# Portuguese Language Models and Word Embeddings: Evaluating on Semantic Similarity Tasks

Ruan Chaves Rodrigues[1(✉)], Jéssica Rodrigues[2], Pedro Vitor Quinta de Castro[1], Nádia Felix Felipe da Silva[1], and Anderson Soares[1]

[1] Institute of Informatics, Federal University of Goiás, Goiânia, Brazil
ruanchaves93@gmail.com, {pedrovitorquinta,nadia,anderson}@inf.ufg.br
[2] Department of Computer Science, Federal University of São Carlos, São Carlos, Brazil
jsc.rodrigues@gmail.com

**Abstract.** Deep neural language models which achieved state-of-the-art results on downstream natural language processing tasks have recently been trained for the Portuguese language. However, studies that systematically evaluate such models are still necessary for several applications. In this paper, we propose to evaluate the performance of deep neural language models on the semantic similarity tasks provided by the ASSIN dataset against classical word embeddings, both for Brazilian Portuguese and for European Portuguese. Our experiments indicate that the ELMo language model was able to achieve better accuracy than any other pretrained model which has been made publicly available for the Portuguese language, and that performing vocabulary reduction on the dataset before training not only improved the standalone performance of ELMo, but also improved its performance while combined with classical word embeddings. We also demonstrate that FastText skip-gram embeddings can have a significantly better performance on semantic similarity tasks than it was indicated by previous studies in this field.

**Keywords:** Deep neural language models · Semantic textual similarity · Portuguese language

## 1 Introduction

The application of deep learning methods to Natural Language Processing (NLP) is possible due to the representation of words as vectors in a low-dimensional continuous space. These traditional word embeddings are static: each word has a single vector, regardless of its context [20,21]. This generates several problems, especially that all the senses of a polysemic word have to share the same

The source code for the experiments described in this paper has been published on GitHub at https://github.com/ruanchaves/elmo.


P. Quaresma et al. (Eds.): PROPOR 2020, LNAI 12037, pp. 239–248, 2020.
https://doi.org/10.1007/978-3-030-41505-1_23

representation. Recent developments in the field produced deep neural language models such as ELMo [23] and BERT [10], which have successfully created contextualized word representations, word vectors that are sensitive to the context in which they appear. Using contextualized representations rather than static embeddings has resulted in significant improvements in a variety of NLP tasks, such as question answering and coreference resolution.

In this paper, we present experiments carried out to evaluate different word representation models for Portuguese, including both Brazilian and European variants, for semantic similarity tasks. To our knowledge, this is the first paper to evaluate deep neural language models on semantic similarity tasks in the Portuguese language.

Our experiments indicate that, if fine-tuning is not applied to any language model, then the ELMo language model is able to achieve better accuracy than any other pretrained model which has been made publicly available for the Portuguese language. We have found that performing vocabulary reduction on the corpus before training not only improved the standalone performance of ELMo, but also improved its performance while combined with classical word embeddings. We also demonstrate that FastText skip-gram embeddings [2] can have a significantly better performance on semantic similarity tasks than it was indicated by previous studies in this field.

In Sect. 2 we describe some of the approaches for generating deep neural language models proposed in the literature. The approaches investigated in this paper are described in Sect. 3. The experiments carried out for evaluating deep neural language models for Portuguese are described in Sect. 4. Section 5 finishes this paper with its conclusions and proposals for future work.

## 2 Related Work

Hartmann et al. [12] trained 31 word embedding models using FastText, GloVe, Wang2Vec and Word2Vec. The authors evaluated them intrinsically on syntactic and semantic analogies and extrinsically on POS tagging and sentence semantic similarity tasks. The authors contribute with a variety of pre-trained word embeddings, intrinsic and extrinsic task comparisons, and preprocessing and evaluation codes. We used this work as a baseline for deep neural language models.

Quinta de Castro et al. [6] evaluated the four different types of word embeddings pre-trained by [12] and performed an extrinsic evaluation of them in the Named Entity Recognition (NER) task. The authors used only 100-dimensional word embeddings, applying them to the same BiLSTM-CRF deep learning architecture from [15], and improved the previous state-of-the-art on the HAREM [26] benchmark for Portuguese language using Wang2Vec [16] embeddings.

