Automatically extracting synonyms in Brazilian Portuguese

Bruno Ferrari Guide
University of Sao Paulo, Department of Linguistics
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.

Main Objectives

1. Have a solid understanding of how word embedding models (CBoW and SG) work and which linguistic model are beneath them.
2. Identify which linguistic limitations are inherent to the models.
3. Explore possible linguistic knowledge databases which can improve synonym extraction, such as WordNet, dictionaries, thesauri.
4. Combine different linguistic databases to improve their depth. (For instance, using Portuguese Dictionaries to improve Wordnet in a semi-supervised way).
5. Create a corpus to specifically deal with the Synonym Extraction in Portuguese.
6. Test different models to achieve a new standard for this task in Portuguese.

Word Embeddings and Synonym

- Word embedding models represent meaning through projecting words in a cartesian space using patterns of cooccurence.

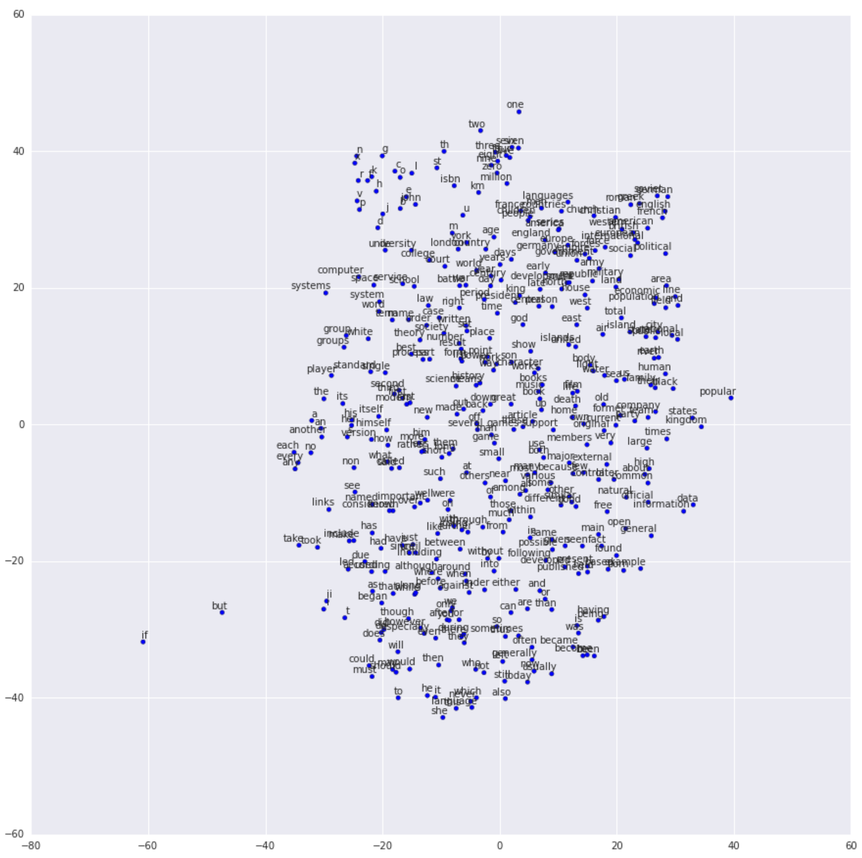

Figure 1: Word2Vec represented in a 2-dimensional space

- It is possible to use linear measurmentes to identify relationships between those represntations, those measurements tend to represent semantic relatednesss in some level, such as those presented in figure 2.

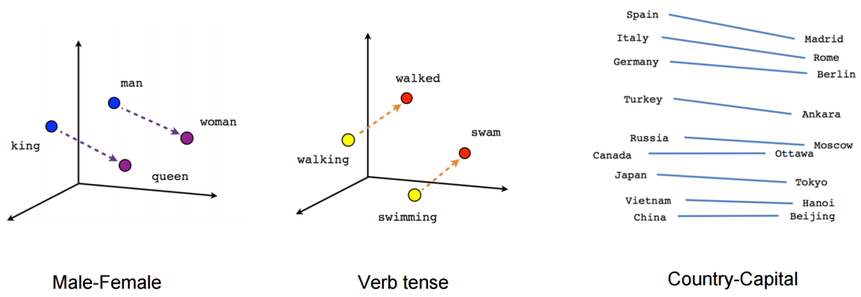

Figure 2: Linear relanshionships between words in word embedding models.

