UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

EDUARDO EIDELWEIN BERLITZ

USING WORD EMBEDDINGS FOR WORD SIMILARITY IN BRAZILIAN PORTUGUESE

São Leopoldo
2018

Eduardo Eidelwein Berlitz

USING WORD EMBEDDINGS FOR WORD SIMILARITY IN BRAZILIAN PORTUGUESE

Artigo apresentado como requisito parcial para obtenção do título de Bacharel em Ciência da Computação pelo Curso de Ciência da Computação da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientador: Prof. Dr. Sandro José Rigo
Coorientador: Prof. PhD. Rodrigo da Rosa Righi

São Leopoldo
2018

USING WORD EMBEDDINGS FOR WORD SIMILARITY IN BRAZILIAN PORTUGUESE

Eduardo Eidelwein Berlitz*

Sandro José Rigo**

Rodrigo da Rosa Righi***

Abstract: The ability to identify the semantic similarity between words has been a subject of research very explored in the last years, because it is related to a series of activities of the area of natural language processing like information retrieval, text summarization, categorization and generation, database schema matching, question answering, machine translation, and others. Most of question answering and information extraction systems use WordNet to search for synonyms in their search engine. However, the expansion of terms using WordNet has several problems. They are of manual construction, time-consuming and expensive, and for this reason, not all links will be present and their quality varies from language to language, as also, it is not available for all languages. Distributional-based approaches like word embeddings have successfully been used to cover out-of-vocabulary items in WordNet. Thus, with the possibility of access to different word embeddings models and the need to improve the way of expanding related terms of query systems to ontological bases used by systems of questions answering and information retrieval, the present work explores the existing techniques regarding word similarity, using a distributional approach called word embeddings, adapting existing works to Brazilian Portuguese. We also experiment with other techniques that are solely based on a lexical database such as WordNet, and we do a qualitative evaluation of all the different techniques over a common dataset called PT65. Proving that word embeddings can cover words out of vocabulary and have slightly better results in comparison with WordNet. We also adapted the studies regarding the addition of syntactic context in the training process of word embeddings to a Brazilian Portuguese corpus, finding similar results through a qualitative evaluation, but for the task of word similarity against the dataset PT65, it had the worst results.

Keywords: Word similarity. WordNet. Word embedding. Computational linguistics. Natural Language Processing.

1 INTRODUCTION

Natural Language Processing or NLP is a field whose purpose is to make computers perform tasks using human languages. In these systems, a series of components must be studied as speech recognition, natural language understanding, and speech synthesis. According to Jurafsky and Martin (2009, p. 29), "What distinguishes

* Aluno do curso de Ciência da Computação. Email: eberlitz@gmail.com
** Orientador, professor da Unisinos. Email: rigo@unisinos.br
*** Coorientador, professor da Unisinos. Email: rrrighi@unisinos.br

language processing applications from other data processing systems is their use of *knowledge of language*.". That is, for several NLP activities you need knowledge about phonetics, phonology, morphology, lexical semantics, compositional semantics. (JURAFSKY; MARTIN, 2009).

The ability to identify the semantic similarity between words has been a subject of research very explored in the last years, because it is related to a series of activities of the area of natural language processing like information retrieval, text summarization, categorization and generation, database schema matching, question answering, machine translation, and others. (ISLAM; INKPEN; KIRINGA, 2007; JURAFSKY; MARTIN, 2009).

1.1 Motivation

The motivation for this work comes from Araujo, Hentges and Rigo (2018) where they describe the ENSEPRO, which is a question answering system for short sentence questions that have it's answers based on ontologies, in their case DBPedia. In short, the system receives a user question that is processed in three main tasks, the first one is to do a natural language understanding, the second is the search engine that consumes an ontology database and finally the third task that generates the response for the use in natural language. The main focus of ENSEPRO is to tackle the Brazilian Portuguese language.

In their search engine, they currently use WordNet for term expansion which is a necessary and important step to make the system work as a whole. Araujo, Hentges and Rigo (2018, our translation) says that "[...] it is necessary to consider that the relevant terms may not be represented in the ontology with the same words of the question, being necessary to search for synonyms of the relevant term.", so for this reason, having other alternatives besides the WordNet could improve the system results.

1.2 Research problem

Most of question answering (QA) and information extraction (IE) systems use WordNet to search for synonyms in their search engine. As we can see according to Araujo, Hentges and Rigo (2018, our translation),

> [...] In the case of the use of semantic technologies, using Wordnet as a linguistic ontology, the use of this resource as a source for the semantic expansion of terms is still noticeable. One fact that draws attention to Wordnet in the construction of the conversational agents in the analyzed works is that all use it only to find synonyms of terms, and this is only one of the possibilities that this linguistic resource makes available.

However, the expansion of terms using WordNet that is a lexical base has several problems, where a word may not be present. Since these lexical bases are of manual construction, they are time-consuming and expensive, and for this reason, not all links will be present, and their quality varies from language to language. There is also no WordNet for all languages. (LEEUWENBERGA et al., 2016). Jurafsky and Martin (2009, p. 297) tell a little bit about the WordNet in the following statement,

> [...] The previous section showed how to compute 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.

So, we intend to change the thesaurus-based approach by a distributional-based one. They have proven to be more competitive than the previous approach, and have been successfully being used to cover out-of-vocabulary items in WordNet. (AGIRRE et al., 2009). In order to do so, WordNet is proposed to be replaced by Word embeddings, which follows a distributional approach and therefore does not depend on manual construction, and can be applied to different languages since its training is unsupervised. Thus, the hypothesis is that for the formulation of queries in QA and IR systems on which they depend on the expansion of similar terms it would be possible to increase the number of relevant results to be found.

