from the input sentences.

The pipeline works in a flexible way so that if a new version of the HAREM dataset is published it is possible to incorporate it in the training dataset, preserving the original content and expanding the volume of data available for training and testing models. Similarly, although three Transformers approaches have been initially used in our experiments, it is also possible to plug in new Transformers-based approaches with no impact to the processing workflow.

5 EXPERIMENTS

In this section we present the experiments we carried out to evaluate our proposed approach, including experimental setup, procedures and results. In particular, the experimental evaluation answer the following research questions:

1. How effective is each one of the multilingual BERT-based algorithm for NER in Portuguese?
2. How does our multilingual approach performs compared to the SOTA single-lingual approach for NER in Portuguese?

In our evaluation we consider for distinct training scenarios: i) 70% of data for training and 30% of data for testing; ii) 80% of data for training and 20% of data for testing; iii) 90% of data for training and 10% of data for testing; iv) 95% of data for training and 5% of data for testing. In each of the scenarios, we evaluate BERT (Devlin et al., 2018), XLM-RoBERTa (Lample and Conneau, 2019) and DistilBERT (Sanh et al., 2019), also performing fine-tuning for NER task. For fair comparison, the same training dataset and setup parameters were used for each BERT-based algorithm.

The large number of batch sizes implies in reducing the number of examples sent for the input of the BERT-based algorithm, consequently negatively impacting in performance. Thus, batches of 128 and 256 have become more suitable for our experiments. Batches smaller than 128 could cause truncation is-

Table 5: Performance of our multilingual approach using different BERT-based algorithms with multiple training set variations and 3 epochs of training.

Approach	Train (%)	SEQ_SIZE	Precision (%)	Recall (%)	F1 (%)
BERT-BASE	95	128	85.00	86.80	85.90
BERT-BASE	95	256	85.70	86.30	86.00
DISTILBERT	95	128	77.10	82.90	79.90
DISTILBERT	95	256	78.50	83.00	80.70
XML-ROBERTA	95	128	**88.00**	87.60	**87.80**
XML-ROBERTA	95	256	86.30	**88.40**	87.30
BERT-BASE	90	128	67.00	74.20	70.40
BERT-BASE	90	256	68.60	75.40	71.80
DISTILBERT	90	128	62.30	68.20	65.10
DISTILBERT	90	256	62.60	69.30	65.80
XML-ROBERTA	90	128	73.00	78.60	75.70
XML-ROBERTA	90	256	**74.60**	**79.80**	**77.10**
BERT-BASE	80	128	66.40	69.90	68.10
BERT-BASE	80	256	68.30	71.20	69.70
DISTILBERT	80	128	59.10	64.60	61.70
DISTILBERT	80	256	60.80	64.70	62.70
XML-ROBERTA	80	128	**67.90**	70.90	69.40
XML-ROBERTA	80	256	67.90	**71.50**	**69.70**
BERT-BASE	70	128	61.40	62.30	61.80
BERT-BASE	70	256	62.50	64.40	63.40
DISTILBERT	70	128	58.00	59.50	58.80
DISTILBERT	70	256	59.30	61.10	60.20
XML-ROBERTA	70	128	**64.40**	**64.80**	**64.60**
XML-ROBERTA	70	256	64.10	64.80	64.40

sues, that is, the sentences would be truncated, generating more data loss.

Transformer-based approaches, particularly BERT, usually require few interactions to converge in a model able to provide efficient results (Wolf et al., 2019). We test different epochs to finally set this parameter to 3, for better balancing between training time and model performance. Table 5 presents the performance of our multilingual approach using different BERT-based algorithms with multiple training set variations and 3 epochs of training.

From Table 5 we observe that XML-ROBERTA outperforms BERT and DISTILBERT in different scenarios. Particularly, the volume of training data impacts the performance of all BERT-based algorithms, with XML-ROBERTA outperforming BERT-BASE in 2.68% in precision, 1.84% in recall and 2.09% in F1, also outperforming DISTILBERT in 12.10% in precision, 6.50% in recall and 8.79% in F1, considering the 95% of training scenario. Additionally, we can observe that the differences in XML-ROBERTA performance with batches of 128 and 256 are negligible (0.57% in F1). Recalling our first research question, these experimental results attest the effectiveness of our multilingual ROBERTA approach for NER in Portuguese.

Table 6 presents the performance of the SOTA single-lingual transformer-based text embeddings approach reported in literature (Souza et al., 2019) in comparison to our proposed multilingual transformer-based text embeddings approach for NER in Portuguese. From Table 6 we observe that our multilingual approach (XML-ROBERTA) outperforms the best single-lingual approach (PT-BERT-LARGE-CRF) in the full scenario with gains of 9.89% in precision, 13.31% in recall, and in 11.60% in F1. Even considering the selective (best) scenario for the single-lingual approach, the gains are still significant of 3.74% in precision, 7.19% in recall, and 5.47% in F1. Recalling our second research question, these experimental results show that multilingual trasformer-based text embeddings approaches fine tuned with a large dataset outperforms SOTA trasformer-based models trained specifically for Portuguese.

