# Reasoning with Portuguese Word Embeddings

**Luís Filipe Cunha**
Department of Informatics, University of Minho, Braga, Portugal

**J. João Almeida**
Centro ALGORITMI, Departamento de Informática, University of Minho, Braga, Portugal

**Alberto Simões**
2Ai – School of Technology, IPCA, Barcelos, Portugal

## Abstract

Representing words with semantic distributions to create ML models is a widely used technique to perform Natural Language processing tasks. In this paper, we trained word embedding models with different types of Portuguese corpora, analyzing the influence of the models' parameterization, the corpora size, and domain. Then we validated each model with the classical evaluation methods available: four words analogies and measurement of the similarity of pairs of words. In addition to these methods, we proposed new alternative techniques to validate word embedding models, presenting new resources for this purpose. Finally, we discussed the obtained results and argued about some limitations of the word embedding models' evaluation methods.

**2012 ACM Subject Classification** Computing methodologies → Natural language processing; Computing methodologies → Machine learning

**Keywords and phrases** Word Embeddings, Word2Vec, Evaluation Methods

**Digital Object Identifier** 10.4230/OASIcs.SLATE.2022.17

**Funding** *J. João Almeida*: This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.
*Alberto Simões*: This project was funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020.

## 1 Introduction

The use of word embedding models to create words' numerical representations is a widely used technique in natural language processing tasks. In this paper, an analysis of the word embedding models' training and validation was performed where we observed the influence of training parameterization and the classical methods of evaluating these models and measuring their quality.

In this paper, to train word embeddings, three corpora were used: a thematic corpus with reviews from the Portuguese trip advisor; The AC/DC corpus from Linguateca that contains cultural content, literature, news, etc. and corpora from Opus Projec; A corpus of an encyclopedic nature extracted from the Portuguese Wikipedia.

Firstly, we created several models with the same corpora with different hyper-parameters such as vector dimensions and epochs and even created models with different architectures: Skip-gram and CBOW. Then, to make some assumptions about these models, we validated them with the usual evaluation methods, four-word analogies, and measuring two-word distance. Furthermore, during the validation process, we also observed the differences in creating word embeddings with different types of corpora with different sizes, analyzing the results of each model.


In addition to these, we also propose new validation methods and resources to make better judgments about the model's behavior. The created resources are based on the minimum distance classification to a set of class classifiers allowing us to calculate words' polarity, classes and similarities to multiple concepts. During the creation of these resources, we tried to use a domain-specific language, so they were easy to use and adapt to new problems.

We also identified some problems that we faced during this work associated with the word embeddings' evaluation methods and the limitations of this word representation method.

## 2 Related Work

Recently, interest in word embedding models has been increasing since they have played an essential role in improving the performance of several natural language processing tasks. In English, there is already a wide variety of word embedding models trained in different contexts, however, for Portuguese, it is not always easy to find pre-trained models that satisfy our requirements.

In recent years, several papers were published where word embeddings created from Portuguese corpora were trained and validated. In [15] they trained word embedding models in Portuguese corpora with 536 805 758 tokens and validated them with a set of resources generated in [9, 7, 8, 1] and [2], later translated into Portuguese in [13].

In [6] 31 word embedding models were trained on a Portuguese corpus with 1 395 926 282 tokens using four different techniques: FastText, GloVe, Wang2Vec and Word2Vec. In this paper they evaluated the word embeddings using semantic analogies and task-specific evaluations such as POS and other sentence semantic similarity tasks. Here they concluded that word analogies were not appropriate for word embedding evaluation and that using task-specific evaluations appeared to be a better option.

In [11] they created a new Portuguese test set for Portuguese word embeddings validation. This validation method is focused on lexical-semantic relations and was compared to previous word embeddings evaluation methods.

Other works have been done recently where pre-trained word embeddings have been used to initialize task-specific models in order to transfer their vocabulary knowledge to other classification tasks. An example of this is [3] where several ML models were initialized with pre-trained Portuguese word embeddings and then trained to perform Named Entity Recognition on Portuguese archival documents.

## 3 Models

In order to train our word embeddings, we used the Word2vec technique with two different models, Continuous Bag-of-Words(CBOW) and Skig-Gram [9]. A CBOW model uses the word's context to predict the current word, i.e., using the $N$ words after and before the token to calculate its representation. Then the model uses a sliding window to go through all the words of the training corpus. Skip-Gram models are similar to CBOW, however, instead of using the word's context to predict the hidden word, given a word, it predicts other words in the sentence, its neighboring words. Training models with both methods allowed the creation of a high dimensional semantic vector space where words with similar semantic meaning are close in distribution.