The ability to identify text similarity is essential for natural language processing segments such as summarization, retrieval of information and question answering. In the search of information through texts, we often do not find the results due to the fact that the texts may not contain the same words used in the search definition, but rather similar words as synonyms. This fact makes the task of identifying similarity/synonyms between words or sentences something very important within the natural language processing area. More precise techniques for identifying word similarity can help in a number of NLP tasks such as dialogue systems, question answering, and information retrieval systems. (ISLAM; INKPEN; KIRINGA, 2007; PILEHVAR; JURGENS; NAVIGLI, 2013; AGIRRE et al., 2009)

1.3 Research focus

With the possibility of access to pre-trained word embeddings including in the Portuguese language and the need to improve the way of expanding related terms of query systems to ontological bases used by systems of questions answering and information retrieval, the present work aims to improve the accuracy and recall of these related terms expansion through the use of word embeddings. For this, the following specific objectives are highlighted:

- Explore the existing techniques regarding word similarity, using a distributional approach called word embeddings, adapting existing works to Brazilian Portuguese.
- Compare the word embeddings approach to other techniques that are solely based on a lexical database such as WordNet.
- Adapt existing studies regarding the addition of syntactic context in the training process of word embeddings to a Brazilian Portuguese corpus, to check if the results will be, similar or not regarding other word embedding models.

- Evaluate the different techniques over a common *dataset*.

1.4 Structure of the thesis

This thesis is structured as follows. The section 2 presents the general concepts and techniques used in this work. In section 3 are described and analyzed the works related to the research area of this work. The section 4 presents the proposed model, as well as the form of the experiment and the necessary tools. The section 5 presents the preliminary results obtained in the case study experiment. Finally, section 6 summarizes the thesis findings, contributions, and discusses.

2 BACKGROUND

Here, the general concepts and techniques used in this work will be presented, in order to guide the reader in a way that he clearly understands what will be addressed in the following chapters.

2.1 Natural Language Processing

Natural Language Processing or NLP is a field whose purpose is to make computers perform tasks using human languages, such as allowing human-machine communication or performing useful processing over text or speech. Within this area, we have an example, **dialogue systems** or **conversational agents** used by chatbots these days. They try to imitate a natural conversation with humans. In these systems, a series of components must be studied as **speech recognition**, **natural language understanding**, and **speech synthesis**. Another important task in NLP is **question answering** that tries to give answers to the search for the users which can be in the form of textual or spoken questions. Currently, these searches can already be answered by web search engines, while they are not yet able to relate multiple sources of information by summarizing or making inferences between them. Currently, these systems use a series of components such as **information extraction** (IE), **word sense disambiguation**, and so on. (JURAFSKY; MARTIN, 2009).

According to Jurafsky and Martin (2009, p. 29), "What distinguishes language processing applications from other data processing systems is their use of ***knowledge of language***.". That is, for several NLP activities you need knowledge about phonetics, phonology, morphology, lexical semantics, compositional semantics. (JURAFSKY; MARTIN, 2009).

2.2 Synonym

According to Zgusta and Cerny (1971, p. 89), "[...]synonyms: they are words which have different forms but identical meaning.". So we can say that synonyms can be defined as expressions with the same meaning. The definition we find in dictionaries like synonyms usually refers generally to any of the different types of synonyms, being near-synonyms and absolute-synonyms. **Near-synonyms** can be defined as expressions that are more or less similar, but not identical in meaning. Common examples in English are 'mist' and 'fog' or 'buy' and 'purchase'. **Absolute synonyms** are one or more words whose meaning is identical and can be used with the same connotation in all different contexts and are equivalently semantic. Therefore, they are extremely rare. (LYONS, 1995).

2.3 Hyponyms and hypernyms

Hyponym can be defined by the lexical relation corresponding to the insertion of one class into another. That is, it shows the relation between a generic term and a specific instance of it, where the most specific term is the Hyponym and the generic class is Hypernym. So if we say that purple is a kind of color, then purple is a hyponym of color and color is the hypernym of purple. Because of this Hypernym is normally referred to as the *is-a* and *is-a-kind-of* relation. (CRUSE; CRUSE, 1986, p. 88).

2.4 Lexical knowledge base - WordNet

WordNet is one of the lexical resources most used over the last few years when it comes to word senses. (FELLBAUM, 1998). It is a large lexical database of English which consists of four sub-nets, one each for nouns, verbs, adjectives, and adverbs. Each of this has a set of lemmas annotated with a set of synonyms. WordNet can either be downloaded for free or accessed via the web. (Princeton University, 2010).

When searching for a word in this database, we will get a list of senses, where for each one we have a set of synonyms (also called **synsets**) and a brief description (gloss) and also sometimes a simple example of use. One of the most important relations of WordNet is the set of near-synonyms for a sense called synset. For example, when searching for 'car' we get the words auto, automobile, machine and motorcar for one specific sense. Also, all these senses are linked with others forming a network. One of the most common links between synsets is the hypernym, hyponymy. With this, each synset is linked with more generic synsets through its hypernym relation and also to more specific synsets through its hyponymy relation. There's also a way to distinguish words between nouns and instances like specific persons, countries and geographic entities.

There are some algorithms that can be used to find similar synsets in WordNet.

