Journal of Computer Science

Original Research Paper

Distributional Models with Syntactic Contexts for the Measurement of Word Similarity in Brazilian Portuguese

Eduardo E. Berlitz, Denis A. Araujo, Allan B. Silva, Rodrigo R. Righi and Sandro J. Rigo

Applied Computing Graduate Program, University of Vale do Rio dos Sinos (UNISINOS), Av. Unisinos, 950-São Leopoldo, RS, Brazil

Article history
Received: 23-04-2019
Revised: 04-05-2019
Accepted: 30-05-2019

Corresponding Author:
Allan B. Silva
Applied Computing Graduate Program, University of Vale do Rio dos Sinos (UNISINOS), Av. Unisinos, 950-São Leopoldo, RS, Brazil
Email: albarsil@gmail.com

Abstract: The similarity between words constitutes significant support to tasks in natural language processing. Several works use Lexical resources such as WordNet for semantic similarity and synonym identification. Nevertheless, words out-of-vocabulary or missing links between senses are perceived problems of this approach. Distributional-based proposals like word embeddings have successfully been used to meet such problems, but the lack of contextual information can prevent the achievement of even better results. The distributional models that include contextual information can bring advantages to this area, but these models are still scarcely explored. Therefore, this work studies the advantages of incorporating syntactic information in the distributional models, fostering for better results in semantic similarity approaches. For that purpose, the current work explore existing lexical and distributional techniques regarding the measurement of word similarity in Brazilian Portuguese. Experiments were carried out with the lexical database WordNet, using different techniques over a standard dataset. The results indicate that word embeddings can cover words out of vocabulary and have better results in comparison with lexical approaches. The main contribution of this article is a new approach to apply syntactic context in the training process of word embeddings to a Brazilian Portuguese corpus. The comparison of this model with the outcome of the previous experiments shows sound results and presents relevant complementary aspects.

Keywords: Word Similarity, WordNet, Word Embeddings, Computational Linguistics, Natural Language Processing

Introduction

The identification of the semantic similarity between words is a research subject explored in the last years (Pawar and Mago, 2018). It is key to a series of tasks in Natural Language Processing (NLP) (Islam *et al*., 2008; Sravanthi and Srinivasu, 2017; Oliveira, 2018). The importance of identifying word similarity is more apparent for some natural language processing tasks such as summarizing, information extraction, information retrieval and question answering. In those tasks, the search for information through texts demands systems to find the correct results considering the flexibility existent in natural language. This can be observed, for instance, in those situations when the texts may not contain the same words used in the search query but yet contain similar words as synonyms. This can also be observed in the opposite situation when the system needs to deal with polysemic words (Jurafsky and Martin, 2009).

Better techniques for identifying word similarity can help in many NLP tasks such as dialogue systems, question answering and information retrieval systems (Agirre *et al*., 2009; Pilehvar *et al*., 2013; Oliveira, 2018). Historically, most of the Question Answering (QA) and Information Extraction (IE) systems use lexical resources such as WordNet to search for synonyms (Jurafsky and Martin, 2009; De Araujo *et al*., 2018). However, the expansion of terms using lexical knowledge base resources, such as WordNet, presents some problems. Since some of these lexical bases are manually constructed, they are time-consuming and expensive. For this reason, eventually not all desired words will be present and the knowledge base quality varies from language to language (Leeuwenberg *et al*., 2016).

In the following statement, Jurafsky and Martin (2009) comment on WordNet aspects:

> [...] The previous section showed how to compute the similarity between any two senses in a thesaurus and by extension between any two words in the thesaurus hierarchy. But of course, we don't have such thesauri for every language. Even for languages where we do have such resources, thesaurus-based methods have a number of limitations. The obvious limitation is that thesauri often lack words, especially new or domain-specific words. In addition, thesaurus-based methods only work if rich hyponymy knowledge is present in the thesaurus. While we have this for nouns, hyponym information for verbs tends to be much sparser and doesn't exist at all for adjectives and adverbs. Finally, it is more difficult with thesaurus-based methods to compare words in different hierarchies, such as nouns with verbs.

These aspects motivate the use of distributional-based models (Oliveira, 2018) as complementary resources to the lexical knowledge-base approach. The distributional models have proven to be more competitive than the Lexical approach and have been successfully used to cover out-of-vocabulary items (Agirre *et al*., 2009; Oliveira, 2018). In order to do so, one possibility is the use of Word embeddings (Hartmann *et al*., 2017; Young *et al*., 2017), which follows a distributional approach and therefore does not depend on manual construction. They also can be applied to different languages since its training is unsupervised.

Nevertheless, the distributional models present some restrictions, regarding the fact that structural relations present in the sentences are not included in these models (Barcelos and Rigo, 2018; Ye and Zhao, 2018). Situations such as polysemic words, in which the meaning is associated with the word context, are not modeled directly. The traditional bag-of-words approach restricts the results of the distributional models to the words co-occurrence, with limitations to other possible semantic contexts present in the sentences (Levy and Goldberg, 2014). The development of distributional models which includes additional contextual information can bring advantages to this area, although these models are still scarcely explored (Levy and Goldberg, 2014; Komninos and Manandhar, 2016; Barcelos and Rigo, 2018; Ye and Zhao, 2018).

