Comparing Semantic Relatedness between Word Pairs in Portuguese Using Wikipedia

Roger Granada[1], Cassia Trojahn[2], and Renata Vieira[3]

[1] PUCRS & IRIT - Toulouse, France
roger.granada@acad.pucrs.br
[2] UTM & IRIT - Toulouse, France
cassia.trojahn@irit.fr
[3] PUCRS - Porto Alegre, Brazil
renata.vieira@pucrs.br

Abstract. The growth of available data in digital format has been facil-itating the development of new models to automatically infer the seman-tic similarity between word pairs. However, there are still many natural languages without sufficient resources to evaluate measures of semantic relatedness. In this paper we translated word pairs from a well-known baseline for evaluating semantic relatedness measures into Portuguese and performed a manual evaluation of each pair. We compared the correlation with similar datasets in other languages and generated LSA models from Wikipedia articles in order to verify the pertinence of each dataset and how semantic similarity conveys across languages.

Keywords: Semantic relatedness, semantic similarity, similarity dataset.

1 Introduction

Discovering similar words in a document collection is still an open problem. The idea of semantic similarity was expressed by Zellig Harris [6] when he formulated the distributional hypothesis. This hypothesis is based on the idea that words that occur in the same contexts tend to have similar meanings. Models built on this assumption are called Distributional Similarity Models (DSMs) and take into account the co-occurrence distributions of the words in order to cluster them together. Several implementations of DSMs have been proposed in the last decades [3,5,8,10,15] and have being used in tasks such as query expansion [1], building bilingual comparable corpora [16], clustering [2], discovering of meaning of noun compounds [14] *etc.*

Although there are many proposals on DSMs, their practical applicability depends on their evaluation. However, evaluation is still an open issue since manual evaluation is a time consuming task and automatic evaluation requires a gold-standard. An approach to overcome this problem is to manually generate a gold-standard containing pairs of terms and a score associated to each pair [4,11,13].

An important resource for English has been defined by Rubenstein and Goodenough [13]. This dataset (from now on called as RG65) was developed to evaluate

semantic similarity measures and contains judgements from 51 human subjects for 65 word pairs. Judgements are scaled from 0 to 4 according to their similarity of meaning, where the greater the similarity between the words, the higher the score. Thus, 0 representing no similarity between words and 4 perfect similarity. The average correlation over the subjects was quite high, achieving $r = .85$.

Miller and Charles [11] repeated the experiments using a subset of RG65 dataset containing 30 word pairs. These pairs were selected according with their score in the original RG65 dataset: 10 pairs have high level of similarity scores (scores between 3 and 4), 10 pairs have intermediate level (scores between 1 and 3) and 10 pairs have low level (scores between 0 and 1). This new dataset (MC30) was evaluated by 38 human subjects who were asked to evaluate specifically the similarity of meaning and to ignore any other semantic relations. Comparing the results obtained using the MC30 dataset with the results obtained by Rubenstein and Goodenough using RG65 dataset the correlation achieved was $r = .97$.

Finkelstein *et al.* [4] expanded the initial MC30 dataset, increasing significantly the number of word pairs. WordSimilarity-353 or just WordSim-353[1] contains 353 pairs of words divided in two sets. The first set contains 153 word pairs along with their similarity scores assigned by 13 subjects. The second set contains 200 word pairs, with their similarity scores assessed by 16 subjects. The subjects were instructed to evaluate the word pairs on a scale ranging from 0 to 10 according to their relatedness, being 0 totally unrelated words and 10 very closely related or identical words. The correlation between MC30 and WorsSim-353 datasets is also quite high, having a Pearson correlation of $r = .95$.

In order to evaluate similarity measures in other natural languages, a translation of some datasets has been made. Joubarne and Inkpen in [9] translated the RG65 dataset into French in order to measure the semantic similarity using second-order co-occurrence measures. After translating all word pairs, 18 human subjects who are French native speakers evaluated the similarity between the word pairs. As the work by Rubenstein and Goodenough [13], evaluators judge the word pairs in a scale ranging from 0 to 4. According to the authors, there was a good agreement amongst the evaluators for 71% of the word pairs and a high disagreement for 23% of the cases. The correlation between RG65 original dataset and the French dataset (JI65) achieved $r = .91$.

Following the work by Joubarne and Inkpen [9], this work attempts to translate into Portuguese all pairs from RG65 and evaluate them using 50 human subjects. Human scores are compared with previous works and an automatic evaluation is performed by comparing LSA generated models from Wikipedia articles with each dataset. These experiments verify the pertinence of each dataset and how the semantic similarity conveys across languages.

2 Data and Methods

In order to generate a dataset for evaluating similarity measures using Portuguese, all word pairs from RG65 were translated into Portuguese by two native

[1] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`

speakers with proficiency in English. Each pair of words was translated separately and their relatedness score in RG65 was used as a hint to disambiguate words when multiple translations were possible. In some cases, the direct translation of each word from the pair resulted in one word, *e.g.*, the pair *cock* and *rooster* from RG65 has the same word *galo* as translation into Portuguese. As performed by Joubarne and Inkpen [9], in these cases the same word was kept as the translation of the word pair.