- there is one caveat, several different kinds of semantic relationships are represented in the same way, which is problematic when those different relations result in entirely different interpretations of some sentence.

- Synonym, the relationship between words which share their meaning, is specially important to not be thrown in the mix, since it is a useful relation to summarise texts, to work in automatic translation and so on.

- It is importatn to put that this work consider that it ir possible to enrich word embeddings in a way that they serve as base to more complex models which are able to differenciate between relatedness and synonym.

- Those models are not necessarely going to be hardwired to identify a previously defined set of synonyms, we aim to find semi-supervised or non-supervised tactics to extract synonyms from the vector space.

Ideas and Issues

- One major issue is one regarding the ambigous strcuture of language itself. There are words which are not synonyms of themselves. This is the case of polysemic words such as *mouse*:

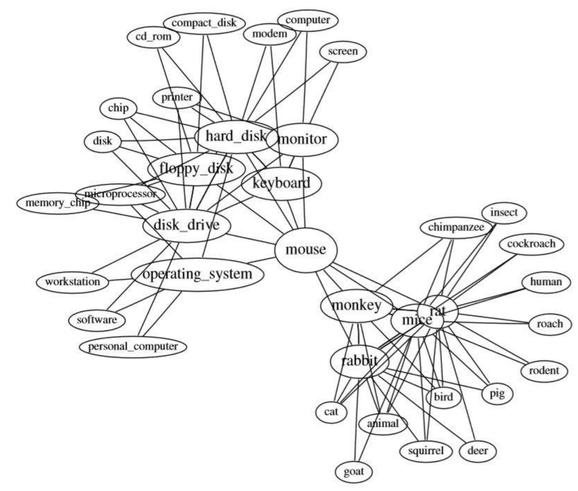

Figure 3: Neighbours of the ambiguous word *mouse* - Widdows 2004

- We will work around this issue using word sense disambiguation and studying consolidated approaches of this field to deal with this.

- To approach synonym extraction, we will use linguistic databases such as dictionaries and thesauri, as well as structured resources such as WordNet.

- Those DBs will be used to support models that differenciate between Synonym and other relationships of words.

- Classifiers like Naive-Bayes or Decision Tree will be used to explore neghbouring word vectors and to identify synonymic relationships.

Forthcoming Research

- Define the models which are going to be used on the task.

- Develop a corpus to train and test synonym extraction models in Portuguese.

- Compare the results to point out limitations and areas to improve.

- Based on the knowledge about the nature of synonym and test results, develop a deep analysis on the possibility of automation of synonym extraction.

References

[1] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[2] Thomas K Landauer. *Latent semantic analysis*. Wiley Online Library, 2006.

[3] Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. Identifying synonyms among distributionally similar words. In *IJCAI*, volume 3, pages 1492–1493, 2003.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[5] Philippe Muller, Nabil Hathout, and Bruno Gaume. Synonym extraction using a semantic distance on a dictionary. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 65–72. Association for Computational Linguistics, 2006.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[7] Silke Scheible, Sabine Schulte Im Walde, and Sylvia Springorum. Uncovering distributional differences between synonyms and antonyms in a word space model. In *IJCNLP*, pages 489–497. Citeseer, 2013.

[8] Bruna Thalenberg. Diferenciao entre sinnimos e antnimos em espaos lexicais, 2017.

[9] Tong Wang and Graeme Hirst. Extracting synonyms from dictionary definitions. In *RANLP*, pages 471–477, 2009.

[10] Dominic Widdows and Dominic Widdows. *Geometry and meaning*, volume 773. CSLI publications Stanford, 2004.