Therefore, this work studies the advantages of incorporating syntactic information in the distributional models, fostering for better results in semantic similarity approaches. Initially, the differences in results for word similarity accuracy and recall through the use of lexical resources and word embeddings were evaluated. After that, the results were compared against the obtained by word similarity based on the use of distributional models enriched with syntactic contexts, using the DEPS (dependency-based syntactic contexts) word embedding model proposed by Levy and Goldberg (2014). For this, the following specific objectives are highlighted. Initially, to explore the existing techniques regarding word similarity, using the distributional approach by adapting existing works to Brazilian Portuguese. After this, compare the distributional approach to other techniques that are solely based on lexical databases. Then complement this comparison with the addition of syntactic context in the training process of word embeddings to a Brazilian Portuguese corpus. Finally, to evaluate all the different techniques over a common dataset.

The contributions presented in this article are the following:

1. The comparison between different approaches do detect word similarity using the Lexical and the distributional resources, adapting existing works to Brazilian Portuguese
2. The addition of syntactic context in the training process of word embeddings to a Brazilian Portuguese corpus, comparing the results obtained with this resource and the results obtained with standard Lexical and Distributional resources. Then, extending the studies of Levy and Goldberg (2014)
3. All the resources used or created from the development of this work were made available on the web, as well as the source code containing the procedures performed to obtain the results achieved in this study. The files are at Github[1]

This article is structured as follows. In Section 0.1 the related works are described and analyzed the works related to the research area of this work. Section 1 presents the methodological aspects of the proposed model, as well as the form of the experiment and the necessary tools. Next, Section 2 presents the results obtained in the case study experiment. Finally, Section 3 summarizes the findings, contributions and discussions.

Related Work

This section describes the related works and analyzes their aspects related to the current work. In particular, the analysis addresses experiments introducing contextual information in distributional models.

A new dataset for evaluating Distributional Similarity Models in Portuguese is presented by Granada *et al*. (2014). For this, they translated the word pairs from the well-known baseline for semantic relatedness evaluation in English called RG65 created by Rubenstein and Goodenough (1965). The original dataset contains judgments from 51 human subjects for 65 word pairs. To

[1] Available at https://github.com/eberlitz/word-embeddings-docker-image.

generate the PT65 they translated all the word-pairs and evaluated them with 50 human subjects. They compared the human scores with previous works and also performed a qualitative evaluation using Latent Semantic Analysis (LSA) models generated from Wikipedia articles. The correlation scores obtained were close to the scores achieved by other works that targeted another language. With the experiment, they observed that the semantic similarity can be transferred across languages, but for Portuguese, a manual evaluation had better results.

Levy and Goldberg (2014) presents a generalized skip-gram model with negative sampling introduced by Mikolov *et al*. (2013), from a linear context of bag-of-words to arbitrary word contexts, specifically syntactic contexts. An interesting fact of this approach in comparison with the original work is that the concept of induced similarity represents a nature of cohyponym. They also describe a way of performing an analysis of the representation learned in the vector space by exploring the contexts of specific words or a group of words. They used the English Wikipedia as a corpus to train the embeddings. This corpus was tagged with Parts-Of-Speech (POS) using the Stanford tagger. For the evaluation, they manually inspected the five most similar words to a hand-picked set of words. One remarkable example is the word "Hogwarts" that in the Bag of Words (BoW) model the most similar words are from the respective domain of Harry Potter and in the developed model it was a list of famous schools, that is, was able to capture the semantic type of the word. The model was evaluated against the WordSim353 dataset from Finkelstein *et al*. (2001), which is a dataset regarding word similarity versus relatedness. They draw a precision-recall curve that describes the embeddings affinity, proving that the results obtained by the developed model were slightly better than the BoW model.

Hartmann *et al*. (2017) present an evaluation of different word embedding models trained on a large Portuguese corpus (Brazilian and European variants together) on syntactic and semantic analogies, POS tagging and sentence semantic similarity tasks. They collected a large corpus from various sources, either from Brazilian or European Portuguese. With that they applied some pre-processing (tokenization and normalization) in order to reduce the vocabulary size. Using the corpus as input, they trained word embedding models using four different algorithms (Word2Vec, Wang2Vec, FastText and GloVe) with varying dimensions (50, 100, 300, 600 and 1000). For the evaluation, first, they used the syntactic and semantic analogies provided by Rodrigues *et al*. (2016), where the FastText model performed better for syntactic analogies. For semantic analogies, GloVe had the best performance. Also, all Continuous Bag Of Words (CBOW) models obtained through word embeddings algorithms, had poor results in semantic analogies (except Wang2Vec). For the POS tagging task evaluation, the Wang2Vec had the best results and higher dimensions had better performance. The worst models in this task were GloVe and FastText. For the sentence semantic similarity task evaluation, they used the ASSIN dataset. With this, they had Word2Vec CBOW model with 1000 dimensions as the best one for European Portuguese. Moreover, for Brazilian Portuguese, the Wang2Vec Skip-Gram model with 1000 dimensions had the best scores. They suggest that word analogies are not very suitable for evaluating word embeddings and task-specific is probably a better approach.

In this article, Agirre *et al*. (2009) compares two categories of techniques used to measure semantic similarity, implementing graph-based algorithms to WordNet and distributional similarities collected from a 1.6 Terabyte Web corpus. The article also describes an integration of the two techniques. For the graph-based algorithm they represent WordNet version 3.0 as a graph, where the relations among synsets are undirected edges and for this graph, they compute the PageRank for each of the words in the corpus, producing a probability distribution over the synsets. Then, this is encoded as vectors by computing the cosine between them. In this work, two WordNet versions were used, the WordNet 3.0 and the Multilingual Central Repository (MCR) aiming to link words between multiple WordNet languages. For cross-linguality, they exchange each non-English word in the dataset with its five best translations into English and then create the vector with the calculated similarities. For the distributional approach of calculating similarities between words the explore the use of a vector space model using three variations as bag-of-words, context-window and syntactic-dependency over a corpus of four billion documents crawled from the web in August 2008. They evaluate all the approaches over two standard datasets (RG65 and WordSim353) and also test a combination of both approaches (WordNet and Distributional) by training an SVM classifier to select the best result of the treedistributional variations for each pair. Thus, achieving state-of-the-art distributional and WordNet-based similarity measures over this datasets.

