*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4828–4834
Marseille, 11–16 May 2020


# Word Embedding Evaluation in Downstream Tasks and Semantic Analogies

**Joaquim Santos, Bernardo Consoli, Renata Vieira**
Pontifical Catholic University of Rio Grande do Sul (PUCRS)
Porto Alegre, Brazil
{joaquim.santos, bernardo.consoli}@acad.pucrs.br,
renata.vieira@pucrs.br

**Abstract**

Language Models have long been a prolific area of study in the field of Natural Language Processing (NLP). One of the newer kinds of language models, and some of the most used, are Word Embeddings (WE). WE are vector space representations of a vocabulary learned by a non-supervised neural network based on the context in which words appear. WE have been widely used in downstream tasks in many areas of study in NLP. These areas usually use these vector models as a feature in the processing of textual data. This paper presents the evaluation of newly released WE models for the Portuguese language, trained with a corpus composed of 4.9 billion tokens. The first evaluation presented an intrinsic task in which WEs had to correctly build semantic and syntactic relations. The second evaluation presented an extrinsic task in which the WE models were used in two downstream tasks: Named Entity Recognition and Semantic Similarity between Sentences. Our results show that a diverse and comprehensive corpus can often outperform a larger, less textually diverse corpus, and that passing the text in parts to the WE generating algorithm may cause loss of quality.

**Keywords:** Language Models, Word Embeddings, Intrinsic Evaluation, Extrinsic Evaluation

## 1. Introduction

Language Models (LM) are an oft studied area of Natural Language Processing (NLP) in current literature. LMs were first developed based on the study of linguistic rules that could not be generalized to languages other than the one they were created for. In the eighties, another generation of LMs were developed, based on statistical models that used conditional probabilities to analyze n-gram sequences. This new strategy became known as N-grams Models, or Statistical LMs (Jing and Xu, 2019). Yet another advancement occurred based on the work of Xu and Rudnicky (2000), which suggests the usage of Neural Networks (NN), more specifically Feedforward NNs, for LMs. The work of Bengio et al. (2003), based on this new advancement, presented a solution to the area's perennial problem: the curse of dimensionality. From there, the usage of Recurrent Neural Networks started becoming more prevalent, which resulted in the adoption of vector space models for word representation, called Word Embeddings (WE) (Mikolov et al., 2009; Mikolov et al., 2010; Mikolov et al., 2013b).

WEs can thus be described as a vector space representing words learned through the non-supervised training of NN. These vector spaces can also be called *embeddings*. Each embedding is the result of successive mathematical operations that happen in the Hidden Layers of the NN, and add semantic meaning to the embeddings.

WEs have been successfully applied to NLP solutions, but the degree of success is heavily dependent on the quality of the learned vectors. Bakarov (2018) discusses the lack of consensus on which WE evaluation method yields the most accurate results when attempting to measure the quality of an *embedding*. Two kinds of evaluation are, however, widely accepted as superior: downstream task evaluation (or using WEs as resources in NLP tasks); and evaluations that explore the semantics and syntax of the WE itself. The former is commonly known as extrinsic evaluation, while the latter is commonly known as intrinsic evaluation.

This work evaluates five WE models for the Portuguese language, trained on a corpus of 4.9 billion tokens. These resources were originally presented in a previous work, were we stacked different Flair Embeddings and WEs for a downstream task (Santos et al., 2019). This work focuses solely on the WEs, evaluating them using both extrinsic and intrinsic methods.

For the intrinsic evaluation, the model answered a set of 17 thousand questions by correctly outputting a word that is related to a second word in the same way that an auxiliary pair of words is related to on another. For the extrinsic evaluation, the model was used as a language resource for two downstream tasks: Named Entity Recognition; and Semantic Similarity between sentence. The results obtained from these were compared to those obtained when using WEs from the NILC WE repository.

All of the models evaluated here are available for free use in our webpage[1].

This work is organized in six sections as follows: Section 2 presents the most relevant works for this area; Section 3 presents the WE models that already exist for the Portuguese language and the agorithms used in their generation; Section 4 presents the evaluation datasets, the results of the experiments and a discussion about them; and Section 5 presents the conclusion and future works.