- **Path Distance Similarity**: It is a scaled metric for measuring the similarity between a pair of senses based on the shortest path that connects the senses in the hypernym/hyponym taxonomy. (MENG; HUANG; GU, 2014).

- **Wu-Palmer Similarity**: Based on the depth of the senses in the taxonomy and of their Least Common Subsumer (most specific ancestor node). (MENG; HUANG; GU, 2014).

2.5 Word embedding

Word embeddings or vector space models of word semantics can be seen as a way of representing words, allowing the creation of NLP applications capable of understanding textual analogies even with few data for training. The use of word embeddings has been exceptionally successful in many NLP tasks over the past few years. In many cases it has completely replaced the traditional models in the distributional field like Brown clusters and LSA.

In many traditional NLP applications, **one-hot vectors** were used to represent words of a vocabulary. In this case, we have a vector for each word of the vocabulary with the same size, filled with zeros beside the position of the word where we have the value one. One problem of using the one-hot representation is that you can't generalize crosswords because of the inner product of any 1-hot vector is always zero. And for this reason, we cannot apply any distance-like metrics to evaluate the similarity. For this reason, a **feature vector** is preferable to word representation. In this case, we have for each word an vector of size d filled with real values between 0 and 1 that represents multiple features. There are several ways to learn these high dimensional feature vectors values.

These feature vectors or embeddings can be generated using neural networks. Bengio et al. (2003) first introduced the term word embeddings with a simple feed forward neural language model to learn these vectors. After this, other models emerged with the creation of a toolkit named *word2vec* presented by Mikolov et al. (2013a).

Word2vec is a predictive embedding model composed by two main architectures to produce word embeddings, as shown in Figure 1:

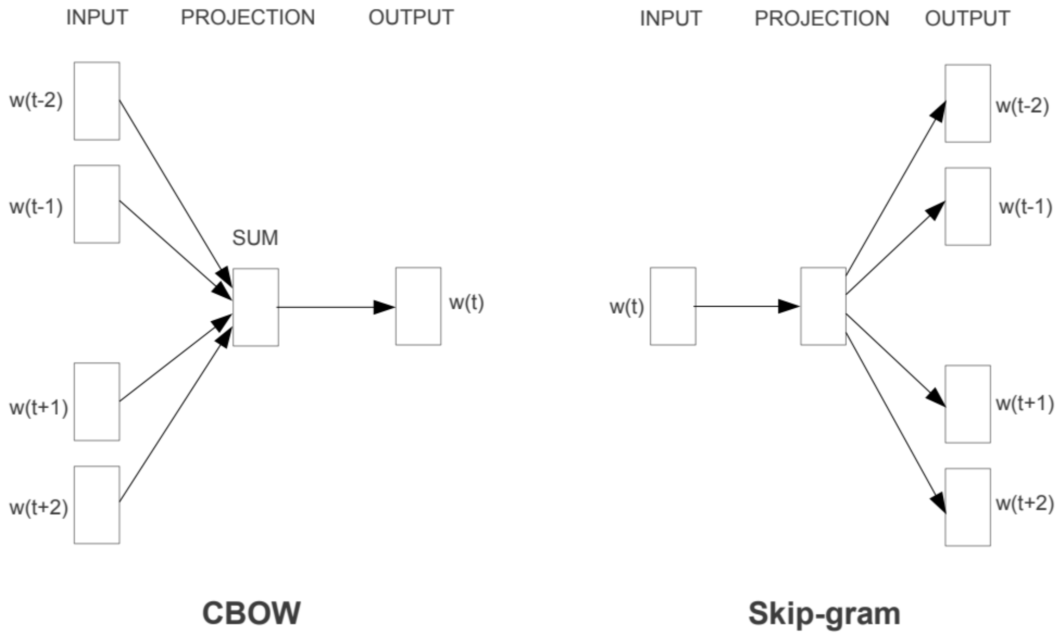
- **Continuous bag-of-words** (CBOW): Learns the embedding by predicting the current word based on their surrounding words (context).
- **Skip-gram** (SG): It is the opposite of CBOW, it learns by predicting the surrounding words (context) given a current word.

After word2vec other algorithms emerged, such as **Global Vectors** (GloVe). Pennington, Socher and Manning (2014) describes their work as "GloVe, is a new global log-bilinear regression model for the unsupervised learning of word representations that outperforms other models on word analogy, word similarity, and named entity recognition tasks.". GloVe is an approach that combines the global statistics of matrix factorization with the context based learning from word2vec.

Ling et al. (2015) presents **Wang2vec** an extensions of the original word2vec models to improve the embeddings obtained for syntactically motivated tasks, by introducing changes that made the network aware of the relative positioning of context words.

Later on, Bojanowski et al. (2016) introduces **FastText**, which is another way to learn word representations by taking into account subword information. They incorporate character n-grams into the Skip-Gram model. By their evaluation the model outperforms baselines that do not take into account subword information.

Figure 1 – Word2Vec training architectures.



Source: Taken from Mikolov et al. (2013b, p. 5).

2.6 Word similarity

We can think of synonyms as a way to determine if two words are similar or not. In other words, they are similar if they have the same meaning, or are near-synonyms. According to Jurafsky and Martin (2009, p. 749),

> Two words are more similar if they share more features of meaning, or are near-synonyms. Two words are less similar, or have greater semantic distance, if they have fewer common meaning elements. Although we have described them as relations between words, synonymy, similarity, and distance are actually relations between word senses. For example of the two senses of bank, we might say that the financial sense is similar to one of the senses of fund while the riparian sense is more similar to one of the senses of slope.