Peters *et al*. (2018) presents in this work a general approach for learning context-dependent representations from Bidirectional Language Models (biLMs), which were called Embeddings from Language Models (ELMo). Instead of learning a word as a vector representation, the ELMo intend to catch the context of a word as a vector representation, meaning that, it learns embeddings with the different nuances of a single word. Models like GloVe, Word2Vec, Wang2Vec and FastText would generalize all the different nuances of a single word in a single word vector having the same representation.

With the release of ELMo, it brought near state-of-the-art results in many downstream NLP tasks, including question answering, textual entailment and sentiment analysis. ELMo induced the current state-of-the-art technique called BERT, which stands for Bidirectional Encoder Representations from Transformers.

BERT, a work by Devlin *et al*. (2018), is a method of pre-training language representations. It outperforms previous methods because it is the first unsupervised, deeply bidirectional system for pre-training NLP. It uses attention transformers instead of bidirectional recursive neural networks to encode the context.

Based on these works, the WordNet was evaluated against the word embedding approach regarding the identification of word similarity using several word embedding algorithms. The evaluation used the pre-trained word embedding models presented by Hartmann *et al.* (2017) for comparison while evaluating the models under the specific task of the measurement of word similarity. Moreover, the work of Levy and Goldberg (2014) was extended to generate a word embedding with syntactic contexts from a Brazilian Portuguese corpus, to check if the results were similar or not regarding other models, but instead of using the WordSim353 which is for English, another dataset that can be considered a gold standard for the Portuguese language was used. Because of the results presented by Granada *et al*. (2014), current work ended up using the PT65 as a gold-standard for evaluating semantic similarity and relatedness between words with word embedding models and the WordNet. Also, as the works from Peters *et al.* (2018) and Devlin *et al.* (2018), represent the state-of-the-art evolution from the first word embedding models and allow pre-trained models to be used for general purpose NLP tasks, the current work intends to explore how these language models behave for word-level tasks such as word similarity.

Materials and Methods

This section presents a description of the adopted approach, as well as the tools and methods used and present an overview of the methodology and how the proposed experiments were performed. Next, the section describes the dataset used in the evaluation process, the Corpus generation and how the most common word embedding models could be obtained. Finally, an explanation of the steps involved on the reproduction the work of Levy and Goldberg (2014) for the Portuguese language was made.

Methodology Overview

This work consists of a methodology to compare different word similarity techniques. Most importantly, the work aims to identify possible outcomes of the joint use of syntactic information with distributional models. Therefore, Fig. 1 defines an overview of the methodology with the intention of comparing several techniques using different algorithms and testing them with a common dataset. In the Figure 1 the elements marked with an asterisk are resources and the tools that were generated by the authors. The other elements are existing components or resources that were used in the methodology.

The proposed methodology compare techniques based on the two main approaches to word similarity: the knowledge-based and the distributional-based. The knowledge-based approach used the Open Multilingual Wordnet (OMW) due to its relevance as a widely known and applied resource. Other motivating aspects for the adoption of OMW in the experiments were its ease of use through the Natural Language Toolkit (NLTK) library available for the Python version 3.6 programming language as well as the availability of the Portuguese language for querying the synsets (Bond and Foster, 2013). The results were obtained with Path Distance and Wu-Palmer similarity techniques (Biggins *et al*., 2012), (Wu and Palmer, 1994).

The distributional approach used generated word embedding models with a corpus obtained from the Brazilian Portuguese Wikipedia dump. The word embeddings were generated using several different model implementations for learning word representations. In this case, the following models were used: FastText (Bojanowski *et al*., 2017), Wang2vec (Ling *et al*., 2015), Word2vec (Mikolov *et al*., 2013) and GloVe (Pennington *et al*., 2014). Also, the obtained models were compared against a set of pre-trained models available from Núcleo Interinstitucional de Linguística Computacional (NILC)[2] in all different implementations (FastText, Wang2vec, Word2vec and GloVe). The cosine distance was the metric used for the comparison of word similarity for all word embedding models. The CBOW and Skip-gram were used for the models that provide this option (Mikolov *et al*., 2013; Pennington *et al*., 2014; Ling *et al*., 2015; Bojanowski *et al*., 2017; Hartmann *et al*., 2017).

The current work generated one additional model to take into account the syntactic tree information of the Portuguese corpus, which were obtained through the PALAVRAS (Bick, 2000). This model used the Levy and Goldberg (2014) algorithm to generates the DEPS model. After the model generation, a quantitative evaluation of all models and techniques was made using the PT65 dataset, which consists of a pair of words and a similarity value given by language specialists (Granada *et al*., 2014).

All the experiments were done using the Semantics server (Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz with 32 cores and 128GB of RAM) granted by the UNISINOS University Applied Computing Graduate Program by running Docker containers.