To train word embedding models, we used the Gensim library [14] that allowed us to fine-tune the model hyper-parameters. In this paper, we trained several different models with different configurations of vector dimensions, epochs and architectures. To keep the models

**Table 1** Corpora used for the training of word embeddings.

| Corpus | Sentences | Tokens | Characters |
|---|---|---|---|
| Wikipedia-PT | 13 551 925 | 298 505 778 | 1 905 522 756 |
| TripAdvisor | 98 778 | 1 040 654 | 6 486 772 |
| LinguOpus | 30 346 096 | 638 513 424 | 3 670 306 994 |

organized, we used a notation to name our models depending on the training hyper-parameters: corpus-dimensions-epochs-architecture. For example the model `wiki-300-20-cbow` is a CBOW model trained with 300 dimensions and 20 epochs.

## 4 Datasets

In this paper, we created several word embedding models using three different datasets (Table 1): Natcorpus-TripAdvisor PT – TripAdvisor reviews, Portuguese Wikipedia and LinguOpus corpus.

### Wikipedia-pt

The Portuguese Wikipedia corpus was obtained from a Wikipedia dump made on 2022-04-01, which contained approximately 152 million sentences made up of 903 million tokens. Despite this, the raw format used by Wikipedia contains a lot of non-linguistic elements, so we had to perform some data cleaning. The resultant dataset is constituted only by the textual elements of the original wiki corpus.

### Natcorpus-TripAdvisor PT

The Natcorpus-TripAdvisor PT is composed of 13 158 reviews of hotels, near 1 million words, written in Portuguese, from TripAdvisor . These reviews also contain a assessment (1..5). It includes Portuguese from Brazil, Portugal, Angola, Moçambique, Cabo Verde and Timor. The reviews contain emojis abbreviations, errors.

### LinguOpus

LinguOpus includes texts in European Portuguese obtained from two main sources: the Linguateca[1] corpus [16] and the Opus Project[2] [17]. From the first, all corpora tagged as European Portuguese were used. From the second, the Portuguese part of European parallel corpora was used. In addition to these two main sources, it also includes some Portuguese parts from the Per-Fide[3] project corpora ([4]), namely Le Monde Diplomatique, The Vatican Corpus and the European Central Bank corpus.

[1] `https://www.linguateca.pt/ACDC/`
[2] `https://opus.nlpl.eu/`
[3] `http://per-fide.di.uminho.pt/`

# 5 Evaluation Methods and Resources

The classic evaluation method of embeddings models consists of computing word analogies such as `vector("King") - vector("Man") + vector("Woman")` expecting that the resulting vector is closest to the `vector("Queen")`. Another way to validate these models is to calculate the proximity between pairs of words and compare the result with a value estimated by humans, normally between 0 and 10.

## 5.1 Available Resources

In English, there are several resources that can be used for evaluation purposes, such as [9] and [1]. Some of these resources were later translated to Portuguese, enabling to validate models trained with Portuguese corpora. In this paper, we use the following test corpus to validate our models: LX-4WAnalogies, LX-WordSim-353, LX-SimLex-999 and LX-Rare-Word from [13] and NRC-EIL [10].