The ability to identify the semantic similarity between words has been a subject of research very explored in the last years, because it is related to a series of activities of the area of natural language processing like information retrieval, text summarization, categorization and generation, database schema matching, question answering, machine translation, and others. (ISLAM; INKPEN; KIRINGA, 2007; JURAFSKY; MAR-

TIN, 2009).

In short, the techniques for identifying similarity can be classified into two main approaches, knowledge-based and distributional-based. **Knowledge-based** similarity models are those that rely on pre-existing knowledge resources, such as thesauri, semantic networks, taxonomies, or encyclopedias. (AGIRRE et al., 2009). Moreover, almost all techniques concerning the **distributional-based** approach come from the basis of statistical semantics in which we have the Distribution Hypothesis which is defined by the fact that words occurring in the same contexts tend to have similar meanings. (HARRIS, 1954). Here the techniques are formed mainly by inducing distributional properties of words from corpora. (AGIRRE et al., 2009).

3 RELATED WORK

In this chapter, will be presented works encountered while doing the bibliographic research. Where, to find the state-of-the-art regarding word similarity, a search with Google Scholar and Semantic Scholar was used. The terms used to find the related work in the field was "word similarity", "semantic similarity", "synonym", "Synonym extraction", "semantic embedding" and "morphological embedding". Also, some search through the Association for Computational Linguistics website revealed some events regarding Semantic Textual Similarity, like the SemEval where we look through the best ranking algorithms used.

3.1 Comparing Semantic Relatedness between Word Pairs in Portuguese Using Wikipedia

In this paper, Granada, Santos and Vieira (2014) presents a new dataset for evaluating Distributional Similarity Models in Portuguese. 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 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.

With this, we intend to use the PT65 as a gold-standard for evaluating semantic similarity and relatedness between words with word embeddings models and the WordNet.

3.2 Dependency-Based Word Embeddings

In this work, Levy and Goldberg (2014) presents a generalized skip-gram model with negative sampling introduced by Mikolov et al. (2013a), 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 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.

We intend to adapt this work to a Brazilian Portuguese corpus, to check if the results will be, similar or not regarding other word embedding models, but instead of using the WordSim353 which is for English, we intend to use another dataset.

3.3 Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks

Hartmann et al. (2017) present in this paper, 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 in Brazilian or European Portuguese. With that they applied some preprocessing (Tokenization and normalization) in order to reduce the vocabulary size. Using the corpus as input, they trained some 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 use 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 CBOW models, except Wang2Vec, had poor results in semantic analogies.

For the POS tagging task evaluation, the Wang2Vec had the best results, and as seen, 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. In the end, they suggest that word analogies are not very suitable for evaluating word embeddings and task-specific is probably a better approach.

We used they pre-trained word embedding models for comparison while evaluating our models under our specific task - word similarity on PT65.

3.4 ELMo and BERT

Peters et al. (2018) presents in this work, a general approach for learning context-dependent representations from bidirectional language models (biLMs). They called it Embeddings from Language Models (ELMo), and we can image it as a new kind of word embedding, that, instead of learning a word as a vector representation it has the intent 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 RNNs to encode context.

As these works represent the state-of-the-art evolution from the first word embeddings and allow pre-trained models to be used for general purpose NLP tasks, we intend to explore how these language models behave for word-level tasks such as word similarity.