An ELMo [23] model trained for Portuguese has been previously evaluated by Quinta de Castro [7] on NER tasks for the IberLEF evaluation [11].

Quinta de Castro [7] also made their model publicly available through the AllenNLP library[1]. The authors experimented different scenarios of NER with Portuguese corpora, using a BiLSTM-CRF network from the AllenNLP library. The results achieved state-of-the-art performance using the optimal values for them.

Santos et al. [27] assessed how different combinations of static word embeddings and contextualized embeddings impact NER for the Portuguese language. The authors show a comparative study of 16 different combinations of static and contextualized embeddings and evaluate NER performance using the HAREM benchmark. The best NER system outperforms the state-of-the-art in Portuguese NER by 5.99 in absolute percentage points.

Quinta de Castro [5] evaluated different combinations of word representations, such as character level embeddings, static word embeddings from [12] and ELMo embeddings [23] on the NER task. The author performed a comparative study on two different domains for the Portuguese language (general and legal), performing the pre-training of the ELMo embeddings for each domain, and comparing them to a fine-tuned version of the model on different NER corpora, for each domain. The author reached a new state-of-the-art for the HAREM benchmark using the fine-tuned ELMo embeddings, combined with 100-dimensional Wang2Vec embeddings.

To our knowledge, this is the first paper to evaluate deep neural language models on semantic similarity tasks in the Portuguese language. The semantic similarity task provided by the ASSIN dataset is equivalent to the Semantic Textual Similarity Benchmark (STS-B), and works that evaluated deep neural language models on the STS-B task, such as [22], can be taken as a reference for what to expect of its performance in other linguistic contexts.

## 3 Word Representations

In this paper, two ways of word representation were evaluated in semantic similarity tasks for Portuguese: contextualized and static word representations. They were tested both individually and also pairwise concatenated with each other, and each approach is explained in the next sections.

### 3.1 Static Word Representations

Word representations are numerical vectors which can represent words or concepts in a low-dimensional continuous space, reducing the inherent sparsity of traditional vector-space representations [25]. These vectors, also known as embeddings, are able to capture useful syntactic and semantic information, such as regularities in natural language. They are based on the distributional hypothesis, which establishes that the meaning of a word is given by its context of occurrence [3]. A numerical vector representing a word can be visualized in a continuous vector space, accepting algebraic operations such as the cosine distance.

[1] https://allennlp.org/elmo.

The ability of static word embeddings to capture knowledge has been exploited in several tasks, such as Machine Translation [20], Word Sense Disambiguation [9] and Language Understanding [18].

Although very useful in many applications, the static word embeddings, like those generated by Word2Vec [19], GloVe [21], Wang2Vec [16] and FastText [2] have an important limitation: each word is associated with only one vector representation, ignoring the fact that polysemous words can assume multiple meanings. This limitation is called *Meaning Conflation Deficiency*, which is a mixture of possible meanings in a single word [4]. For instance, in the phrase "*My mouse was broken, so I bought a new one yesterday.*" the word "*mouse*" should be associated with its meaning of being a *computer device*, rather than the *animal called mouse*. Figure 1 is an illustration of this Meaning Conflation Deficiency in a 2D semantic space.