- **LX-SimLex-999** – Portuguese translation of [7]. In this dataset, the proximity between pairs of words was calculated on a scale between 0 and 10. During the translation of this dataset, the proximity value was recalculated to adjust to the Portuguese language. In total, it contains 999 pairs of words: 666 pairs of noun-noun, 222 pairs of verb-verb and 111 pairs of adjective-adjective.
- **LX-Rare Word Similarity** – Created from [8]. It contains about 1 017 pairs of words extracted from Wikipedia and WordNet and the proximity score associated with each pair. During the translation of this dataset, the proximity value was recalculated and adjusted to the Portuguese language.
- **LX-WordSim-353** – Created from WordSim-353 [1] containing 353 pairs of words and their proximity scores. All these pairs received a human judgment on a proximity scale from 0 to 10. In this case, the proximity scores were preserved during translation.
- **LX-4WAnalogies** – Created from the Portuguese translation of a test dataset used in [9]. This dataset contains analogies represented by sets of four words, for example, `"Brussels Belgium Lisbon Portugal"`. In total, this dataset contains five sections with 8 869 semantic analogies and nine sections with 10 675 syntactic analogies.
- **NRC-EIL (PT-BR)** – The NRC Emotion Intensity Lexicon [10] contains a list of words paired with eighth basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) and their proximity score. In this paper, we used the Portuguese version of this resource that was translated automatically using Google Translate in 2018.
- **LX-Battig** – Portuguese translation of an English dataset [2]. This resource contains 83 terms associated with ten different categories: mammals, birds, fish, vegetables, fruit, trees, vehicles, clothes, tools and kitchenware.
- **TALES** – *Teste para Analogias Lexico-Semanticas* (TALES) [11] is a Portuguese testset focused on lexical-semantic relations, that uses the analogies method, however, accepting more than one acceptable answer for each analogy.
  The creation of this resource was inspired by [5] but created from scratch for the Portuguese Language, containing relations such as Hypernymy, Meronymy, Synonymy and Antonymy. It is a balanced testset with a total of 14 test files with 50 entries each.

## 5.2 Created Resources

In addition to using several already available resources to validate the quality of our word embedding models, we decided to create our own resources and methods in order to have more information about our models' behavior. Thus, we present Class-Sim (Classifier by Similarity) and Analogy-DD (Analogy Declarative Description).

**Listing 1** NAT-WordSim-1 dataset (YAML) used for Class-Sim method.

```
- section: polarity
    clas:
      - [bom]
      - [mau,horrivel,fujam]
    testes:
      - [ acolhedor , 0 ]
      - [ adorei , 0 ]
      - [ ridiculo , 1 ]
      - [ amavel , 0 ]
      - [ ruidoso , 1 ]
      - [ horror , 1 ]
```

```
- section: class-animais
    clas:
      - [ave]
      - [mamifero, cao, gato]
      - [peixe]
    testes:
      - [ melro , 0]
      - [ aguia , 0]
      - [ macaco , 1 ]
      - [ elefante , 1 ]
      - [ tubarao, 2 ]
```

## Classifier by Similarity

The first method that we created was a classifier by similarity (Class-Sim), which tests whether the model is able to associate words with its corresponding class. For that, we created a new YAML dataset **NAT-WordSim-1** (Listing 1) that is divided into sections. Each section contains multiple classes and a set of words associated with one of those classes. The principle behind this method is for the model to measure the cosine similarity of each word with all the classes of the section. Then it assigns each word to a class by choosing the minimum distance value of the word to all the available classes. Note that each class can be made up by more than one word. Then, we validate if the words get assigned with the correct classes.

The resources used to perform this evaluation method are publicly available [4] containing a dataset with three classes: Geo (27 samples), class of Animals (22 samples) and Polarity (86 samples) with a total of 135 word samples. The polarity samples were extracted from the TripAdvisorPT corpus.

Finally, the corpus LX-Battig was also parsed into the YAML format present in Listing 1 to use it with the Class-Sim evaluation method.

## Analogy-DD

Analogies are a very elegant way to show / see if a word-embedding captured a relation (semantic, syntactic, ...) between terms.

The analogy tests, previously described, focus in a set of relations that related to encyclopedic knowledge (country-capital, country-currency, city-in-state), grammatical relations, and family relation.

With Analogy-DD initiative, we intend:

- to discuss analogies, their underlying relations and properties;
- to define a declarative notation to describe facts and analogy types;
- to create a sets of facts based on relations;
- to build a set of tools that generates analogies from the previous facts. The generated analogies tests

The problem of analogies and relations is in fact complex. Relations have different signatures and properties.

[4] https://github.com/lfcc1/slate22-wemb

Consider the following situations:

- multiple answers. Example:
  - country-people: $England : English :: Poland : x \in \{polaco, polonês\}$
- transitive relation. Example "is-a" (classification) is a transitive relation; cat is a feline, but also a mammal and an animal:
  - is-a: $ant : insect :: cat : x \in \{feline, mammal, animal\}$
- multi word elements. Example:
  - capital-country: $Lisboa : Portugal :: Londres : x \in \{$Reino Unido, $Inglaterra\}$

Consider the following extract from a Analogy-DD in Portuguese:

```
# Relações para uso na geração de Analogias (e outros)
# V0.3  2022-05-01

# $1(ferramenta) é_usado_para $2(actividade,V)

martelo :: pregar, martelar
verruma :: furar
chave de fendas :: aparafusar
berbequim :: furar
lixa :: lixar
serrote, serra :: serrar
```