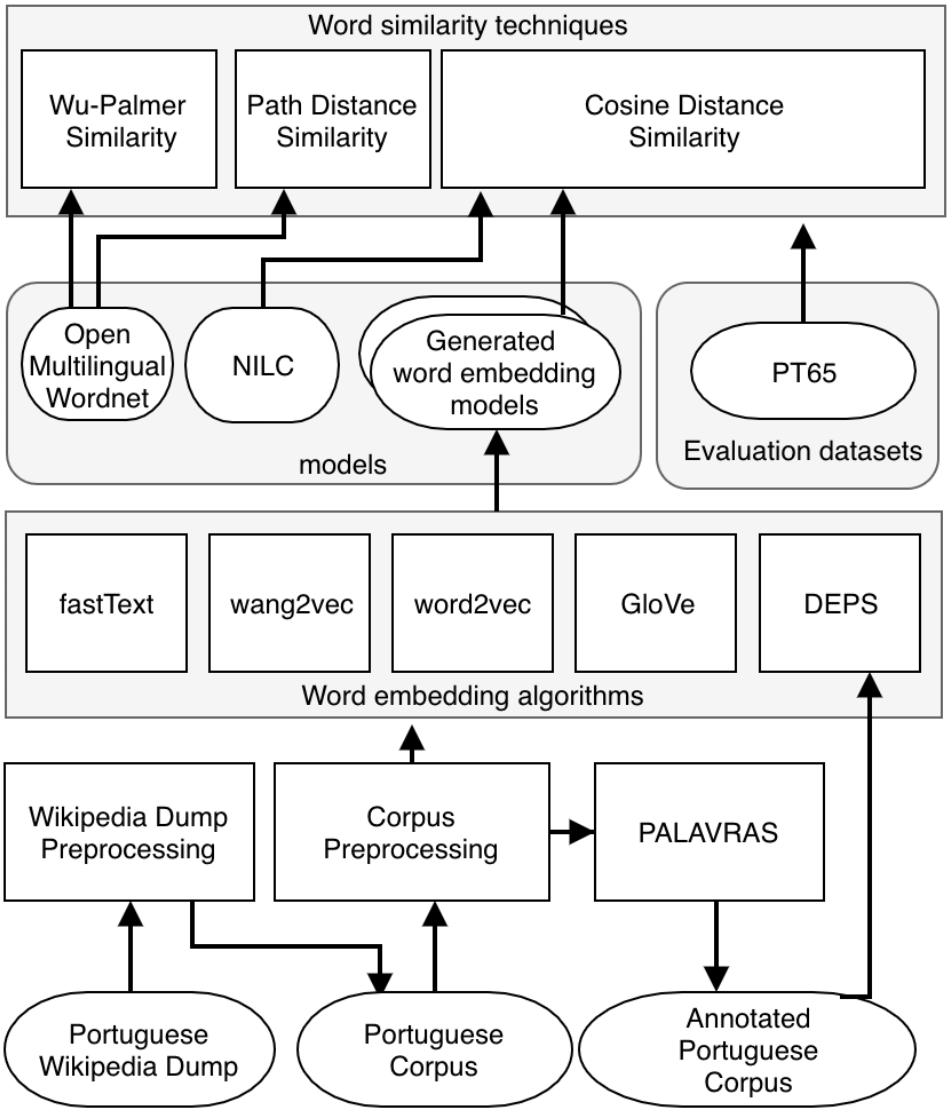
4 METHODS AND MATERIALS

This chapter will present a description of what and how this work was done, as well as the tools and methods used. First subsection 4.1 presents an overview of the architecture and how the proposed experiment was realized. Then, the subsection 4.2 presents the dataset used in the evaluation process. After that we start with a detailed explanation of the Corpus generation in subsection 4.3, of the syntactic parsing in subsection 4.4. Then we explain how we generate the most common word embedding models in subsection 4.5 and at last we explain how we reproduced the work of Levy and Goldberg (2014) for Portuguese in subsection 4.6.

4.1 Architecture overview

The proposed work consists of comparing different word similarity techniques. Therefore, Figure 2 defines an overview of the architecture with the intention of comparing several techniques using different algorithms and testing them with a common dataset.

Figure 2 – Proposed architecture



Source: Made by the author.

In this work, we compared techniques based on the two main approaches to word similarity, the knowledge-based and the distributional-based.

Regarding the knowledge-based approach we utilized a lexical base, in this case, **Open Multilingual Wordnet** (OMW) wore used with **Path Distance** and **Wu-Palmer** similarity techniques. It was decided to use OMW due to 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; FOSTER, 2013).

For the distributional approach we **generated word embedding models** with a corpus obtained from the **Brazilian Portuguese Wikipedia dump** of articles. The word embeddings wore generated using several different model implementations for learning word representations. In this case, **FastText**, **Wang2vec**, **Word2vec** and **GloVe**. Also, we compare our word embedding models with a set of pre-trained models available from **Núcleo Interinstitucional de Linguística Computacional** (NILC) in all different implementations (FastText, Wang2vec, Word2vec, and GloVe). One thing to note is that the metric used for the comparison of similarity between one word and another for all word embedding models was the **cosine distance**. The **CBOW** and **Skip-gram** were used for the models that has this option. (BOJANOWSKI et al., 2016; LING et al., 2015; MIKOLOV et al., 2013a; PENNINGTON; SOCHER; MANNING, 2014; HARTMANN et al., 2017)

We also generated one more model in order to take into account the syntactic tree information of the sentences from the Portuguese corpus using the algorithm implementation by (Levy2014) which generates the **DEPS** model. In order to do this, we used the **PALAVRAS syntactic parser** to annotate the corpus with syntactic information. (BICK, 2000).

In the end, we do a quantitative evaluation of all models and techniques using the **PT65** dataset, which consists of a pair of words and a similarity value given by persons. (GRANADA; SANTOS; VIEIRA, 2014).

All the experiments were done using the *Semantics* computer (Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz with 32 cores and 128GB of RAM) granted by the *UNISINOS Programa de Pós-graduação em Computação Aplicada* (PPGCA) by running **Docker** containers.

4.2 PT65 Dataset

This dataset is composed of 65 word pairs, initially generated by Rubenstein and Goodenough (1965) on the name of *RG65*. This word pairs wore translated to Portuguese by Granada, Santos and Vieira (2014) and evaluated with 50 persons.

The initial idea 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, we would have to translate to Portuguese and then the human-assigned similarity judgments would not fit entirely regarding the semantic changes involved in the translation process. So the PT65 dataset was used in the evaluation process.

4.3 Corpus generation

In this section, we will present the process involved in generating a corpus that can be used on NLP tasks from a Wikipedia dump. We have used this process to generate the Word Embeddings for evaluation in this thesis. While we focus on the Portuguese language, you could easily do the same thing for the other available languages in Wikipedia.

4.3.1 Getting the Wikipedia PT-BR dump

First, we downloaded the latest Portuguese Wikipedia articles dump[1]. The file is a big, compressed XML file that contains all articles in the wiki text format, just like markdown but with some special tokens that deal with some specific Wikipedia features. For example: "*[[Imagem:Starsinthesky.jpg|thumb|[[Estrela|Formação estrelar]] na [[Grande Nuvem de Magalhães]], uma [[galáxia irregular]].]]*"

More detailed information about the dump formats and different languages can be found in their website[2].