## 2. Related Work

Mikolov et al. (2013a) developed a non-supervised NN, called Word2Vec (W2V), capable of capable of capturing and distributing the semantic and syntactic properties of words that are automatically learned by the NN. W2V uses two architectures (also called strategies) to train their vector space models: the first is Skip-gram, and the second CBOW. The authors generated one model for each architecture with an input corpus of 6 billion tokens. In or-

[1] http://www.inf.pucrs.br/linatural/wordpress/pucrs-bbp-embeddings/

| Language | Type of Relation | $(w_1, w_2)$ | $(w_3, w_4)$ |
|---|---|---|---|
| PT-BR | Syntactic (plural-verbs) | (trabalhou, trabalham) | (embaralhar, embaralham) |
| PT-EU | Syntactic (plural-verbs) | (trabalhou, trabalham) | (baralhar, embaralham) |
| PT-BR | Semantic (family) | (padrasto,madrasta) | (rapaz, moça) |
| PT-EU | Semantic (family) | (padrasto, madrasta) | (rapaz, rapariga) |

Table 1: Example of Semantic and Syntactic Relations

der to evaluate these models, a benchmark questionnaire composed of 8,869 semantic questions and 10,675 syntactic question was created. The idea behind this benchmark is to verify whether or not a model correctly learned the expected syntactic and semantic relations from the training corpus.

dos Santos and Zadrozny (2014) used W2V's Skip-gram architecture to develop a Portuguese language WE model. The training corpus contained 401 million tokens, and was composed of texts from the Portuguese Wikipedia, CENTEMFolha, and CENTEMPublico. The model was trained for word-level representation in order to perform the tasks of Part-of-Speech tagging and NER (dos Santos and Guimarães, 2015). This was meant to show that the non-supervised learning of the model enhanced semantic and syntactic representation, thereby improving token classification tasks. This work did not make use of intrinsic evaluation, however.

Fonseca and Aluísio (2016) generated a W2V model with a 228 million token corpus composed of CETENFolha and CETEMPublico. It was evaluated extrinsically through Part-of-Speech tagging.

Rodrigues et al. (2016) trained a Skip-gram W2V model with a 1.7 billion token corpus which combined European Portuguese texts and Brazilian Portuguese texts. The authors only performed intrinsic evaluations, and created the first intrinsic benchmark questionnaire for the Portuguese language.

Hartmann et al. (2017) trained a total of 35 WE models for the Portuguese language or a corpus of 1.3 billion tokens. The authors evaluated them on the same benchmark questionnaire as Rodrigues et al. (2016). They also performed extrinsic evaluations with Part-of-Speech tagging and semantic similarity.

Rodrigues and Branco (2018) trained a Skip-gram W2V model with a 2.2 billion token corpus which combined European Portuguese texts and Brazilian Portuguese texts. This corpus is an expansion on the corpus produced by Rodrigues et al. (2016). The model was evaluated on two new benchmarks proposed in the work, as well as the one described in Rodrigues et al. (2016).

Santos et al. (2019) trained W2V and FastText WE models using a 4.9 billion token corpus with texts in European and Brazilian Portuguese. NER was used for an extrinsic evaluation, but tests were performed stacked embeddings which joined WEs with Contextualized Embeddings (Akbik et al., 2018). That is, the evaluation was not representative of the quality of the WEs, but rather that of the stacked embeddings. Furthermore, no intrinsic evaluations were performed.

In general, the works that involve creating and evaluating Word Embeddings differ both in the corpora used to train the model, and, mainly, in the evaluation method. Of the studied works, only (Hartmann et al., 2017) evaluated its WE intrinsically and extrinsically.

## 3. Word Embeddings Models

Many NLP tasks benefit greatly from the addition of pre-trained WE models. These are usualy loaded into an NN's *embedding* layer that transforms tokens into their vector space equivalents.

Recurrent NNs have been widely used for the generation of WE models since the publication of Mikolov et al. (2010)'s contributions to the area. This kind of NN has the inherent advantage of being capable of capturing larger quantities of context (Mikolov et al., 2011; Sundermeyer et al., 2012).

Since the training of WEs is non-supervised, corpora used to train them do not need any kind of annotation. It also means that a great quantity of text is necessary during training in order to produce a representative word vector space. These training corpora usually have over a billion tokens in total.