Notes:

- line 1,2 – metadata
- line 3 – defines a schema of a new section:
  - `$1(ferramenta)` – a value selected from the the first column (this value is a *ferramenta* (tool),
  - `é_usado_para` – (is used to) name or the relation,
  - `$2(actividade,V)` – value selected from second column (a activity, POS=verb)
- line 4...9 – tuples. Example: *lixa :: lixar* can be read *lixa é_usada para lixar*.
- An analogy is built by joining 2 different tuples. (Ex: serra : serrar :: verruma : ?furar).
- Some assessment tools may need explicit constrains (no multi-word allowed on the elements (like *chave de fendas*), or no variants allowed on the answer (like *pregar* or *martelar*) – .

Analogy-DD may have:

- several sections;
- with several relation schema;
- tuples may have 2 or more fields;
- each field may have several values (separated by ",").

The current version of the Portuguese Analogy-DD has 37 sections (near 37 relations), 410 tuples, and generates more then 5000 analogy lines[5].

[5] This number may change according to the selected options.

## 6 Results

In this section, the results of the experiments made in this work are presented. Initially, we trained several models with vectors of different dimensions in order to analyze the impact of this hyper-parameter on the model's performance and computational cost.

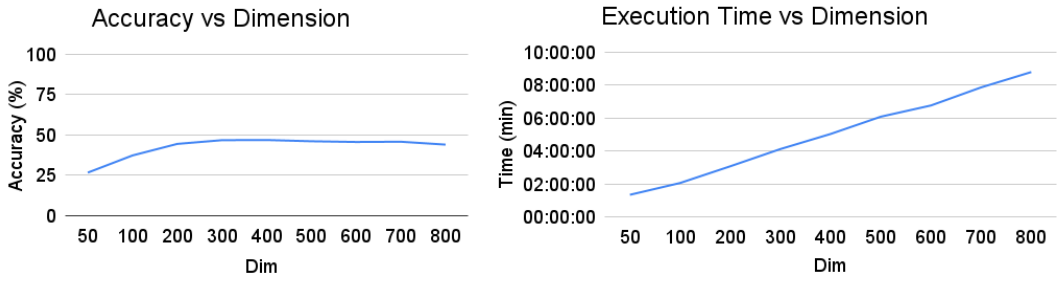


**Figure 1** Comparison of vector dimension with accuracy and execution time on LX-4Analogies task.

In Figure 1 we have a performance analysis of models trained on the Portuguese Wikipedia corpus, where several models with different word embeddings dimensions were trained and then validated with the LX-4Analogies dataset. As we can see in Figure 1 (Left), the model trained with 50 dimensions obtained an accuracy of 25%. As we increased the dimensions of the word embeddings, the model's accuracy also increased, reaching values of approximately 50% with 300 dimensions vectors. However, models trained with dimensions number greater than 300 were not able to obtain higher results. In fact we registered an accuracy decrease in models trained with 700 and 800 vector dimensions.

On the other hand, in Figure 1 (Right), we performed an analysis at each model's validation execution time. This Figure shows that the execution time is proportional to the number of dimensions used to train the models. Thus, since models trained with dimensions greater than 300 did not obtain superior performances, we decided to use models with 300 dimensions to perform the tests in this paper. This dimensional value is consistent with most of the related work done in this field where several researches [12],[9],[6] concluded that the models achieved the best performance by using 300 dimensions.

Then we also analyzed the models' behavior with different epoch values.

