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

EDUARDO EIDELWEIN BERLITZ

SEMANTIC MODEL FOR WORD SIMILARITY

São Leopoldo
2018

Eduardo Eidelwein Berlitz

SEMANTIC MODEL FOR WORD SIMILARITY

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

SEMANTIC MODEL FOR WORD SIMILARITY

Eduardo Eidelwein Berlitz*
Sandro José Rigo**
Rodrigo da Rosa Righi***

1 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 1.1 presents an overview of the architecture and how the proposed experiment was realized. Then, the subsection 1.2 presents the dataset used in the evaluation process. After that we start with a detailed explanation of the Corpus generation in subsection 1.3.

1.1 Architecture overview

The proposed work basically consists of comparing different word similarity techniques. Therefore, Figure 1 defines an overview of the architecture with the intention of comparing several techniques using different algorithms and testing them with a common dataset.

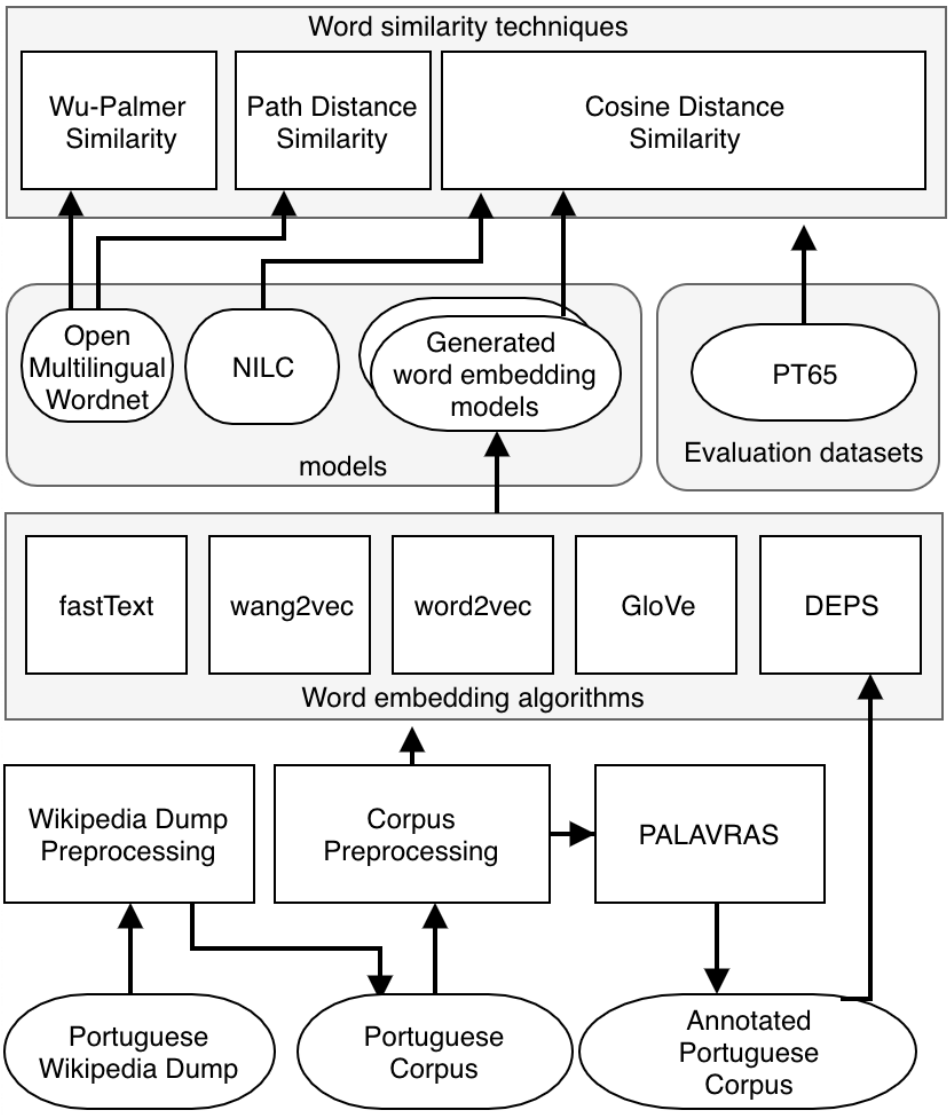
In this work we compared techniques based on the two main approaches to word similarity, the knowledge-based and the distributional-based.

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.

For the distributional approach we **generated word embedding models** with a corpus obtained from the **brazilian portuguese Wikipedia dump** of articles. The word

* 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

Figure 1 – Proposed architecture



Source: Made by the author.

embeddings wore generated using several diferent 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 by **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.

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 imple-

mentation by (Levy2014) wich generates the **DEPS** (Word2vecf) model. In order to do this, we used the **PALAVRAS syntactic parser** to annotate the corpus with syntactic information.

In the end we do a quantitative evaluation of all models and techniques using the **PT65** dataset, which consists in a pair of words and a similarity value given by persons.

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.

1.2 PT65 Dataset

This dataset is composed by 65 word pairs, initialy 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 judgements. But we would have to translate to portuguese and then the human-assigned similarity judgements would not fit entirelly regarding the semantic changes involved in the translation proccess. So the PT65 dataset was used in the evaluation process.

1.3 Corpus generation

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

1.3.1 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]].]]
```

You can find more detailed information about the dump formats and different languages in their website[2].

1.3.2 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 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 in a later stage. So, we just cloned the repository and executed the **wikiextractor**:

```
git clone https://github.com/attardi/wikiextractor.git
cd ./wikiextractor
python ./WikiExtractor.py --no-templates -o ../data/ptwiki-articles-text/ -b 10M
  ↪  -c ../data/ptwiki-latest-pages-articles-multistream.xml.bz2
cd ..
```

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 expands these templates, because it will just add duplicated sentences to the content.

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

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

So, it is really important to use the *–no-templates* flag. This tool generated multiple compressed 10MB files of wiki articles sentences in the following format:

```
<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>
```

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.

1.3.3 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 the 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 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 I'll use a Syntactic parser that has better accuracy if I 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.

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

- 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.

1.4 PALAVRAS annotaded corpus generation

TODO

1.4.1 converted sentences to a sqlite to be able to recover progress of palavras parser in case of errors

TODO

1.4.2 python code to run the PALAVRAS Parser in multiple cores to reduce from 14 days to 24 hours

TODO

1.5 Common Word Embeddings generation

TODO

1.5.1 GloVe,FastText,word2vec,wang2vec model generations

TODO

1.6 DEPS Word Embedding generation

TODO

1.6.1 script to generate word2vecf(DEPS) input files from the palavras anotated sentences

TODO

1.6.2 deps model generation

TODO

1.7 WordNet evaluation

TODO

1.7.1 script to quantitative evaluate two WordNet similarity metrics (Wu-Palmer and Path-distance) against the PT65 dataset

TODO

1.8 Word embeddings Evaluation

TODO

1.8.1 script to quantitative evaluate all models against the PT65 dataset

TODO

1.8.2 script to query the top most similar words to a given set of target words for manual qualitative evaluation.

TODO

1.8.3 script to download all NILC pre-trained word embeddings and quantitative evaluate all models against the PT65 dataset

TODO

REFERENCES

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.

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

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