According to Mikolov et al. (2013c), WEs are capable of abstracting semantic and syntactic structures. As an example, 1 represents a vector operation between the vectors for the words *Norway, Oslo* and *Havana*. The result of this operation should be the word vector for *Cuba*. In other words, equation 1 can also be read as a pair of *country-capital* type semantic relations: $(Oslo, Norway)$ and $(Havana, Cuba)$.

$$[Norway] - [Oslo] + [Havana] \simeq [Cuba] \quad (1)$$

There are many NNs that generate WEs beyond W2V. The most well known alternatives are FastText (Grave et al., 2017), Glove (Pennington et al., 2014) and Wang2Vec (Ling et al., 2015).

**Word2Vec (W2V)** (Mikolov et al., 2013a) is, as previously mentioned, a recurrent NN capable of learning vector space representations of words, as well as capturing semantic and syntactic information, from a textual training corpus. That is, it is meant to produce WEs. It has two training architectures: Continuous Bag-of-Words (CBOW) and Skip-gram. The former seeks to predict a word given a context while the latter seeks to predict a context given a word.

**FastText (FT)** (Grave et al., 2017) is very similar to W2V, also being divided into the CBOW and Skip-gram architectures. The main difference between the two, and FT's main feature, is that it can estimate vectors for tokens which were not part of the training corpus. This is due to its character-

level training model, as opposed to W2V's word level training.

### 3.1. Pre-trained Word Embeddings

This work's evaluations are performed using WEs from two repositories: NILC Embeddings (Hartmann et al., 2017) and PUCRS-BBP (Santos et al., 2019).

**NILC Embeddings** Nilc Embeddings[2] is a repository for WE models trained in Portuguese language texts. The models were generated using W2V, FT, Wang2Vec and Glove. The W2V, FT and Wang2Vec models are available in both Skip-gram and CBOW versions. All models are available in 50, 100, 300, 600 and 1000 dimensions. The training corpus used to train them is composed by 17 Portuguese language textual corpora, some Brazilian and others European. The corpus has a total of 1.3 billion tokens. This work only used 300-dimensional W2V and FT models.

**PUCRS-BBP Embeddings** PUCRS-BBP Embeddings is another WE repository for the Portuguese language. It was created for or previous work (Santos et al., 2019), and has four models: 300-dimensional Skip-gram and CBOW trained with W2V and 300-dimensional Skip-gram and CBOW trained with FT. They were trained with the Gensim framework and, to minimize the use of RAM, we divided the corpus into batches of 10 million tokens. That is, each batch of text is loaded into memory (RAM) only when necessary and removed when Gensim requestes a new batch. An experimental FT model was also trained with the entire corpus at once, that is, we loaded the entire corpus of 4.9 billion tokens into RAM, making Gensim understand that there is only a single text file. The training corpus was composed of 4.9 billion tokens and is composed of three corpora: BrWaC (Wagner Filho et al., 2018), BlogSet-Br (dos Santos et al., 2018) and ptwiki-20190301[3].

- **BrWaC:** is a Brazilian Portuguese language corpus composed of the collection of *.br* domain webpages. It has, in total, 3.53 million documents, or 2.6 billion tokens (Wagner Filho et al., 2018). After pre-processing, the corpus had 2.9 billion tokens.
- **BlogSet-Br:** is a Brazilian Portuguese language corpus composed of blog pages. It has, in total, 7.4 million posts, or 2.7 billion tokens (dos Santos et al., 2018). After pre-processing, the corpus had 1.8 billion tokens.
- **ptwiki-20190301[4]:** is the Portuguese Wikipedia dump from March 2019, with a total of 162 million tokens (after pre-processing).

## 4. Evaluations

Our evaluation method follows the approach of Hartmann et al. (2017), as they evaluated WEs both extrinsically and intrinsically. While this work's intrinsic evaluation we perform is the same as Hartmann et al. (2017)'s, this work's extrinsic evaluation is composed of two tasks: NER and semantic similarity between sentences (SSS). Since the WEs trained by Hartmann et al. (2017) (found in NILC's repository) have been updated since publication of the original paper, we performed the tests therein described again, having obtained updated results from those reported.

Below, we will present the evaluation datasets. The benchmark dataset Analogies was adopted for the intrinsic evaluation, while First HAREM (Santos and Cardoso, 2007) and the ASSIN dataset (Fonseca et al., 2016) were used for the NER and SSS intrinsic evaluations, respectively.