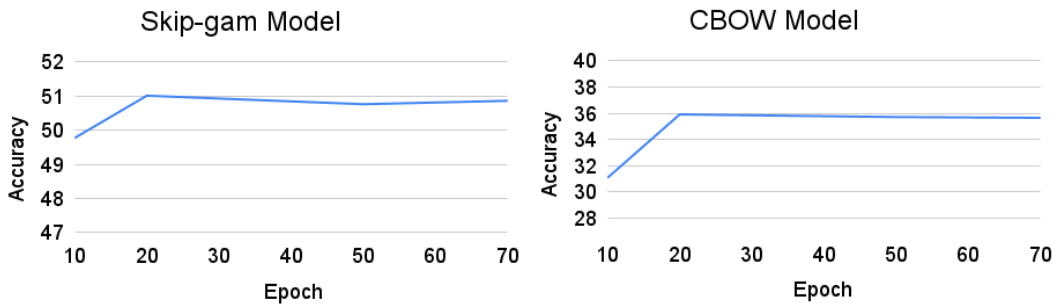


**Figure 2** Comparison of models epochs with accuracy on LX-4Analogies task.

**Table 2** Models' accuracy results with Class-Sim method and Analogies.

| Model | NAT-WordSim-1 | LX-Battig | Analogies-DD | LX-4WAnalogies |
|---|---|---|---|---|
| TripAdvisorPT | **95.06** | oov | 5.19 | 4.13 |
| LinguOpus | 79.55 | 41.77 | 21.04 | 29.86 |
| wiki-300-70-sg | 86.36 | 66.67 | 23.91 | 50.86 |
| wiki-300-50-sg | 86.36 | 62.82 | **24.29** | 50.76 |
| wiki-300-20-sg | 84.78 | **67.95** | 22.74 | **51.01** |
| wiki-300-10-sg | 84.85 | 66.67 | 22.57 | 49.77 |
| wiki-300-70-cbow | 75 | 38.46 | 21.26 | 35.65 |
| wiki-300-50-cbow | 71.21 | 38.46 | 21.43 | 35.70 |
| wiki-300-20-cbow | 74.24 | 39.74 | 21.65 | 35.91 |
| wiki-300-10-cbow | 75 | 34.62 | 20.35 | 31.08 |

In Figure 2 we can see the accuracy of the model trained with different epoch numbers, validated on the LX-4Analogies dataset. Increasing the model epochs from 10 to 20, we see that the models increased their performance, especially the CBOW model from 32% to 36%, which is a considerable performance gain. However, going upwards with the epoch number, the models' performance remained stable in this concrete task.

## 6.1 Analogies and Class-Sim

Then we validated the models with our evaluation resources, starting with the Class-Sim and analogies resources. In Table 2 we can observe the results of the Class-Sim and analogies evaluations on eight models, six of them trained on the Portuguese Wikipedia corpus with the Skip-gam (sg) and CBOW (cbow) architectures, vector dimensions of 300 trained with 10, 20, 50 and 70 epochs. Note that some models were not able to perform some validation tests due to the excess of out of vocabulary words (oov).

Firstly we have the NAT-WordSim-1 test corpus evaluated with the Class-Sim method. In this evaluation, we can see that the TripAdvisorPT was the model that obtained higher results. In fact, this evaluation dataset contains a high number of polarity test samples extracted from the TripAdvisor reviews. One could say that this model was the one to create better polarity associations between the words used in this test resource. This is expected since the TripAdvisorPT model was trained with the same data as this evaluation test.

We can also see that the models trained on the Wikipedia corpus with Skip-gram architecture could obtain better results than those trained with CBOW.

Then we have another Class-Sim evaluation test with the LX-Battig. In this case, the TripAdvisor model was not able to produce any results because it was trained in a very specific domain (reviews from Trip Advisor). As such, it cannot recognize most of this test resource vocabulary. The model that obtained higher results in this test case was the wiki-300-20-sg.

Secondly, we have analogies evaluation resources. Again, the TripAdvisorPT model obtained low results in these test cases due to its limited vocabulary. On the other hand, some models were able to achieve results of 51.01% in the LX-4WAnalogies dataset and 24.29% on Analogies-DD. As stated in Section 5, the Analogies-DD resource was created in

**Table 3** LX-4WAnalogies accuracy results by section.

| Section | wiki-300-20-sg | wiki-300-20-cbow | TripAdvisorPT | LinguOpus |
|---|---|---|---|---|
| capital-common-countries | **83.55** | 47.4 | oov | 25.97 |
| capital-world | **67.65** | 30.48 | oov | 19.09 |
| currency | **28.92** | 6.59 | oov | 5.58 |
| city-in-state | **51.01** | 13.94 | oov | 4.29 |
| family | **59.29** | 57.62 | 13.33 | 48.57 |
| gram1-adjective-to-adverb | 9.85 | **13.05** | 0 | 10.47 |
| gram2-opposite | 23.19 | **24.82** | 3.57 | 19.38 |
| gram3-comparative | 56.67 | **63.33** | 16.67 | 43.33 |
| gram4-superlative | 8.18 | **14.55** | 3.57 | 9.52 |
| gram5-present-participle | 69.57 | **72.15** | 0 | 65.44 |
| gram6-nationality-adjective | **76.46** | 61.87 | 0 | 54.19 |
| gram7-past-tense | 58.11 | **59.32** | 1.39 | 45.31 |
| gram8-plural | **47.7** | 29.6 | 3.33 | 31.01 |
| gram9-plural-verbs | 39.15 | 46.3 | 4.17 | **50.49** |
| Total accuracy | **51.01** | 35.91 | 4.13 | 29.86 |

order to generate a higher variety of analogies. We consider that the generated analogies are harder to predict since most of them would accept multiple words as a correct answer. Considering this property, the models are expected to obtain lower validation values.