Multilingual transformer-based approaches for NER becomes particularly interesting in scenarios where the amount of computational resources is limited to train single-lingual approaches but the amount of training data is abundant for fine tuning. In addition, the fine tuning step can be generalized for any multilingual approach based on BERT. Therefore, ALBERT (Lan et al., 2019) and BART (Lewis et al., 2019) for instance, can be easily implemented in our proposed transformed-based approach, similarly we implemented DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019).

Table 6: Performance of the SOTA single-lingual and the proposed multilingual transformed-based text embeddings approaches for NER in Portuguese.

Approach	**Prec.**	**Rec.**	**F1**
Single-lingual (Full Scenario)			
LSTM-CRF	72.78	68.03	70.33
BiLSTM-CRF+FlairBBP	74.91	74.37	74.64
ML-BERT-Base-CRF	74.82	73.49	74.15
PT-BERT-Base-CRF	78.60	76.89	77.73
PT-BERT-Large-CRF	80.08	77.31	78.67
Single-lingual (Selective Scenario)			
LSTM-CRF	78.26	74.39	76.27
BiLSTM-CRF+FlairBBP	83.38	81.17	82.26
ML-BERT-Base-CRF	80.10	78.78	79.44
PT-BERT-Base-CRF	83.89	81.50	82.68
PT-BERT-Large-CRF	84.82	81.72	83.24
Multilingual (Full Scenario)			
DistilBERT	78.50	83.00	80.70
BERT-Base	85.70	86.30	85.90
XLM-RoBERTa	**88.00**	**87.60**	**87.80**

6 CONCLUSIONS

In this article, we assessed the effectiveness of multilingual transformer-based text embeddings for named entity recognition in Portuguese. Particularly, we fine-tuned our approach using a large Portuguese dataset, and we carried out experiments comparing our approach with the state of the art single-lingual approach trained specifically for Portuguese.

Experimental results showed that our multilingual trasformer-based approach outperformed the state of the art approach, achieving 88.0% of precision and 87.8% in F1 in named entity recognition for Portuguese, with gains of up to 9.89% of precision and 11.60% in F1. Additionally, even considering a selective scenario, where the state of the art approach performed better, our approach outperformed it by 3.74% of precision and 5.47% in F1. Thus, our experiments showed that pre-trained multilingual generic language models based on BERT and fine-tuned with a larger dataset can outperforms single-lingual specific language models that requires a lot of time and computational resources to be trained.

In future work, we intent to evaluate the impact of the size of the unified dataset over the effectiveness of the NER model, as well to improve the transformer-based algorithms so that it is possible to adjust the batches to smaller sizes, such as 64, allowing to increase the number of sentences analyzed and possibly get outstanding results. In addition, similarly to the state of the art single-lingual approach, we intent to add a CRF layer to our multilingual approach, which can further improve the precision.

ACKNOWLEDGEMENTS

The present work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing Code 001. The authors thank the partial support of the CNPq (Brazilian National Council for Scientific and Technological Development), FAPEMIG (Foundation for Research and Scientific and Technological Development of Minas Gerais), CEMIG, FUMEC, LIAISE and PUC Minas.

REFERENCES

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

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.

Borthwick, A. E. (1999). *A maximum entropy approach to named entity recognition*. PhD thesis, New York University, USA.

Consoli, B. and Vieira, R. (2019). Multidomain contextual embeddings for named entity recognition. *Proceedings of the Iberian Languages Evaluation Forum*, 2421:434–441.

de Castro, P. V. Q., da Silva, N. F. F., and da Silva Soares, A. (2018). Portuguese named entity recognition using LSTM-CRF. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 83–92.

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

Fernandes, I., Cardoso, H. L., and Oliveira, E. (2018). Applying deep neural networks to named entity recognition in portuguese texts. In *Procedings of the 5th In-*

ternational Conference on Social Networks Analysis, Management and Security, pages 284–289.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57(1):345–420.

Gomez, A. N., Ren, M., Urtasun, R., and Grosse, R. B. (2017). The reversible residual network: Backpropagation without storing activations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 2211–2221.

Gonçalo Oliveira, H. and Cardoso, N. (2009). Sahara: An online service for harem named entity recognition evaluation. In *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology*, pages 171–174.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13(1):307–361.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-Thought vectors. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, pages 1532–1543.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, page 282–289.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

Marsh, E. and Perzanowski, D. (1998). MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the 7th Message Understanding Conference*.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3111–3119.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 246–252.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 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*, pages 2227–2237.

Pires, A. R. O. (2017). Named entity extraction from portuguese web text. Master's thesis, Porto University.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Santos, D. and Cardoso, N. (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Shazeer, N., Cheng, Y., Parmar, N., Tran, D., Vaswani, A., Koanantakool, P., Hawkins, P., Lee, H., Hong, M., Young, C., et al. (2018). Mesh-TensorFlow: Deep learning for supercomputers. In *Procedings of the Advances in Neural Information Processing Systems*, pages 10414–10423.

Souza, F., Nogueira, R. F., and de Alencar Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. (2017). A correlated topic model using word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4207–4213.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *CoRR*, 1906.08237.