### 4.1. Evaluation datasets

**Analogies** is a intrinsic evaluation benchmark dataset for WEs. It was initially developed by Mikolov et al. (2013a) for the English language, but was later translated into two Portuguese datasets (one European and one Brazilian) by Rodrigues et al. (2016). The English dataset has 19,544 questions, while the Portuguese datasets have 17,558 questions each. The discrepancy comes from the fact that the translation of certain terms resulted in multigrams, but WEs can only handle unigrams, so those had to be removed. The dataset presents two types of questions: semantic and syntactic. Semantic questions are divided into 5 types of relations while syntactic questions are divided into 9 types of relations. Some examples of these can be seen in Table 1 for both Brazilian (PT-BR) and European (PT-EU) Portuguese.

**First HAREM and Mini-Harem** are Gold corpora for the NER task in the Portuguese language. They were used in our previous work (Santos et al., 2019) to train and test NER models. More specifically, First HAREM was used to train the models, while Mini-Harem was used to test them. These corpora are divided into two scenarios: total and selective. The Total scenario is composed of ten Named Entity (NE) categories, while the Selective scenario is composed of the five most represented NE categories of those ten. These five categories are: PERSON, LOCATION, ORGANIZATION, TIME and VALUE. Only the Selective scenarios was used in our evaluation.

**ASSIN Corpus** is a Portuguese language annotated dataset for the SSS task. The SSS extrinsic evaluation benchmark in our evaluation was the ASSIN Corpus' test set. It is composed of 2000 pairs of sentences in Brazilian Portuguese and 2000 pairs of sentences in European Portuguese. Each of these sentence is graded with a similarity score of 1 to 5, with 1 being no similarity and 5 being complete similarity.

### 4.2. Intrinsic evaluation

Intrinsic evaluations of WEs consist of evaluating the WE by itself, and not its performance in a downstream task. This section presents the results achieved by this work in this category of evaluation. The Analogies benchmark dataset and scripts used for these evaluations were made available by Hartmann et al. (2017). The test itself consists of the WE model correctly predicting the fourth element of two word pairs, as exemplified in equation 1, for each two

[2] http://nilc.icmc.usp.br/embeddings

[3] https://dumps.wikimedia.org/ptwiki/20190301/

[4] https://dumps.wikimedia.org/ptwiki/20190301/

pairs presented in the Analogies dataset. This enables the calculation of accuracy, presented in Table 2.

It is immediately clear that the best performing NILC WE model was the FT-SKPG, for both variations of Portuguese, while NILC's worst performing model was W2V-CBOW. As for the PUCRS-BBP models, BigFT-SKPG obtained the best results for the semantic question sets and the complete question sets. FT-CBOW, however, obtained better results for the syntactic question set.

It is important to not that there was not much variation in results between the Brazilian Portuguese and European Portuguese datasets. These observations were mathematically confirmed with the calculation of variance and standard deviation between the results for Brazilian and European Portuguese. NILC's WEs had a maximum variance of $s^2 = 1.13$ and a maximum standard deviation of $s = 1.06$. PUCRS-BBP's WEs had a maximum variance of $s^2 = 4.81$ and a maximum standard deviation of $s = 2.19$.

As the variance was not found to be significative, the averages of the variances between the results for the two Portuguese datasets were calculated, and are shown in Figure 1. Equations 2, 3, and 4 are each a formalization of the calculations for the syntactic question set, the semantic question set and the complete question set, respectively. $i$ represents a given WE model.

$$M_{Syntactic} = \frac{1}{2}(Syntactic_i^{BR} + Syntactic_i^{EU}) \quad (2)$$

$$M_{Semantic} = \frac{1}{2}(Semantic_i^{BR} + Semantic_i^{EU}) \quad (3)$$

$$M_{all} = \frac{1}{2}(All_i^{BR} + All_i^{EU}) \quad (4)$$

As a rule, NILC models performed better when answering questions involving syntactic relations, while showing average performance when answering questions involving semantic relations. This disparity can be seen in the performance for the complete question set, with its results between the high syntactic and average semantic questions sets' results. Additionally, it can be seen that NILC's FT models obtained superior results to their W2V models.

PUCRS-BBP's models' results were similar to those of NILC's models. The best results were achieved in the syntactic question dataset, while the semantic question dataset's results were poor. Figure 1 shows that PUCRS-BBP's W2V models performed better than NILC's, while PUCRS-BBP's FT models performed worse. It can also be observed that BigFT-SKPG's performance in the semantic question datasets far exceeded those of FT-SKPG and FT-CBOW, which resulted in better results for the complete question dataset.