After this, we decided that it was crucial to create a view of these evaluations that was able to show the analogies' results by section. In Table 3, we can analyze the results of the models' validation on the LX-4WAnalogies resource, being possible to observe the model's accuracy associated with each domain section. A view that allows us to analyze the results of each section instead of total accuracy gives us more accurate information about where our model fails, allowing us to make better judgments about its behavior.

Analyzing the models trained on Wikipedia, the Skip-Gram model obtained better overall results than the CBOW model, achieving an accuracy of 51.01% and 35.91%, respectively. In fact, the Skip-Gram model managed to obtain results above 80% in the section of capitals and common countries, 67.65% in the remaining capitals of the world, and 76.46% in the nationality-adjective section. Since the Portuguese Wikipedia corpus is encyclopedic, it contains information about all countries and the relationships of their corresponding capitals and nationalities, making this type of evaluation strongly favorable to this model. However, when we look at the grammatical sections, we see that this model obtains lower results, reaching accuracy values of 9.85% and 8.18% in the gram1-adjective-to-adverb and gram4-superlative sections. These results may occur because Wikipedia is not a grammatical corpus and does not contain enough information about the words' morphology.

To learn the word's grammatical relations the model would need a larger variety of morphological and grammatical samples. These are more frequent in literary and everyday texts, and less frequent in encyclopedic corpora. For example, the adjectives' superlatives are rarely present in Wikipedia texts.

On the other hand, we verified that the model trained with the Portuguese TripAdvisor corpus obtained lower results in this test, 4.13% of accuracy overall. Again, the reason for this could be that this model was trained with fewer data in a specific context, limiting

**Table 4** Models' word similarity validation results with Pearson/Spearman correlation coefficient.

| Model | LX-WordSim-353 | LX-SimLex-999 | LX-Rare Word | NRC-PT |
|---|---|---|---|---|
| TripAdvisorPT | 28.70 / 25.27 | 22.84 / 21.73 | 33.08/ 29.80 | 14.85 / 12.57 |
| LinguOpus | 42.49 / 41.69 | 25.65 / 22.74 | 49.67/ 48.34 | 22.93 / 21.97 |
| wiki-300-70-sg | **52.53 / 56.00** | 34.16 / 32.41 | 50.69 / 49.79 | 28.16 / 26.63 |
| wiki-300-50-sg | **52.53 / 56.00** | 34.16 / 32.41 | 50.76/ 49.75 | 28.16 / 26.63 |
| wiki-300-20-sg | 52.50 / 55.67 | 34.90 / 33.06 | 51.63 / 50.67 | 28.28 / 26.70 |
| wiki-300-10-sg | 51.97 / 55.15 | **35.63 / 33.96** | **51.89 / 50.70** | **28.77/ 27.39** |
| wiki-300-70-cbow | 41.26 / 41.57 | 26.74 / 24.29 | 48.16 / 46.46 | 26.08 / 24.42 |
| wiki-300-50-cbow | 41.63 / 42.42 | 26.86 / 24.45 | 48.02 / 46.28 | 26.15 / 24.57 |
| wiki-300-20-cbow | 41.77 / 42.50 | 27.91 / 25.54 | 47.65 / 45.65 | 25.81 / 24.38 |
| wiki-300-10-cbow | 41.73 / 42.56 | 26.02 / 23.80 | 44.63 /42.80 | 24.03 / 22.79 |

its acquired knowledge to the domain of its training data. We also verified that it did not obtain a classification in some sections. This was because the model's vocabulary is not extensive enough to represent words of some domains, such as the countries, capitals, and nationalities present in the LX-4WAnalogies dataset. In this way, we could not perform some of the validation tests.

## 6.2 Word Similarity Results

With the Class-Sim and analogies tests validated, we proceeded with another evaluation method, Word similarity.

