Automatically Extracting Synonyms in Brazilian Portuguese*

Bruno Ferrari Guide[1,2]

[1] University of Sao Paulo, Sao Paulo, Brazil
[2] Verbio Technologies, Brazil
bruno.fguide@gmail.com

Abstract. While new models to represent meaning as multidimensional vectors (word embedding models) such as Word2Vec [4] are consistently accurate to determine measures such as meaning similarity and relatedness between words, those models are not efficient to identify if two words are related because they have the same meaning (synonyms), the opposite meaning (antonyms) or some group relation (such as hyponyms and hypernyms). The main goal of this research is to investigate the way vector models work to represent meaning, how they are limited and if their capacities can be complemented with other approaches in order to enrich the way they represent meaning. We will work not only by methodically describing meaning and the vector representation of meaning, but also by testing several different tools, such as WordNet, in a pipeline to see if substantial results in identifying synonyms can be obtained in an efficient way. This task will also require compiling a corpus of Synonym identification in Brazilian Portuguese.

Keywords: Word Embedding · Synonym · WordNet.

1 Word Embeddings - Models

The project aims to describe several of the most prominent models of word embedding, not only due to the importance of mapping their inner works[3] but also to determine their limitations to represent meaning in natural language.

Word embedding models are a way to represent words as vectors in vector spaces (usually with a big number of dimensions) that are based on some method of statistically representing language. A good model is one which the mathematical relations between different points on space have a resemblance with relations between specific words [10].

Those models are sold as a way to efficiently represent meaning in Cartesian spaces, which allows a sort of soft transposition between linguistic operations

* Supported by the Linguistics Department of Sao Paulo University.
[3] Specially since there is a lack of material in Portuguese explaining those models to linguists.

and algebraic ones. This becomes clear when we see, for instance, the idea of measuring vector distance as a way to determine if two words are closely related or not.

We aim to describe and scrutinize the functioning of Word2Vec [4], LSA [2], and GloVe [6] models. This list is far from being exhaustive and is clearly open to further analysis and development.

Of course, meaning in natural language is an extremely complex concept and a highly pervasive one. Word embeddings are a way to formally express one of the aspects of meaning, and those sort of models are quite successful to do so, one initial goal of this project is to specify what dimensions of meaning are not represented by those models and then to proceed by analysing what other types of models can be combined with word embedding in creating more robust meaning representations.

We aim to clearly establish what can be expected and what cannot in terms of representing meaning using this sort of tool.

2 Semantic Relations

From all the different ways that two or more words can be related, such as being part of the same group, having family relatedness, being opposites in meaning, one being a member or containing the other, we will focus on one specific semantic relation: Synonym.

Synonym occurs when two words have the same meaning. A formal definition will be derived in this work in order to delimit the experimental phase that will come later in the research process.

This definition will also be of vital importance to formulate adequate expectations from how synonym can even be represented in a way by word embedding models, a working hypothesis in this research is that this sort of model cannot formalize the aspect of meaning fundamental to recognizing synonyms.

This semantic relation was chosen due to the fact that there is already some strong research work done about it, such as [1], [7], [3],[5], [9], [8]. This provides a solid ground from which to build up, and we intend to develop a framework to allow further research about synonym and meaning representation to be done in a more consistent way.

3 WordNet and other tools

Parallel to the theoretical discussion already presented, we will work on implementing all the word embedding models that will be analyzed. Those implementations will feed the next phase of the research, which will combine word embedding with automatic synonym detection models.

Those models will be built by allying word embedding models with knowledge tools such as Wordnet, thesauri, dictionaries, and ontologies [4]. The idea here is to come up with several possible solutions to the matter of automatic synonym extraction.

We believe that if the computational models are not quite adequate to the matter at hand, there is a possibility that if they are combined with knowledge-rich solutions, the results will be worthy.

4 Corpus and Testing

A final important aspect of this research is that it is not only focused on a theoretical discussion regarding meaning. We also intend to build a corpus that will be useful to test models in recognizing semantic relations between words and also to identify specifically synonyms amongst related words.

From this development, we expect to be able to test different approaches to this task and compare the results, which would generate quantitative input to the theoretical discussion, a recurrent theme in the author's research.

The initial idea of the corpus is to gather a vast amount of groups of sentences, each group would consist of sentences that are almost identical, except for one noun or verb. The varying words of each group would be synonyms, antonyms, non-related or related words amongst themselves. Each sentence of each group would be paired with all the other sentences of that group and that pair would get a label representing if their meaning is equal, opposite or just different. From that, a successful model would correctly predict the label of a sentence pair when it should be equal (that is, the only different word between those two sentences are synonyms.).

From this corpus, each and every proposed model will be tested, cross-validated and then a quantitative discussion about the success rates of the models will be made.

[4] There is also a possibility to implement naive probabilistic models to try to automatically identify Synonyms.

References

1. Fellbaum, C.: WordNet. Wiley Online Library (1998)
2. Landauer, T.K.: Latent semantic analysis. Wiley Online Library (2006)
3. Lin, D., Zhao, S., Qin, L., Zhou, M.: Identifying synonyms among distributionally similar words. In: IJCAI. vol. 3, pp. 1492–1493 (2003)
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
5. Muller, P., Hathout, N., Gaume, B.: Synonym extraction using a semantic distance on a dictionary. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. pp. 65–72. Association for Computational Linguistics (2006)
6. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
7. Scheible, S., Im Walde, S.S., Springorum, S.: Uncovering distributional differences between synonyms and antonyms in a word space model. In: IJCNLP. pp. 489–497. Citeseer (2013)
8. Thalenberg, B.: Diferenciao entre sinnimos e antnimos em espaos lexicais (2017)
9. Wang, T., Hirst, G.: Extracting synonyms from dictionary definitions. In: RANLP. pp. 471–477 (2009)
10. Widdows, D., Widdows, D.: Geometry and meaning, vol. 773. CSLI publications Stanford (2004)