[2] http://www.nilc.icmc.usp.br/nilc/index.php

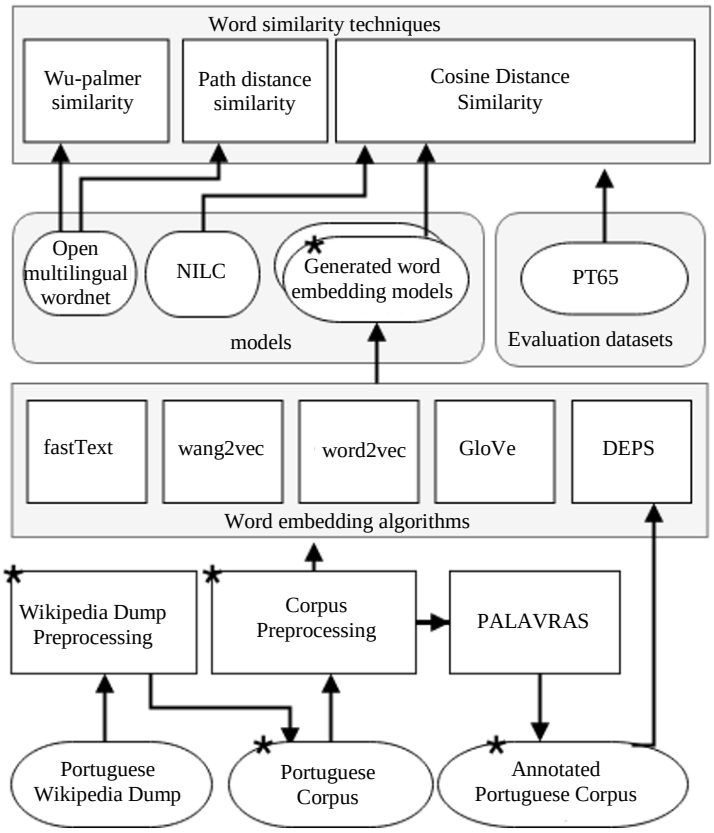

Fig. 1: Proposed methodology; the elements marked with an asterisk in the image are resources/tools that were generated by the authors

PT65 Dataset

The PT65 dataset is composed of 65 word pairs, initially generated by Rubenstein and Goodenough (1965) on the name of RG65. These word pairs wore translated to Portuguese by Granada *et al*. (2014) and evaluated with 50 language specialists.

The initial idea for the experiments was to use the WordSimilarity-353 Test Collection developed by Finkelstein *et al*. (2001) which consists of two sets of English word pairs along with human-assigned similarity judgments. However, the sentences were transled to Portuguese and then the human-assigned similarity judgments would not fit entirely, because the semantic changes involved in the translation process. So the PT65 dataset was considered more adequate and used in the evaluation process.

Corpus Generation

This subsection present the process involved in generating the corpus that was used on this article. It was obtained from a Wikipedia dump. While the current work focus on the Portuguese language, this could also be done for the other available languages in Wikipedia.

Getting the Wikipedia PT-BR Dump

The latest Portuguese Wikipedia articles dump is a big, compressed XML file that contains all articles in the wikitext format, but with some special tokens that deal with some specific Wikipedia features. For example, the file contains this kind of format: a) Original Portuguese version *"[[Imagem:Starsinthesky.jpg|thumb| [[Estrela | Formação estrelar]] na [[Grande Nuvem de Magalhães]], uma [[galáxia irregular]].]]"; b) English version (authors translation): "[[Image: Starsinthesky.jpg |thumb|[[Star | Star Formation]] in the [[Great Magalhães cloud]], a [[irregular galaxy]].]]"* More detailed information about the dump formats and different languages can be found in their website[3].

Pre-Processing with Wikiextractor

As described in the previous step, the format of the dump is not suitable for most of NLP tasks. Then, there is a need to parse the wiki text format to transform it in raw text. In order to do this, the current work used the wikiextractor project that just reads the XML file and outputs all the documents in parsed text.

[3] Available at https://en.wikipedia.org/wiki/Wikipedia:Database_download.

DOI: 10.3844/jcssp.2019.1378.1389

```
<doc id="220" url="https://pt.wikipedia.org/wiki?curid=220"
title="Astronomia">
Astronomia
Astronomia é uma ciência natural que estuda corpos celestes (como
estrelas, planetas, cometas, nebulosas, aglomerados de estrelas,
galáxias) e fenômenos que se originam fora da atmosfera da Terra (como
a radiação cósmica de fundo em micro-ondas). Preocupada com a evolução,
a física, a química e o movimento de objetos celestes, bem como a
formação e o desenvolvimento do universo.
...
</doc>
```

Fig. 2: WikiExtractor output sample

*English Translation (authors version): Astronomy. Astronomy is a natural science that studies celestial bodies (such as stars, planets, comets, nebulae, clusters, stars, galaxies) and phenomena that originate outside the Earth's atmosphere (like cosmic microwave background radiation). Worried about evolution, physics, chemistry and the movement of celestial objects, as well as the formation and development of the universe

Wikipedia has a concept of Templates, which consists of using other documents inside of a given one. For the objective of this corpus, it is not desired that the tool expand these templates because it will just add duplicated sentences to the content. So, it is really important to use the-no-templates flag. This tool generated multiple compressed 10MB files of wiki articles sentences as seen in Fig. 2.

It is also possible to save this contents as only one text file just by changing the tool arguments. At the time of writing, there were 1.000.400 documents available in the ptwiki-dump.