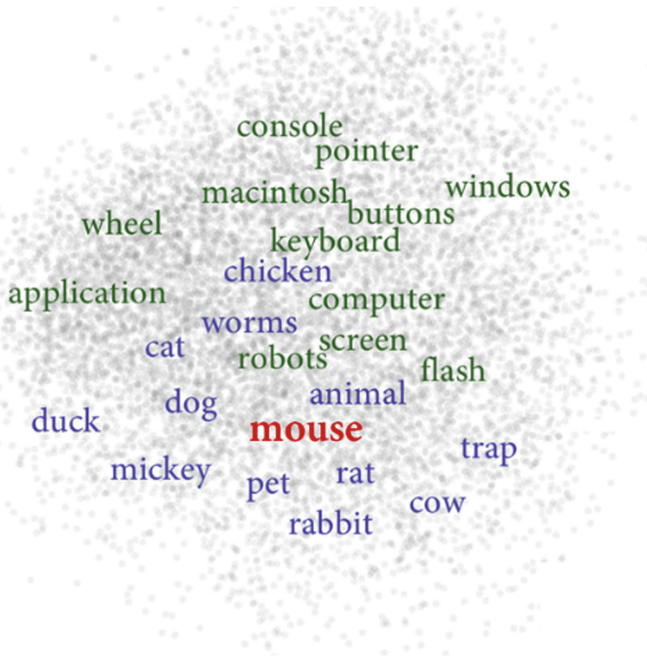


**Fig. 1.** Example of meaning conflation deficiency of ambiguous word "mouse". The words in blue refer to the sense of animal and the words in green to the sense of device. (Color figure online)

Because they create a single representation for each word, a notable problem with static word embeddings is that all senses of a polysemous word must share a single vector.

### 3.2 Contextualized Word Representations

The limitations of static word embeddings have led to the creation of context-sensitive word representations. ELMo [23], BERT [10], and GPT-2 [24] are examples of deep neural language models that are fine-tuned to create models for a wide variety of downstream NLP tasks. As GPT-2 is not yet available for the Portuguese language, we performed our experiments solely on ELMo and a multilingual version of BERT. The internal representations of words for these language models are called contextualized word representations because they are a function of the entire input sentence, and in this study, sentence embeddings were built through the summation of these representations. The success of this approach suggests that these representations capture highly transferable and task-agnostic properties of natural languages [17].

**ELMo.** [23] is a two-layer bidirectional LSTM language model, built over a context independent character CNN layer and originally trained on the Billion Word Benchmark dataset [8], consisting primarily of newswire text. In order to obtain a representation for each word, we performed a linear concatenation of all three ELMo layers, without learning any task-specific weights. During our experiments, we considered two ELMo language models that were exclusively trained for the Portuguese language. The first model has been made publicly available through the AllenNLP library. The second model was trained by ourselves in an attempt to improve on the accuracy of this public model: although it took the same dataset used by the first model as its starting point, words that occurred less than three times were removed from the dataset before training the model. Such additional vocabulary reduction step was accompanied by suitable adjustments on the softmax layer and the network architecture.

**BERT.** [10] is a deep Transformer [28] encoder trained jointly as a masked language model and on next-sentence prediction, originally trained on the concatenation of the Toronto Books Corpus [29] and the English Wikipedia. As with GPT, we do not fine-tune the encoder weights. We utilized the publicly released BERT-multilingual model, which was simultaneously trained on the Wikipedia dumps for 104 different languages. In order to achieve better accuracy on the semantic similarity task, we considered only the final layer of the model for generating its sentence embeddings.

## 4 Experiments and Results

In this section we show the experiments carried out to evaluate the two approaches under investigation: word embeddings (Word2Vec, FastText) and deep neural language models (ELMo, BERT) on semantic similarity tasks.

### 4.1 Evaluation

Based on [12], this experiment is a task of semantic similarity between sentences where the use of neural language models is evaluated. Word embeddings were chosen as baselines.

***Dataset.*** ASSIN (Avaliação de Similaridade Semântica e Inferência Textual) was a workshop co-located with PROPOR-2016 which encompassed two shared-tasks regarding: (i) semantic similarity and (ii) entailment. We chose the first one to evaluate our contextualized vectors extrinsically in a semantic task. In ASSIN, the participants of the semantic similarity shared-task were asked to assign similarity values between 1 and 5 to pairs of sentences (gold score). The workshop made available the training and test sets for Brazilian (PT-BR) and European (PT-EU) Portuguese.

***Algorithm.*** The objective of this task is to predict, through a linear regression, the similarity score between two sentences. The model is trained in the training set, which contains sentence pairs with the gold score. The prediction occurs in the test set, which contains sentence pairs without the gold score. As we have this same test set with the gold score, it is possible to calculate Pearson's Correlation ($\rho$) and Mean Squared Error (MSE) between them. These results show how much the automatic prediction has approached the human prediction.