4.3.2 Preprocessing with Wikiextractor

As described in the previous step, the format of the dump is not suitable for most of NLP tasks. That's why we need to parse the wiki text format to raw text. In order to do

[1] `https://dumps.wikimedia.org/ptwiki/latest/ptwiki-latest-pages-articles-multistream.xml.bz2`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Database_download`

this, we have a few options. We could use the python **gensim.corpora.WikiCorpus** class but its tokenizer is not so good for Portuguese (In our case we need to have words separated by '-', like *guarda-chuva*, which is very common in Portuguese). So, we ended up using the **wikiextractor** project that just reads the XML file and outputs all the documents in parsed text. We chose to cleanup and tokenize the corpus at a later stage. So, we just cloned the repository and executed the **wikiextractor**.

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 Figure 3.

Figure 3 – WikiExtractor output sample.

```
<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>
```

Source: Made by the author.

It is also possible to save this as only one text file just by changing the tool arguments. At the time of writing, there were 1,000,400 documents in the ptwiki-dump.

4.3.3 Custom preprocessing

In order to cleanup the sentences for generating the Word embedding models we did some custom pre-processing[3] based on Hartmann et al. (2017) preprocessing scripts. Some changes were made to do some 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

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

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 we'll use a Syntactic parser that has better accuracy if we maintain this)
- Remove sentences with less than 4 tokens (as it does not add meaningful value to the corpus we can remove very short sentences).
- 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, we ended up with the final 1.6GB PT-BR corpus file which contains 9.896.520 sentences, 251.193.592 tokens, and 3.137.040 unique tokens.

4.4 PALAVRAS annotated corpus generation

To annotate all sentences of the corpus with syntactic tags, we used the software PALAVRAS developed by Bick (2000), which is an automatic parser for Portuguese.

First, we tried to use the parser with multiple sentence files of 1MB. However, the parser was taking too much time to execute and sometimes errors occurred. So we wrote a Python script that sends sentences in batches to the PALAVRAS parser and saves the results. Also, we have used parallel computing doing this process times the number of cores on the machine. Although, we first run this on a i5 2.4GHz computer with 4 cores, achieving an average speed of 16 sentences per second, which means that for all 9896520 sentences it would take 7 days to complete. We have tried other techniques attempting to increase the speed, but the bottleneck was indeed in the parser tool.

With this problem at hand, *UNISINOS Programa de Pós-graduação em Computação Aplicada* (PPGCA) granted us access to the *Semantics* computer (Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz with 32 cores and 128GB of RAM). With 32 cores the parsing step should be concluded in 24 hours.

One more problem that we had is that the PALAVRAS could not parse some of the sentences. Since we were running the parser in batches, this means that if one sentence failed, we lost all the parsed sentences in the batch. Also, as this process would take too long, we had to implement some way to continue the process if some fatal error occurred. With this in mind, we converted the sentences file to an SQLite table with three columns (id, text and parsed text). With this whenever we start the parsing process, we can continue from where it stopped.

With this solution implemented, we created a docker image and started running in the Semantics machine. The overall process took 38 hours. 24 hours to process 8916000 sentences using batches of 30, and 14 hours to process the remaining ones without sending in batches. Resulting in a 15GB corpus file.

4.5 Common Word Embeddings generation

In order to have a base for comparison, we generated all models that were used by Hartmann et al. (2017). In this case, FastText, Wang2vec, Word2vec and GloVe using different dimensions values like 50, 100, 300, 600 and 1000. (BOJANOWSKI et al., 2016; LING et al., 2015; MIKOLOV et al., 2013a; PENNINGTON; SOCHER; MANNING, 2014). Also, the CBOW and Skip-gram were used for the models that have this option.

For this, we downloaded all those model generation tools from GitHub, compiled into a docker image, and use as input our corpus text file. We only created some script to run this several times with different dimension sizes as it took several hours to complete.

4.6 DEPS Word Embedding generation

In order to generate the DEPS (dependency-based syntactic contexts) word embedding model proposed by Levy and Goldberg (2014), we got the source code called *word2vecf* from their website. The input required by this model is three files, a word and 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*> <*dependency-relation*>_<*referred-word*>".

In order to generate this file from our corpus we got the annotated output from the PALAVRAS parser and extracted the syntactic tags. This process is shown by in Figure 4.

Figure 4 – 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.". Right: Sample of our generated contexts file for an annotated sentence.

<ß>					
A	[o] <*> <artd> DET F S	**@>N**	**#1->2**	a	>n_astronomia
astronomia	[astronomia] <domain> N F S	**@SUBJ>**	**#2->3**	astronomia	subj>_é
é	[ser] <fmc> <vK> <mv> V PR 3S IND VFIN	**@FS-STA**	**#3->0**		
uma	[um] <card> NUM F S	**@<SC**	**#4->3**	uma	<sc_é
de	[de] <sam-> <np-close> PRP	**@N<**	**#5->4**	de	n<_uma
as	[o] <-sam> <artd> DET F P	**@>N**	**#6->9**	as	>n_ciências
mais	[mais] <quant> <KOMP> ADV	**@>A**	**#7->8**	mais	>a_antigas
antigas	[antigo] <jh> ADJ F P	**@>N**	**#8->9**	antigas	>n_ciências
ciências	[ciência] <domain> N F P	**@P<**	**#9->5**	ciências	p<_de
$.			**#10->0**		
</ß>					

Source: Made by the author.

The generation of this 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 9896000 parsed sentences with an average speed of 196.2 sentences per second.

After we had the input files, we just run the *word2vecf* tool which took some hours to complete. We generated models with different dimensions values as 50, 100, 300, 600 and 1000.