Custom Pre-Processing

The cleanup of the sentences for generating the Word embedding models used a custom pre-processing[4] based on Hartmann *et al.* (2017) pre-processing scripts. Some changes were made to do some necessary cleaning, as follows:

- Breaks an entire document into multiple sentences using the nltk.data.load ('tokenizers/punkt/portuguese.pickle'). Natural Language Toolkit - NLTK is a leading platform for building Python programs to work with human language data and it has a sentence segmentation tool called punkt
- Does not change the current letter case. Later the Syntactic parser that obtained better accuracy with this choice
- Remove sentences with less than 4 tokens. As it does not add meaningful value to the corpus, this sentences can be removed
- Allow abbreviations, like 'Dr.'
- Keep words with '-', like 'guarda-chuva' (which means umbrella in English)
- All emails are mapped to EMAIL token
- All numbers are mapped to 0 token
- All URLs are mapped to URL token
- Different quotes are standardized.
- Different kinds of hyphenation are standardized
- HTML strings are removed
- All text between brackets is removed

With this, a file of 1.6GB PT-BR for the corpus was obtained. It contains the number of 9.896.520 sentences, 251.193.592 tokens and 3.137.040 unique tokens.

PALAVRAS Annotated Corpus Generation

The current work used PALAVRAS (Bick, 2000) software to annotate all sentences of the corpus with syntactic tags, which is an automatic parser for Portuguese, broadly known by its high accuracy in the results.

In order to deal with the massive volume of text in the corpus, some proceedings were necessary. At first, the parser was used with multiple sentence files of 1MB. However, it was taking too long to execute and sometimes throws errors. So a Python script was created to send the sentences in batches to the PALAVRAS and save the results. Also, parallel computing was used spreading the process according to the number of cores on the computer. First, it was executed on an i5 2.4GHz computer with four cores, achieving an average speed of 16 sentences per second, which means that for all 9.896.520 sentences it would take seven days to complete. Although the use of other techniques to attempt to increase the speed, the bottleneck was indeed in the parser tool.

With this problem at hand, UNISINOS University Applied Computing Graduate Program granted us access to the Semantics server (Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz with 32 cores and 128GB of RAM).

[4]https://github.com/eberlitz/pt-br-word-embeddings/blob/master/scripts/preprocess.py

Using the available 32 cores the parsing step should be concluded in 24 hours.

On the processing step, the log files noticed that the PALAVRAS parser could not parse some of the sentences. Since the application was running the parser in batches, this means that if one sentence failed, then it lost all the parsed sentences in the batch. Also, as this process would take too long, it needs another implementation to continue the process if some fatal error occurred. With this in mind, the sentences were converted to an SQLite table with three columns (id, text and parsed text). With this whenever the application starts the parsing process, it can continue from where it stopped.

Once the application development finished, a Docker image was created and started running in the Semantics server. The overall process took 38 h, which was divided in 24 h to process 8.916.000 sentences using batches of 30 and 14 h to process the remaining ones without sending in batches. Resulting in a 15GB corpus file.

Common Word Embeddings Generation

In order to have a base for comparison, all of Hartmann *et al*. (2017) generated models were used. In this case, FastText, Wang2vec, Word2vec and GloVe usin different dimensions values like 50, 100, 300, 600 and 1000 (Mikolov *et al*., 2013; Pennington *et al*., 2014; Ling *et al*., 2015; Bojanowski *et al*., 2017). Also, the CBOW and Skip-gram were used for the models that have this option.

All of the model generation tools were downloaded and compiled into a docker image, which uses as input the generated corpus text file.

DEPS Word Embedding Generation

In order to generate the DEPS (dependency-based syntactic contexts) word embedding model proposed by Levy and Goldberg (2014), the source code called word2vecf[5] was used. The input required by this model is three files, a word, context vocabulary and a contexts file.

The vocabulary files are just a list of words or contexts with the total number of occurrences. The contexts file is a multiline context per word, where one word can have multiple contexts. In the form of *"<word> <dependencyrelation>_<referred-word>"*.

To generate the file from the used corpus, the PALAVRAS output was used to extracted the syntactic tags. This process is shown by in Fig. 3.

The generation of these three files was not trivial and the implemented code to do so had to use a map-reduce approach in order to use all computation resources available. It took 14 hours to process all 9.896.000 parsed sentences with an average speed of 196, 2 sentences per second. After, the resulting files were used as input for word2vecf tool, which took some hours to complete. The models were generated with different dimensions values as 50, 100, 300, 600 and 1000.

[5] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

Results and Discussion

The evaluation step was divided on three steps, described in this section. Two quantitative evaluations and one qualitative evaluation were made. First, the quantitative evaluation of the Open

Multilingual WordNet was developed by using the WordSim353 dataset (Finkelstein *et al*., 2001), which is a dataset regarding word similarity and relatedness. As all the pairs in this dataset are in English, they were manually translated to Portuguese to be able to compare the results with all the techniques. Then, the same evaluation was made with the word embeddings models. At last, a qualitative evaluation regarding the DEPS model was presented.

For the Portuguese language DEPS Model evaluation, a qualitative measure based on a manual inspection of the most similar words to a hand picked set of words and see if they are in fact similar or not to the other words regarding similarity, relatedness and also the syntactic similarity.

To better visualize the performance of each technique the precision-recall curve draws was used to describes the embeddings affinity.

Open Multilingual WordNet Evaluation

The quantitative evaluation of the knowledge-based approach for word similarity used the Open Multilingual Wordnet (OMW) (Bond and Foster, 2013), which was loaded with the Natural Language Toolkit (NLTK) library. Then, the resources was used to calculate the similarity between the pair of words from the PT65 dataset using two distances: Path Distance and Wu-Palmer. The both distances allow to measure the Pearson's Correlation (ρ) for each of the techniques. Table 1 shows the results.

The Path Distance algorithm gave a relatively high score, but we noticed the occurrence of some out of vocabulary words Jurafsky and Martin (2009), in this case, a number of 15.38% of the words.