In Table 4 we have the validation results of the models trained with different epochs. We performed word similarity tests in this validation, measuring two different metrics to validate the word embedding models, the Pearson and Spearman correlation coefficient. As can be seen, contrary to tests with analogies, in the word similarity tests, CBOW models are able to obtain results similar to Skip-gram models, sometimes even superior.

Looking at the results' variation based on the number of epochs of the models, we can say that this parameter did not have significant relevance to the results, mainly from 20 epochs upwards.

Again the TripAdvisorPT was the model that obtained the worst results for the same reason we already discussed in the analogies validation. On the other hand, the Wikipedia model was the one that obtained the higher correlation coefficients, followed by the LinguOpus model.

Looking at the validation tests, we can see that the NRC-PT was the test where the models produced the worst results. One possible reason could be that this dataset was translated from English to Portuguese using Google Translate, keeping the word proximity scores. However, the semantic similarity between pairs of words may change across different languages due to linguistic and cultural differences, negatively affecting the models' results.

**Table 5** TALES accuracy results by relation category.

| Relations | wiki-300-20-sg | Correct Questions | Total Questions |
|---|---|---|---|
| Synonym-of-N | 12.01 | 3824 | 31852 |
| Synonym-of-V | 12.64 | 4337 | 34302 |
| Synonym-of-ADJ | 10.47 | 3079 | 29402 |
| Antonym-of-ADJ | 20.20 | 495 | 2450 |
| Purpose-of-D | 12.27 | 601 | 4900 |
| Purpose-of-Inv | 11.12 | 817 | 7350 |
| Part-of-D | 10.54 | 2583 | 24500 |
| Part-of-Inv | 9.90 | 2668 | 26950 |
| Hypernym-of-abstract | 11.07 | 1085 | 9800 |
| Hypernym-of-concrete | 12.20 | 1495 | 12250 |
| Hypernym-of-ACCAO | 12.69 | 1866 | 14700 |
| Hypernym-of-ACCAO-Inv | 13.09 | 2245 | 17150 |
| Hypernym-of-Inv-abstract | 12.03 | 2357 | 19600 |
| Hypernym-of-Inv-concrete | 11.53 | 2542 | 22050 |
| Total accuracy | 10.62 | 27326 | 257256 |

## 6.3 TALES

In order to use TALES resource we had to use the Vecto package[6]. Vecto was used with default parameters i.e., accuracy for the performance measurement and vector offset (3CosAdd) [9] for the analogy solving method. Only one model was used in the TALES validation method, which was the `wiki-300-20-sg` model. In this validation test we can observe the performance of our model in 14 different lexical-semantic relations subcategories, Table 5.

In this test we were able to achieve 10.62% of accuracy, with our best score being the Antonym-of-ADJ relations with 20.20% and the Part-of-Inv relations being the validation test with lower results, 9.9% accuracy. Comparing this results with the original TALES paper, in [11], using the word2vec-SkipGram architecture they were able to achieve 10% of accuracy which is almost the same value achieved by our model.

# 7 Problems during validation

During the word embeddings validation, we encountered some problems that could negatively affect the models' performance measurement.

Starting with the already available datasets used in this paper, in the case of the words similarity evaluation, we have seen that some of these resources are translations from English to Portuguese, preserving the similarity score between pairs of words. Despite some of the translations being made carefully involving more than one human translator, keeping

[6] `https://github.com/vecto-ai`

the same similarity value can be problematic due to the cultural and semantic linguistics differences between different languages. Sometimes it is impossible to map words between different languages while keeping the exact same meaning.

For example, in the resource LX-WordSim-353 (English version) we have the following pairs of words, `"football"` and `"soccer"` with a similarity score of 9.03, which makes sense since football and soccer are both sports and can share much semantic meaning between them. However, looking at the Portuguese version of this resource, this entry was translated to `futebol futebol 9.03`. In Portugal, soccer is one of the most influential national sports, and when the word football is mentioned, it usually refers to the English word soccer. In this example, both `"football"` and `"soccer"` were translated to the same word, `"futebol"`, however, the similarity score remained the same. In this case, the model will measure the distance between the same word vector, which will return a proximity score of 10 (maximum proximity value), however, the validation expects it to return a proximity score of 9.03.