5 RESULTS

We separate the evaluation in three steps. In subsection 5.1 we do a quantitative evaluating of the Open Multilingual WordNet. In subsection 5.2 we do the same evaluation but with the word embeddings models. At last, we do a qualitative evaluation regarding the DEPS model in subsection 5.3

5.1 Open Multilingual WordNet evaluation

In order to do a quantitative evaluation of the knowledge-based approach for word similarity. We used the Open Multilingual Wordnet (OMW) from Bond and Foster (2013) and loaded it with the Natural Language Toolkit (NLTK) library. We then calculated the similarity between the pair of words from the PT65 dataset using two algorithms, Path Distance and Wu-Palmer. With this, we calculated the Pearson's Correlation (ρ) for each of the techniques.

Table 1 shows the results, and as we can see, the Path Distance algorithm gave a relatively high score, but as stated by Jurafsky and Martin (2009, p. 297) we indeed have out of vocabulary words, in this case, 15.38% of the words.

Table 1 – OMW evaluation on PT65. 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.

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

Source: Made by the author.

5.2 Word embeddings Evaluation

To do a quantitative evaluation of the distributional approach for word similarity we did the same experiment as the WordNet evaluation but with our word embeddings models. We loaded the PT65 dataset, and for each word pair, we compared the expected result with the Cosine similarity given by the model. With this, we calculated the 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) we can use word embeddings to cover out-of-vocabulary words. In comparison with the WordNet, we can also see, that it has slightly better results, which is a good thing considering that it has no manual construction as WordNet.

Also, we can see that 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, maybe 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) from Núcleo Interinstitucional de Linguística Computacional (NILC).

5.3 Qualitative Evaluation of DEPS model

For evaluating our DEPS model we did a qualitative evaluation were we manually inspect the five most similar words (by cosine similarity) to a given set of target words (Board 1), and we compared it with other models, just like Levy and Goldberg (2014) did in their experiment.

The first target word, *longe* (Far), we can see similar results provided by all the different models. The word *inglês* (English) have the same behavior. However, for some specific words, like *faculdade* (College) we can see that 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.

Table 2 – Word embeddings evaluation on PT65. $\rho(ours)$ is the Pearson's Correlation value from our trained models. $\rho(nilc)$ is the Pearson's Correlation values from the NILC pre-trained models.

Embedding Models		**Size**	$\rho(ours)$	$\rho(nilc)$
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	

Source: Made by the author.

Board 1 – Target words and their five most similar words per word embedding models.

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

Source: Made by the author.

6 FINAL CONSIDERATIONS

The study carried out in this work, proves that the detection of similarity between words is a very important topic for the NLP segments as summarization, information retrieval, and question answering. More precisely techniques for identifying word similarity can help in a number of NLP tasks such as dialogue systems, question answering, and information retrieval systems. (ISLAM; INKPEN; KIRINGA, 2007; PILEHVAR; JURGENS; NAVIGLI, 2013; AGIRRE et al., 2009).

It was also found indications that the expansion of terms using WordNet has several problems, where a word may not be present. Since these lexical bases are of manual construction, they are time-consuming and expensive, and for this reason, not all links will be present and their quality varies from language to language. There is also no WordNet for all languages. (LEEUWENBERGA et al., 2016). Distributional approaches regarding word similarity have been proven to be more competitive than the thesaurus-based approach, and have been successfully being used to cover out-of-vocabulary items in WordNet. (AGIRRE et al., 2009).

Thus, we explored the existing techniques regarding word similarity, using a distributional approach called word embeddings, adapting existing works to Brazilian Portuguese. We also experiment with other techniques that are solely based on a lexical database such as WordNet, and we evaluate all the different techniques over a common dataset called PT65. Proving that word embeddings can cover words out of vocabulary and have slightly better results in comparison with WordNet. We also 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, but for the task of word similarity against the dataset PT65, it had the worst results.

As a future work, we intend to evaluate the different techniques against the work of Araujo, Hentges and Rigo (2018) to see if we can improve the ENSEPRO results. As well as to explore how we could use BERT and ELMo language models for word-level tasks such as word similarity as these works represent the state-of-the-art.

UTILIZANDO WORD EMBEDDINGS PARA SIMILARIDADE DE PALAVRAS NO PORTUGUÊS BRASILEIRO

Resumo: A capacidade de identificar a similaridade semântica entre palavras tem sido objeto de pesquisa nos últimos anos, pois está relacionada a uma série de atividades da área de processamento de linguagem natural, como recuperação de informação, sumarização de texto, categorização e geração, tradução automática e outros. A maioria dos sistemas de resposta a perguntas e extração de informações usa o WordNet para procurar sinônimos em seu mecanismo de busca. No entanto, a expansão de termos usando o WordNet tem vários problemas. Eles são de construção manual, demorados e caros, e por esse motivo, nem todos os links estarão presentes e sua qualidade varia de idioma para idioma, assim como não está disponível para todos os idiomas. Abordagens baseadas em distribuição, como a *word embedding*, foram usadas para cobrir itens fora do vocabulário no WordNet. Assim, com a possibilidade de acesso a diferentes *word embeddings models* e a necessidade de melhorar a forma de expandir os termos relacionados aos sistemas de consulta para bases ontológicas utilizadas por sistemas de perguntas e respostas e recuperação de informação, o presente trabalho explora as técnicas existentes para identificação de similaridade entre palavras, usando a abordagem distribuicional chamada word embeddings, adaptandando trabalhos existentes para o português brasileiro. Também é realizado experiementos com outras tecnicas que são basesadas em bases lexicas como WordNet, aonde uma avaliação qualitativa é realizada de todas as técnicas sobre um *dataset* comun PT65. Provando que word embeddings podem de fato cobrir as palavras faltantes e tem um resultado ligeiramente melhor quando comparado com o WordNet. Também é realizado uma adaptação de estudos sobre a adição do contexto sintático no processo de trainamento do word embeddings a partir e um corpus português brasileiro, aonde obtivemos resultados similares atraves de uma avaliação qualitativa, porem para a atividade de identificar palavras similares utilizando o dataset PT65 os resultados foram piores em considersão aos outros modelos.