Table 1: OMW evaluation on PT65

Algorithms	*r*	*ρ*	Out of Vocabulary ratio
Path Distance	**0.76**	0.67	15.38
Wu-Palmer	0.62	0.51	15.38

*r is the Pearson's Correlation considering only the words in vocabulary;

*ρ is the Pearson's Correlation considering all the words, given a similarity value of zero for words out of vocabulary;

*The higher value is in bold for better readability.

Word Embeddings Evaluation

To do a quantitative evaluation of the distributional approach for word similarity, the same experiments as the WordNet evaluation were made but with the embeddings models. For each word pair of PT65 dataset, the expected result with the Cosine similarity given by the model were compared. As the previous evaluation, this also allows the measurement of Pearson's Correlation (ρ). Table 2 shows the results for all the 40 generated models.

There was no out of vocabulary words in this approach which in comparison with the WordNet approach is better. Just like mentioned by Agirre *et al*. (2009), that the word embeddings can cover out-of-vocabulary words. In comparison with the WordNet, it has slightly better results, which is a good thing considering that it has no manual construction as WordNet.

As we can see, the better word embedding model for this task is the FastText Skip-Gram. In all of them, Skip-gram was slightly better than the others. Moreover, in overall the models with 300-600 dimensions got higher values. We can also note that the DEPS model has an inferior performance in this particular task, probably because the dataset does not differentiate between relatedness and similarity. We also repeated the same experiment with the pre-trained models by Hartmann *et al*. (2017).

Qualitative Evaluation of DEPS Model

A qualitative evaluation of DEPS model was made, which had a manually inspect of the five most similar words (by cosine similarity) to a given set of target words (Table 3). Also, the DEPS was compared against other models, just like Levy and Goldberg (2014) did in their experiment.

Table 2: Word embeddings evaluation on PT65

Embedding models		Size	ρ(*ours*)	ρ(*nilc*)
FastText	CBOW	50	0.67	0.63
		100	0.72	0.67
		300	**0.75**	0.73
		600	0.73	0.74
		1000	0.71	0.74
	Skip-Gram	50	0.74	0.64
		100	**0.77**	0.73
		300	**0.79**	**0.78**
		600	**0.77**	**0.76**
		1000	0.72	0.74
Wang2vec	CBOW	50	0.57	0.59
		100	0.61	0.69
		300	0.69	0.74
		600	0.69	0.66
		1000	0.68	0.65
	Skip-Gram	50	0.65	0.60
		100	**0.74**	0.70
		300	**0.75**	**0.77**
		600	0.72	**0.76**
		1000	0.69	0.71
Word2vec	CBOW	50	0.58	0.34
		100	0.63	0.43
		300	0.68	0.58
		600	0.69	0.62
		1000	0.68	0.61
	Skip-Gram	50	0.65	0.48
		100	**0.75**	0.54
		300	**0.76**	0.64
		600	0.74	0.68
		1000	0.69	0.67
GloVe		50	0.63	0.63
		100	0.69	0.71
		300	0.69	0.72
		600	0.67	0.71
		1000	0.65	0.68
DEPS		50	0.47	-
		100	0.44	-
		300	0.43	-
		600	0.45	-
		1000	0.44	-

*r(*ours*) is the Pearson's Correlation value from current work trained models
*ρ(*nilc*) is the Pearson's Correlation values from the NILC pre-trained models
*All values equals or greater than 0.75 are in bold for better readability

Table 3: Excerpt of most similar words per word embedding models

Target word	DEPS	FastText	Wang2vec	Word2vec	GloVe
longe	perto	próximo	perto	distante	perto
	lá	distante	distante	fora	fora
	abaixo	afastado-se	afastada	perto	ficar
	cá	afastada	fora	afastada	lá
	debaixo	afastados	distantes	tirá-los	tão
guarda-chuva	caneta	guarda-chuvas	lenço	galhardete	égide
	tampão	manda-chuva	sobre-tudo	xale	cinza
	espingarda	manda-chuvas	quepe	chapeu	casaco
	cetro	guarda-chaves	moletom	abajur	disfarce
	carregador	guarda-copos	pulôver	paletó	crachá
correr	viajar	correndo	correndo	correndo	caminhar
	aceitar	correrem	caminhar	correu	nadar
	ganhar	correu	pedalando	caminhar	correndo
	aprender	correra	correu	pedalando	saltar
	realizar	correria	agachar-se	pular	pular
inglês	espanhol	inglêss	ingles	ingles	português
	francês	inglêz	português	espanhol	francês
	norueguês	inglês-the	espanhol	francês	inglesa
	sueco	inglêsa	francês	português	espanhol
	italiano	francês-inglês	galês	irlandês	britânico
faculdade	universidade	faculda	universidade	universidade	
	universidade				
	escola	faculdad	bacharelando-se	histórico-filosóficas	medicina
	liceu	ex-faculdade	politécnica	bacharelando-se	ciências
	conservatório	universidade	pós-graduação	politécnica	curso
	colégio	faculde	puc	puc-pr	usp