### 4.3. Extrinsic Evaluation

The goal of these extrinsic evaluations is to observe the changes in performance between NNs trained for downstream tasks when they are exactly the same but for the WE model used to represent words as vectors. The difference in quality is most perceptible when the tasks in question require high quality contextual word representations, disambiguation, and syntactic and semantic knowledge (all of which a WE is supposed to automatically learn).

The NER experiments were performed using a BiLSTM-CRF NN from the Flair framework (Akbik et al., 2018; Akbik et al., 2019). A BiLSTM-CRF network is composed of two modules: the first module receives natural language sentences as sequences of characters and recovers vector space representations for each word through the use of an embedding model (a WE model, in this case); the second module uses these vector space word representations to incorporate new features into the words and classify each one with a *Conditional Random Fields* (CRF) probabilistic classifier.

Nine NER tests were performed using this NN, one for each WE model. Table 3 presents the results from these tests for both NILC and PUCRS-BBP repositories.

The best performing model, as measured by F-measure, from the NILC repository was FT-CBOW ($F_1 = 69.11\%$), while the worst performing model was W2V-CBOW ($F_1 = 67.02\%$). The variance of NILC model's F-measures is $s^2 = 0.9577$, and the standard deviation of NILC model's F-measures is $s = 0.9786$. These measures show that these models are stable, and don't vary much based on the architecture used to create the model.

The best performing model from PUCRS-BBP was the experimental BigFT-SKPG ($F_1 = 69.93\%$), having also achieved better results than any of the NILC models. The worst performing model was W2V-SKPG ($F_1 = 58.33\%$), having also achieved worse results than any of the NILC models. Excluding BigFT-SKPG, the variance and standard deviation of PUCRS-BBP model's F-measures were $s^2 = 3.3248$ and $s = 1.8234$, respectively. These measures show that our corpus causes more WE model instability in the NER task than NILC's models. Additionally, *BigFT-SKPG*'s better performance over its batch-trained counterpart (FT-SKPG) indicates that training with the entire corpus at once may result in better models.

For the SSS task, tests were performed using the same approach and scripts as Hartmann et al. (2017). The results for these tasks are presented in Table 4. This table is divided into two benchmarks: the Brazilian Portuguese benchmark and the European Portuguese benchmark. the evaluation metrics for this task were Pearson Correlation ($\rho$) and Mean Squared Error (MSE). The closer to 1, the better the ($\rho$), while ideal MSE scores approach 0.

Table 4 shows that the best NILC model was W2V-CBOW for both variations of Portuguese. PUCRS-BBP's best performing model was FT-SKPG, and, curiously, its worst performing model was W2V-CBOW. PUCRS-BBP best models achieved comparable or superior results when compared to NILC's models.

## 5. Conclusion

In this work, we presented extrinsic and intrinsic evaluations of five new Word Embedding models for the Portuguese language using the Word2Vec and FastText neural networks from the Gensim framework. The intrinsic benchmarks involved testing of semantic and syntactic knowledge and resulted in 26 tests performed. Extrinsic evalu-

| | | BR | | | EU | | |
|---|---|---|---|---|---|---|---|
| **Repository** | **Model** | **Syntactic** | **Semantic** | **Complete** | **Syntactic** | **Semantic** | **Complete** |
| **NILC** | W2V-SKPG | 33.0% | 15.6% | 24.3% | 32.2% | 14.1% | 23.2% |
| | W2V-CBOW | 24.7% | 4.6% | 14.7% | 24.5% | 4.5% | 14.5% |
| | FT-SKPG | **58.7%** | **32.3%** | **45.5%** | **58.6%** | **31.2%** | **44.8%** |
| | FT-CBOW | 52.0% | 8.4% | 30.2% | 52.0% | 9.2% | 30.5% |
| **PUCRS-BBP** | W2V-SKPG | 26.3% | 17.7% | 22.0% | 26.3% | 14.6% | 20.5% |
| | W2V-CBOW | 45.9% | 18.1% | 32.1% | 45.7% | 16.3% | 31.1% |
| | FT-SKPG | 43.8% | 5.8% | 24.9% | 44.1% | 5.2% | 24.7% |
| | FT-CBOW | **50.4%** | 6.4% | 28.6% | **50.6%** | 5.6% | 28.2% |
| | BigFT-SKPG | 46.2% | **21.3%** | **33.9%** | 46.1% | **18.5%** | **32.4%** |