Palavras-chave: Similaridade de palavras. WordNet. Word embedding. Linguística computacional. Processamento de Linguagem Natural.

REFERENCES

AGIRRE, E. et al. A study on similarity and relatedness using distributional and WordNet-based approaches. **Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09**, [S.l.], n. June, p. 19, 2009.

ARAUJO, D. de; HENTGES, A.; RIGO, S. Uma abordagem linguística para sistemas de perguntas e respostas curtas. In: Simpósio Brasileiro de Sistemas de Informação, Caxias do Sul/RS, 2018. **Anais...** [S.l.: s.n.], 2018.

BENGIO, Y. et al. A neural probabilistic language model. **J. Mach. Learn. Res.**, [S.l.], v. 3, p. 1137–1155, Mar. 2003.

BICK, E. **The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Århus: University of Arhus, 2000.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. **arXiv preprint arXiv:1607.04606**, [S.l.], 2016.

BOND, F.; FOSTER, R. Linking and extending an open multilingual wordnet. In: ACL, 2013. **Anais...** [S.l.: s.n.], 2013.

CRUSE, D.; CRUSE, D. **Lexical semantics**. [S.l.]: Cambridge University Press, 1986. (Cambridge Textbooks in Linguistics).

DEVLIN, J. et al. Bert: pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, [S.l.], 2018.

FELLBAUM, C. **Wordnet**: an electronic lexical database. [S.l.]: Bradford Books, 1998.

FINKELSTEIN, L. et al. Placing search in context: the concept revisited. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 10., 2001, New York, NY, USA. **Proceedings...** ACM, 2001. p. 406–414. (WWW '01).

GRANADA, R.; SANTOS, C. T. dos; VIEIRA, R. Comparing semantic relatedness between word pairs in portuguese using wikipedia. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE - 11TH INTERNATIONAL CONFERENCE, PROPOR 2014, SÃO CARLOS/SP, BRAZIL, OCTOBER 6-8, 2014. PROCEEDINGS, 2014. **Anais...** [S.l.: s.n.], 2014. p. 170–175.

HARRIS, Z. S. Distributional structure. **WORD**, [S.l.], v. 10, n. 2-3, p. 146–162, 1954.

HARTMANN, N. et al. Portuguese word embeddings: evaluating on word analogies and natural language tasks. **CoRR**, [S.l.], v. abs/1708.06025, 2017.

ISLAM, A.; INKPEN, D.; KIRINGA, I. Applications of corpus-based semantic similarity and word segmentation to database schema matching. **The VLDB Journal**, [S.l.], v. 17, p. 1293–1320, 2007.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing (2nd edition)**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009.

LEEUWENBERGA, A. et al. A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. **The Prague Bulletin of Mathematical Linguistics**, [S.l.], v. 105, n. 105, p. 111–142, 2016.

LEVY, O.; GOLDBERG, Y. Dependency-Based Word Embeddings. **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, [S.l.], p. 302–308, 2014.

LING, W. et al. Two/too simple adaptations of word2vec for syntax problems. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2015., 2015. **Proceedings...** Association for Computational Linguistics, 2015.

LYONS, J. **Linguistic semantics**: an introduction. [S.l.]: Cambridge University Press, 1995. (Cambridge Approaches to Lingui).

MENG, L.; HUANG, R.; GU, J. Measuring semantic similarity of word pairs using path and information content. **Int. J. Future Gener. Commun. Netw**, [S.l.], v. 7, n. 3, p. 183–194, 2014.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: NIPS, 2013. **Anais...** [S.l.: s.n.], 2013.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **CoRR**, [S.l.], v. abs/1301.3781, 2013.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: global vectors for word representation. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, [S.l.], p. 1532–1543, 2014.

PETERS, M. E. et al. Deep contextualized word representations. In: NAACL, 2018. **Proceedings...** [S.l.: s.n.], 2018.

PILEHVAR, M. T.; JURGENS, D.; NAVIGLI, R. Align, disambiguate and walk: a unified approach for measuring semantic similarity. **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**, [S.l.], p. 1341–1351, 2013.

Princeton University. **About wordnet.** [Online; accessed 31-May-2018], `https://wordnet.princeton.edu/`.

RODRIGUES, J. A. et al. Lx-dsemvectors: distributional semantics models for portuguese. In: PROPOR, 2016. **Anais...** [S.l.: s.n.], 2016.

RUBENSTEIN, H.; GOODENOUGH, J. B. Contextual correlates of synonymy. **Commun. ACM**, [S.l.], v. 8, p. 627–633, 1965.

ZGUSTA, L.; CERNY, V. **Manual of lexicography**. [S.l.]: De Gruyter, 1971. (Janua Linguarum. Series Maior).