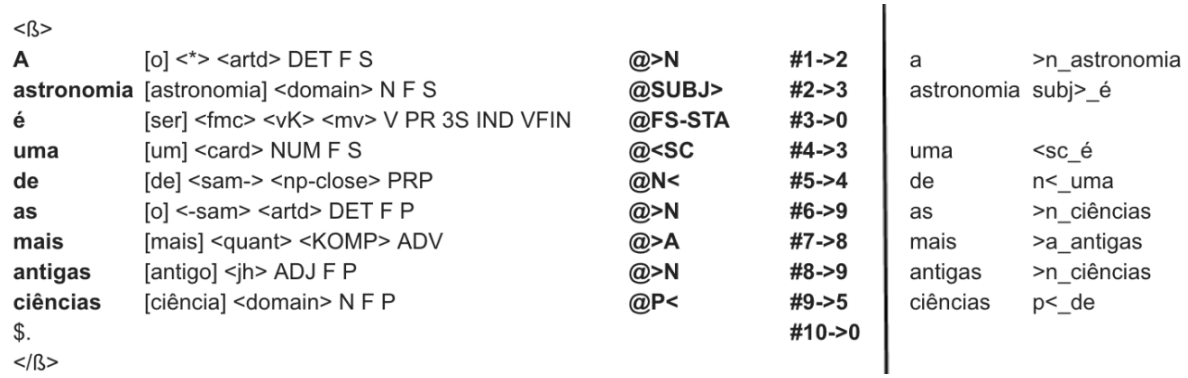

Fig. 3: DEPS contexts generation for a single parsed sentence
*Left: Sample output from the PALAVRAS parser for the sentence "A astronomia é uma das mais antigas ciências." (Astronomy is one of the oldest sciences). Right: Sample of authors generated contexts file for an annotated sentence

The first target word, *longe* (Far), have similar results by all the different models. The word *inglês* (English) have the same behavior. However, for some specific words, like *faculdade* (College), the DEPS model returned other types of languages or colleges while the other models could just bring words related to the same domain. This is similar to the target word *Hogwarts* from the Levy and Goldberg (2014) work.

The other two words, *guarda-chuva* (Umbrella) and correr (Run), demonstrates that the DEPS model find other words with the same syntactic function (verb and noun) like a classifier, which in terms of semantic similarity or relatedness is not so good, just as we saw in the qualitative experiment (Table 2) where the DEPS model had the worst results in that particular task for Portuguese.

Conclusion

The detection of similarity between words is a vital topic for NLP tasks such as summarization, information retrieval and question answering, among others (Islam *et al*., 2008; Agirre *et al*., 2009; Pilehvar *et al*., 2013). This motivated the study carried out and related in this paper.

Some identified evidences in the literature indicates that the expansion of terms using lexical resources such as WordNet has several problems since some of these lexical bases are manually constructed, time-consuming and expensive. For this reason, not all desired words will be available and their quality varies from language to language (Leeuwenberg *et al*., 2016). Distributional approaches regarding word similarity proved to be more competitive than the thesaurus-based approach and have been successfully used to cover out-of-vocabulary items in lexical resources (Agirre *et al*., 2009; Barcelos and Rigo, 2018; Oliveira, 2018). Nevertheless, limitations are also observed in Distributional approaches, mainly due to the lack of contextual information in this models (Barcelos and Rigo, 2018; Ye and Zhao, 2018).

This article describes the investigation of techniques regarding word similarity, using a distributional approach (word embeddings), adapting existing works to Brazilian Portuguese. Also, experiments with other techniques that are solely based on a lexical database such as WordNet were made and evaluated over a common dataset called PT65. The results showed that word embeddings can cover words out of vocabulary and have slightly better results in comparison with WordNet in this particular task.

As main contribution, we adapted the studies of Levy and Goldberg (2014) regarding the addition of syntactic context in the training process of word embeddings to a Brazilian Portuguese corpus, finding similar results for the task of word similarity against the dataset PT65.

As a limitation, this work did not compare differences between similarity and relatedness since the dataset does not specifically distinguish between them. Querido *et al*. (2017), have recently adapted the SimLex-999 and WordSim-353 datasets to Portuguese so it's possible to do an evaluation comparing the performance between similarity and relatedness. Also, this work only used the Brazilian Portuguese corpus from the Wikipedia, but accordingly to Fonseca and Aluísio (2016) the bigger the corpus is, the better the embeddings, even with mixed Portuguese variants, so it could also be possible to evaluate with a much bigger corpus by joining the European with the Portuguese Wikipedia dumps.

In future works, the authors intend to evaluate the different techniques against the work of De Araujo *et al*. (2018), in a real situation of support for a Question-Answering system regarding word synonymy identification. Also, future approaches could explore the use of BERT and ELMo language models for word-level for Brazilian Portuguese word similarity.

Acknowledgment

This research paper was made possible through the help and support from PhD Sandro José Rigo. First and foremost, thanks for his most support and encouragement.

Author's Contributions

Eduardo E. Berlitz: He participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Denis A. Araujo: He participated in all experiments and contributed to the writing of the manuscript.

Rodrigo R. Righi: He contributed reviewing the article critically for significant intellectual content.

Allan B. Silva: He contributed to the writing of the manuscript and contributed reviewing the article critically.

Sandro J. Rigo: Designed the research plan and organized the study. He also contributed in drafting the article and reviewing it critically for significant intellectual content. Also, he gave the final approval of the version to be submitted.

Ethics

We testify that this research paper submitted to the Journal of Computer Science has not been published elsewhere and that has no ethical issues. All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

References

Agirre, E., E. Alfonseca, K. Hall, J. Kravalova and M. Pa¸sca *et al*., 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. Proceedings of the Human Language Technologies: The Annual Conference of the North American Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, (NAACL '09), PA, USA, pp: 19-27.

Barcelos, A. and S.J. Rigo, 2018. Enhancing brazilian portuguese textual entailment recognition with a hybrid approach. J. Comput. Sci., 14: 945-956. DOI: 10.3844/OFSP.12054

Bick, E., 2000. The parsing system "palavras". Ph.D. Thesis, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Århus.