The results were obtained after training a linear regressor with the cosine similarity between the summations of the word representations of each sentences' words, in a procedure almost equivalent to what has been performed by [12]. However, we applied the following changes to his original approach: for word embeddings, we avoided most occurrences of out of vocabulary words by applying to the test set the same tokenization and normalization steps which were performed on the training set before the word embeddings were trained. These steps were described by [12] and we performed them through their standard implementation.

This approach significantly reduced the amount of out-of-vocabulary words for word embeddings, and the remaining ones were simply ignored, instead of being replaced by a single UNKNOWN token. In the case of language models, such as ELMo and BERT, no preprocessing was applied, and whenever evaluating the combination of a language model and a word embedding, we simply performed the concatenation of the sentence embeddings produced from each source.

***Evaluation Metrics.*** The Pearson correlation coefficient measures the linear relationship between two datasets: one annotated by the participants and another which is output by the system. Like other correlation coefficients, this one varies between $-1$ and $+1$ with 0 meaning no correlation. Correlations of $-1$ or $+1$ mean an exact linear relationship. The Mean Squared Error (MSE) of an estimator measures the average of the squares of the errors, that is, the average squared difference between the estimated value and what was expected.

***Discussion of Results.*** Table 1 shows the performance of our models for the Brazilian Portuguese and European Portuguese test sets, through the Pearson's Correlation ($\rho$) and mean squared error (MSE).

All semantic similarity tests for word embeddings listed by [12] were repeated during our experiments. Although most word embeddings retained exactly the same relative accuracy, FastText skip-gram embeddings have exhibited a noticeable increase in performance; in fact, the FastText skip-gram embedding at 1000 dimensions achieved the best standalone accuracy among all word embeddings considered for the ASSIN semantic similarity task. This happened because the approach deployed by [12] produced a higher amount of out-of-vocabulary (OOV) words, and the FastText embeddings were abnormally sensitive to their adopted strategy of replacing all OOV words by a single UNKNOWN token. [14] provides a survey of OOV word replacement techniques that can avoid this handicap. We therefore conclude that the performance oscillations in FastText word embeddings reported by [12] should first be regarded as a result of their approach to word preprocessing and OOV word replacement after the training stage, rather than as a by-product of intrinsic properties of the word embeddings themselves.

Furthermore, three language models were evaluated, both in isolation and in combination with each one of the word embeddings made publicly available by [12]: BERT, and ELMo with and without vocabulary reduction. While the concatenation of ELMo without vocabulary reduction with any word embeddings resulted in a worse result than using ELMo by itself, the reduced version of ELMo significantly improved its accuracy after being concatenated with Word2Vec embeddings. Such an improvement has not been achieved by any other word embedding. The best combination of Word2Vec and reduced ELMo is reported in Table 1: results belong to the concatenation of the reduced version of ELMo with a Word2Vec embedding that has 1000 dimensions and follows the Continuous Bag of Words (CBOW) model.

It is also important to notice that, while ELMo retained a relatively stable performance across the Brazilian and European versions of the dataset, BERT-

**Table 1.** Best results for extrinsic evaluation on the semantic similarity task. Arrows indicate whether lower (↓) or higher (↑) is better. A hyphen (-) indicates the absence of either a word embedding or a language model. All word embeddings present in the table below have 1000 dimensions.

| Word embedding | Language model | PT-BR | | PT-EU | |
|---|---|---|---|---|---|
| | | $\rho$ (↑) | MSE (↓) | $\rho$ (↑) | MSE (↓) |
| Word2Vec (CBOW) | ELMo, reduced | **0.632** | **0.457** | **0.654** | **0.709** |
| - | ELMo, reduced | 0.618 | 0.472 | 0.627 | 0.735 |
| - | ELMo | 0.611 | 0.479 | 0.620 | 0.745 |
| - | BERT-multilingual | 0.604 | 0.483 | 0.564 | 0.813 |
| FastText (skip-gram) | - | 0.590 | 0.496 | 0.571 | 0.796 |

multilingual loses a measurable portion of its accuracy while performing semantic similarity tasks in European Portuguese. In all likelihood, such a steep decline happens due to the imbalanced proportion between Brazilian and European Portuguese articles in Wikipedia, on which BERT-multilingual was trained.