Another problem found in these resources during the translation process was that some words were translated into multiple words. For example the the entry `"gem jewel 8.96"` was translated to *"pedra preciosa joia 8.96"* where the word `"gem"` is translating to `"pedra preciosa"`. This evaluation entry could be invalid for evaluating word embedding models that use vectors to represent single words.

Finally, when we validated our models with the LX-Battig resource with the Class-Sim method, we encountered a limitation in the word embeddings mechanism. This evaluation classified each sample class by minimizing the distance between the words and all the classes. Using the Wikipedia model, we were able to achieve results of 67.95%. Then we decided to analyze the samples that the model failed to classify to draw conclusions about the model's behavior. Some words the model was unable to classify included "cat", "lion" and "tiger", which the model classified as fish instead of mammal. In this case, we found out that the model considers the words cat, lion, and tiger closer to the class fish because it is association these tokens with the cat fish, lion fish, and tiger fish which are all described in detail in the Portuguese Wikipedia.

Other words the model failed to identify were "*Pereira*" and "*Oliveira*" (pear and olive tree), which were identified as person names instead of tree names. In Portugal, due to historical reasons, some tree names are common names of people, which confuses the model. In this case, using only one vector to represent a word can cause ambiguity when a word can be interpreted differently depending on its context domain.

## 8 Conclusions and future work

Some final claims:

- Word embeddings' results are very sensitive to changes on the configuration parameters used in the creation.
- The preprocess stage is very important.
- The available public resources for testing have several important issues, and often are concerned with irrelevant questions. Nevertheless, they are very useful for tuning and improving models. In some of these evaluation tests, 50% of their entries test whether a successful model can determine the capital of a country. We believe that this type of test may not be an optimal indicator to generalize the performance of the model.
- It is crucial to analyze specific examples of the discordant results in order to understand what is being tested and is being learned (surprises are frequent).

- Designing new tests and new use cases proved to be challenging and rich. We end up building a set of functions that will probably constitute a python toolkit in the near future.

For future work, one experiment that would be interesting is to use all the corpora used in this paper to train only one final model and re-evaluate it to understand how the merge of data would affect its performance. Another experiment would be to create contextualized word embeddings, generating multiple word vectors for the same word depending on its context domain. This would allow overcoming some problems we observed in this paper, where a word with different meanings could be hard to disambiguate.

As for our validations resources, we intend to increase the size of the Analogias-DD and NAT-WordSim-1 datasets, adding more samples from a wider variety of domains. By doing so, we would be improving the quality of this evaluation method.

## References


1. Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL: `https://aclanthology.org/N09-1003`.
2. Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254, 2010. `doi:10.1111/j.1551-6709.2009.01068.x`.
3. Luís Filipe da Costa Cunha and José Carlos Ramalho. Ner in archival finding aids: Extended. *Machine Learning and Knowledge Extraction*, 4(1):42–65, 2022. `doi:10.3390/make4010003`.
4. Idalete Dias, Sílvia Araújo, Alberto Simões, José Almeida, Nuno Carvalho, Ana Oliveira, and André Santos. *The Per-Fide Corpus: A New Resource for Corpus-Based Terminology, Contrastive Linguistics and Translation Studies*, pages 177–200. Bloomsbury, April 2014.
5. Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL: `https://aclanthology.org/C16-1332`.
6. Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks, 2017. `doi:10.48550/ARXIV.1708.06025`.
7. Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. `doi:10.1162/COLI_a_00237`.
8. Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL: `https://aclanthology.org/W13-3512`.
9. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.
10. Saif M. Mohammad. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018.
11. Hugo Gonçalo Oliveira, Tiago Sousa, and Ana Oliveira Alves. Tales: Test set of portuguese lexical-semantic relations for assessingword embeddings. In *HI4NLP@ECAI*, 2020.

12 Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL: http://www.aclweb.org/anthology/D14-1162.

13 Andreia Querido, Rita Carvalho, Joao Rodrigues, Marcos Garcia, Joao Silva, Catarina Correia, Nuno Rendeiro, Rita Pereira, Marisa Campos, and António Branco. Lx-lr4distsemeval: a collection of language resources for the evaluation of distributional semantic models of portuguese. *Revista da Associação Portuguesa de Linguística*, 3:265–283, September 2017. doi:10.26334/2183.

14 Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. URL: http://is.muni.cz/publication/884893/en.

15 João Rodrigues and António Branco. Finely tuned, 2 billion token based word embeddings for Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1382.

16 Diana Santos and Eckhard Bick. Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/85.pdf.

17 Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).