Biggins, S., S. Mohammed, S. Oakley, L. Stringer and M. Stevenson *et al*., 2012. Two approaches to semantic text similarity. Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SEM), Association for Computational Linguistics, (ACL' 12), pp: 655-661

Bojanowski, P., E. Grave, A. Joulin and T. Mikolov, 2017. Enriching word vectors with subword information. Transact. Associat. Computat. Linguistics, 5: 135-146. DOI: 10.1162/tacl_a_00051

Bond, F. and R. Foster, 2013. Linking and extending an open multilingual wordnet. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Aug. 4-9, Sofia, Bulgaria, pp: 1352-1362.

De Araujo, D., A. Hentges and S. Rigo, 2018. Uma abordagem linguística para sistemas de perguntas e respostas curtas. In: Simpósio Brasileiro de Sistemas de Informação, Caxias do Sul/RS, SBSI.

Devlin, J, M.W. Chang, K. Lee and K. Toutanova, 2018 Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin and Z. Solan *et al*., 2001. Placing search in context: The concept revisited. Proceedings of the 10th International Conference on World Wide Web, (WWW '01), ACM, New York, NY, USA, pp: 406-414.
DOI 10.1145/371920.372094

Fonseca, E.R. and S.M. Aluísio, 2016. Improving Pos Tagging Across Portuguese Variants with Word Embeddings. In: Computational Processing of the Portuguese Language, Silva, J.R., R. Ribeiro, P. Quaresma, A. Adami and A. Branco (Eds.), Springer, Cham, ISBN-10: 978-3-319-41552-9, pp: 227-232.

Granada, R., C.T. Dos Santos and R. Vieira, 2014. Comparing semantic relatedness between word pairs in portuguese using Wikipedia. In: Computational Processing of the Portuguese Language, Baptista, J., N.J. Mamede, S. Candeias, I. Paraboni and T.A.S. Pardo *et al*. (Eds.), Springer.
ISBN-10: 978-3-319-09761-9, pp: 170-175.

Hartmann, N., E.R. Fonseca, C. Shulby, M.V. Treviso and J. Silva *et al*., 2017. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: STIL, Sociedade Brasileira de Computação, Paetzold, G.H. and V. Pinheiro (Eds.), pp: 122-131.

Islam, A., D.Z. Inkpen and I. Kiringa, 2008. Applications of corpus-based semantic similarity and word segmentation to database schema matching. VLDB J., 17: 1293-1320. DOI: 10.1007/s00778-007-0067-9

Jurafsky, D. and J.H. Martin, 2009. Speech and Language Processing. 2nd Edn., Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Komninos, A. and S. Manandhar, 2016. Dependency based embeddings for sentence classification tasks. Proceedings of NAACL-HLT the Association for Computational Linguistics, Jun. 12-17, San Diego, California, pp: 1490-1500. DOI: 10.18653/v1/N16-1175

Leeuwenberg, A., M. Vela, J. Dehdari and J. Genabithb, 2016. A minimally supervised approach for synonym extraction with word embeddings. Prague Bull. Mathematic. Linguist., 105: 111-142.

Levy, O. and Y. Goldberg, 2014. Dependency-based word embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics, (ACL' 14), Baltimore, Maryland, pp: 302-308.

Ling, W., C. Dyer, A. Black and I. Trancoso, 2015. Two/too simple adaptations of word2vec for syntax problems. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, (ACL' 15).

Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado and J. Dean, 2013. Distributed representations of words and phrases and their compositionality. Neural Information Processing Systems Foundation, Inc.

Oliveira, H.G., 2018. Distributional and knowledge-based approaches for computing Portuguese word similarity. Information, 9: 35-35.

Pawar, A. and V. Mago, 2018. Calculating the similarity between words and sentences using a lexical database and corpus statistics. CoRR abs/1802.05667

Pennington, J., R. Socher and C. Manning, 2014. Glove: Global vectors for word representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (NLP' 14).

Peters, M.E., M. Neumann, M. Iyyer, M. Gardner and C. Clark *et al*., 2018. Deep Contextualized Word Representations. In: NAACL-HLT, Association for Computational Linguistics, Walker, M.A., H. Ji, A. Stent (Eds.), pp: 2227-2237.

Pilehvar, M.T., D. Jurgens and R. Navigli, 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. Proceedings of the Association for Computer Linguistics, (ACL' 13), ACL.

Querido, A., R. De Carvalho, J. Rodrigues, M. Garcia and J. Silva *et al*., 2017. Lx-lr4distsemeval: A collection of language resources for the evaluation of distributional semantic models of portuguese. Revista Associação Portuguesa Linguística 3: 265-283.
DOI: 10.26334/2183

Rodrigues, J.A., A. Branco, S. Neale and J.R. Silva, 2016. Lx-dsemvectors: Distributional Semantics Models for Portuguese. In: Lecture Notes in Computer Science, Silva, J.R., R. Ribeiro, P. Quaresma, A. Adami and A. Branco (Eds.), Springer, pp: 259-270

Rubenstein, H. and J.B. Goodenough, 1965. Contextual correlates of synonymy. Commun. ACM, 8: 627-633.

Sravanthi, P. and B. Srinivasu, 2017. Semantic similarity between sentences. Int. Res. J. Eng. Technol., 4: 156-161.

Wu, Z. and M. Palmer, 1994. Verb semantics and lexical selection. Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, Association for Computational Linguistics, (ACL' 94).

Ye, Z. and H. Zhao, 2018. Syntactic word embedding based on dependency syntax and polysemous analysis. Frontiers IT EE, 19: 524-535

Young, T., D. Hazarika, S. Poria and E. Cambria, 2017. Recent trends in deep learning based natural language processing. CoRR. abs/1708.02709