Table 2: Intrinsic Evaluation

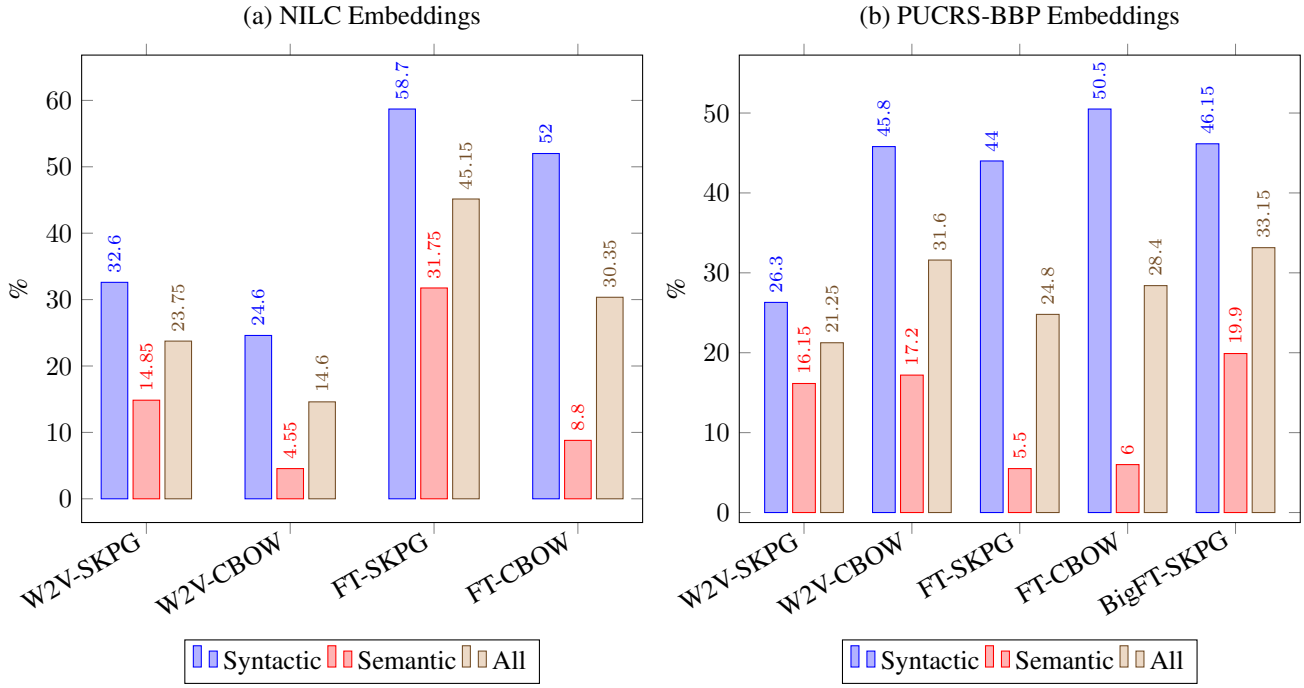


Figure 1: Average between the results of the intrinsic assessment in the Brazilian and Portuguese variations, according to equations 2, 3 and 4.

ations were performed with the use of two tasks, Named Entity Recognition and Semantic Similarity between Sentences, for a total of 18 tests.

Our results show that NILC's WE models outperform our new models in intrinsic evaluation, but some of them, such as BigFT-SKPG in the NER task, can outperform them in extrinsic tasks. It was also found that, in general, FastText models achieved better results than Word2Vec models, with the exception of the results achieved in the Semantic Similarity task.

It must be reported that, even though the training corpus of PUCRS-BBP was much larger than NILC's, the results when applied the the evaluation's tasks were generally inferior. This leads us to believe that the quality and textual diversity of NILC's corpus is significantly greater than PUCRS-BBP's. This can be clearly seen in the ammount of different corpora used to compose each corpus: PUCRS-BBP is composed of 3 corpora, while NILC's is composed of 17.

Furthermore, we've demonstrated that batch training of WE models may cause significant losses in the quality of representations relevant to some tasks. This can be seen in how BigFT-SKPG outperformed FT-SKPG in the NER task.