## 5 Conclusion and Future Work

Our experiments have shown that the ELMo model that has been made publicly available through the AllenNLP library is already able to consistently perform better on semantic similarity tasks than multilingual versions of BERT that have not been subject to fine-tuning, or classical word embeddings, even when both the Brazilian and European dialects of the Portuguese language are taken into account. Although results for word embeddings superior to those achieved by ELMo have already been reported in the literature, they combine multiple word embeddings [13] or combine a word embedding with several linguistic features [1].

Furthermore, we have also seen that vocabulary reduction not only improved its standalone performance, but made it suitable to be concatenated with Word2Vec embeddings on semantic similarity tasks, which seems to suggest that vocabulary reduction made ELMo favorable to ensemble approaches for improving on its accuracy.

In the future, we should also evaluate if similar results would happen on other downstream tasks, such as sentiment analysis, part-of-speech tagging and named entity recognition. And given the current lack of pretrained deep language models in the Portuguese language, we may also consider introducing in our next experiments not only existing multilingual models, but also more deep language models trained by ourselves, optimized to work exclusively with the Portuguese language.

## References

1. Alves, A., Gonçalo Oliveira, H., Rodrigues, R., Encarnaçao, R.: ASAPP 2.0: advancing the state-of-the-art of semantic textual similarity for Portuguese. In: 7th Symposium on Languages, Applications and Technologies (SLATE 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2018)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
3. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res. **49**, 1–47 (2014)
4. Camacho-Collados, J., Pilehvar, M.T.: From word to sense embeddings: a survey on vector representations of meaning. J. Artif. Intell. Res. **63**, 743–788 (2018)
5. de Castro, P.V.Q.: Aprendizagem Profunda para Reconhecimento de Entidades Nomeadas em Domínio Jurídico. Master's thesis, Universidade Federal de Goiás (2019)

6. Quinta de Castro, P.V., Félix Felipe da Silva, N., da Silva Soares, A.: Portuguese named entity recognition using LSTM-CRF. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 83–92. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99722-3_9
7. de Castro, P.V.Q., da Silva, N.F.F., da Silva Soares, A.: Contextual representations and semi-supervised named entity recognition for Portuguese language (2019)
8. Chelba, C., et al.: One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005 (2013)
9. Chen, X., Liu, Z., Sun, M.: A unified model for word sense representation and disambiguation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1025–1035. Association for Computational Linguistics, October 2014. https://doi.org/10.3115/v1/D14-1110, https://www.aclweb.org/anthology/D14-1110
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2018)
11. Glauber, R.: IberLEF 2019 Portuguese named entity recognition and relation extraction tasks (2019)
12. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025 (2017)
13. Hartmann, N.S.: Solo queue at assin: Combinando abordagens tradicionais e emergentes. Linguamática **8**(2), 59–64 (2016)
14. Hu, Z., Chen, T., Chang, K.W., Sun, Y.: Few-shot representation learning for out-of-vocabulary words (2019)
15. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
16. Ling, W., Dyer, C., Black, A.W., Trancoso, I.: Two/too simple adaptations of Word2Vec for syntax problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1299–1304 (2015)
17. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
18. Mesnil, G., He, X., Deng, L., Bengio, Y.: Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Interspeech, pp. 3771–3775 (2013)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
20. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013)
21. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
22. Peters, M., Ruder, S., Smith, N.A.: To tune or not to tune? adapting pretrained representations to diverse tasks. arXiv preprint arXiv:1903.05987 (2019)
23. Peters, M.E., et al.: Deep contextualized word representations (2018)
24. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018). https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
25. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. ACM Commun. **18**(11), 613–620 (1975). https://doi.org/10.1145/361219.361220

26. Santos, D., Cardoso, N.: Reconhecimento de entidades mencionadas em português. Linguateca **7**(7), 1 (2007). Portugal
27. Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for Portuguese named entity recognition. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), pp. 437–442. IEEE (2019)
28. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
29. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27 (2015)