Future work includes a study of how batch training in the Gensim framework impacts embedding quality and a study into how corpus diversity impacts the quality of word embeddings, and why it might be more impactful than mere corpus size.

## Acknowledgments

We would like to thank CNPq and CAPES for their financial support in the execution of this work.

| | NILC | | | PUCRS-BBP | | |
|---|---|---|---|---|---|---|
| **Model** | **PRE** | **REC** | **$F_1$** | **PRE** | **REC** | **$F_1$** |
| W2V-SKPG | 74.84% | 60.72% | 67.05% | 74.11% | 48.09% | 58.33% |
| W2V-CBOW | **74.94%** | 60.62% | 67.02% | 73.45% | 54.29% | 62.43% |
| FT-SKPG | 74.81% | 61.96% | 67.78% | 72.19% | 52.36% | 60.69% |
| FT-CBOW | 73.18% | **65.47%** | **69.11%** | 73.24% | 53.60% | 61.90% |
| BigFT-SKPG | - | - | - | **74.83%** | **65.64%** | **69.93%** |

Table 3: Extrinsic Evaluation with the NER task

| | NILC | | | | PUCRS-BBP | | | |
|---|---|---|---|---|---|---|---|---|
| | **BR** | | **EU** | | **BR** | | **EU** | |
| **Model** | $\rho$ | **MSE** | $\rho$ | **MSE** | $\rho$ | **MSE** | $\rho$ | **MSE** |
| W2V-SKPG | 0.52 | 0.56 | 0.48 | 0.93 | 0.54 | 0.54 | 0.53 | 0.85 |
| W2V-CBOW | **0.55** | **0.53** | **0.54** | **0.87** | 0.51 | 0.56 | 0.53 | 0.87 |
| FT-SKPG | 0.51 | 0.56 | 0.52 | 0.87 | **0.55** | **0.53** | **0.55** | **0.83** |
| FT-CBOW | 0.37 | 0.66 | 0.37 | 1.02 | 0.53 | 0.55 | 0.55 | 0.85 |
| BigFT-SKPG | - | - | - | - | **0.55** | **0.53** | 0.54 | 0.84 |

Table 4: Extrinsic Evaluation with the SSS task

## 6. References


Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 54–59.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. *Computing research repository - arXiv*, abs/1801.09536.

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

dos Santos, C. N. and Guimarães, V. (2015). Boosting named entity recognition with neural character embeddings. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics*, pages 25–33.

dos Santos, C. N. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1818–1826.

dos Santos, H. D., Woloszyn, V., and Vieira, R. (2018). Blogset-br: A brazilian portuguese blog corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Fonseca, E. R. and Aluísio, S. M. (2016). Improving POS tagging across portuguese variants with word embeddings. In *Computational Processing of the 12th International Conference on the Computational Processing of Portuguese*, pages 227–232.

Fonseca, E. R., dos Santos, L. B., Criscuolo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8–2:3–13.

Grave, E., Mikolov, T., Joulin, A., and Bojanowski, P. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 427–431.

Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131.

Jing, K. and Xu, J. (2019). A survey on neural network language models. *Computing research repository - arXiv*, abs/1906.03591.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304.

Mikolov, T., Kopecký, J., Burget, L., Glembek, O., and Cernocký, J. (2009). Neural network based language models for highly inflective languages. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 4725–4728.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048.

Mikolov, T., Kombrink, S., Burget, L., Cernocký, J., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *Proceedings of the Interna-*

*tional Conference on Acoustics, Speech, and Signal Processing*, pages 5528–5531.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

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 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Rodrigues, J. and Branco, A. (2018). Finely tuned, 2 billion token based word embeddings for portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Rodrigues, J. A., Branco, A., Neale, S., and Silva, J. R. (2016). Lx-dsemvectors: Distributional semantics models for portuguese. In *Computational Processing of the 12th International Conference on the Computational Processing of Portuguese*, pages 259–270.

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

Santos, J., Consoli, B., dos Santos, C., Terra, J., Collovini, S., and Vieira, R. (2019). Assessing the impact of contextual embeddings for portuguese named entity recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, pages 437–442.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). LSTM neural networks for language modeling. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association*, pages 194–197.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Xu, W. and Rudnicky, A. (2000). Can artificial neural networks learn language models? In *Proceedings of the 6th International Conference on Spoken Language Processing*, pages 202–205.