The evaluation process was performed by 50 undergraduate and graduate students who were asked to evaluate each pair according with their semantic relatedness. Following Rubenstein and Goodenough [13] our scores also range from 0 to 4. Results were averaged over all 50 subjects and the whole dataset (hereafter named as PT65) is freely available[2]. The average agreement among subjects was $r = .71$ having a standard deviation $\sigma = .13$. We use Pearson (r) and Spearman (ρ) correlation coefficients to measure the relation between scores, since Pearson correlation is highly dependent on the linear relationship between the distributions in question and Spearman mainly emphasizes the ability of the distributions to maintain their relative ranking. Table 1 presents the correlation scores between datasets for 65 word pairs datasets and for 30 word pairs datasets.

Table 1. Correlation between datasets evaluations

	r		ρ	
	JI65	PT65	JI65	PT65
RG65	.91	.90	.91	.83
JI65	-	.89	-	.85

	r			ρ		
	RG30	JI30	PT30	RG30	JI30	PT30
MC30	.97	.92	.87	.95	.89	.87
RG30		.91	.89	-	.90	.90
JI30			.86	-	-	.87

As reported by Miller and Charles [11] the correlation between MC30 and RG30 dataset was $r = .97$. RG65 and PT65 datasets achieved $\rho = .83$, the lowest correlation among datasets. On the other hand, their correlation using Pearson achieved $r = .90$, which is almost the same of the French dataset. Although the correlations of the Portuguese dataset have the lowest scores, their values are still relatively high (greater than $r = .80$).

In order to evaluate the pertinence with respect to the representativity of the word pairs in the languages, experiments using these datasets and Wikipedia dumps dating from February 2013 were performed. Each Wikipedia dump was pre-processed by WikiExtractor[3] (version 2.6) in order to extract and clean its content.

Each Wikipedia article was tokenized and a bag-of-words model was generated. Thus, each article is represented as an attribute vector of words that occur in the corresponding article. In order to remove noisy words, a threshold was applied removing words that appear less than 10 times in the whole Wikipedia.

[2] `http://www.inf.pucrs.br/linatural/wikimodels/similarity.html`
[3] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

The resulting vectors are weighted using Term Frequency - Inverse Domain Frequency (TFIDF) scheme and transformed into LSA models using the Gensim [12] tool.

LSA model uses Singular Value Decomposition (SVD) on a word-document matrix to extract its reduced representation by truncating the matrix to a certain size (also called the semantic dimension of the model). It is justified because it often improves the quality of the semantic space [10]. In this model, two words end up with similar vectors if they co-occur multiple times in similar documents. Thus, a similarity measure can be used between word vectors in order to measure the similarity between word pairs.

3 Experiments and Results

In our experiments, the LSA model was generated reducing the original matrix to a matrix containing 250 dimensions, *i.e.*, rank $k = 250$, and the cosine of the angle between word vectors was used to measure their similarity. Scores from Wikipedia were compared in terms of Pearson and Spearman correlations as presented on Table 2. The correlation between scores in different languages allows to see whether it is possible to transfer semantic similarity across languages.

Table 2. Pearson and Spearman correlation between datasets and Wikipedia data

		RG65	JI65	PT65	MC30	RG30	JI30	PT30
Pearson (r)	Wikipedia (EN)	.65	.59	.63	.69	.72	.62	.70
	Wikipedia (FR)	.57	.55	.53	.55	.53	.48	.50
	Wikipedia (PT)	.47	.56	.57	.67	.63	.72	.77
Spearman (ρ)	Wikipedia (EN)	.69	.61	.61	.71	.77	.66	.67
	Wikipedia (FR)	.52	.50	.38	.52	.51	.46	.39
	Wikipedia (PT)	.43	.42	.53	.66	.66	.69	.79

Joubarne and Inkpen in [9] suggest that it might be possible to transfer semantic similarity across languages. As Joubarne and Inkpen, the correlation found in our experiments using data in French suggests that it would be possible to transfer semantic similarity across languages. For example, looking at Table 2 it would be possible to use RG65 scores to find similar terms in French Wikipedia, since it achieved almost the same correlation when compared with JI65 dataset. On the other hand, scores found using Wikipedia in Portuguese achieved the highest correlation using Portuguese datasets, which is an evidence that using translated words evaluated by native speakers would get better results when compared with approaches that transfer the human scores across languages.

Observing the distributional similarity between the evaluations, the correlation using English and Portuguese Wikipedias has a similar behavior. Both languages presented an increase in correlation scores when the number of terms decreased, *i.e.*, when changing datasets from 65 to 30 word pairs. On the other hand, French Wikipedia had a decrease in correlation scores when the number of terms decreased (except for Spearman score using Portuguese dataset which increased $\rho = .01$). This decrease might be due to the fact that the MC30 dataset contains terms that are less related in the French Wikipedia.

Our correlation scores are close to the scores achieved by Hassan and Milhalcea [7] when using the MC30 dataset to evaluate a method based on Explicit Semantic Analysis (ESA). In that work, the authors achieved a Pearson correlation of $r = .58$ and a Spearman correlation of $\rho = .75$ for the English Wikipedia. In our experiments the correlation between the MC30 dataset and the English Wikipedia achieved a Pearson correlation of $r = .69$ and a Spearman correlation of $\rho = .71$. A comparison using other languages is not applicable since Hassan and Milhalcea used Arabic, Romanian and Spanish Wikipedias while our work used French and Portuguese Wikipedias. Joubarne and Inkpen in [9] used French to evaluate an automatic similarity measure, but unfortunately a comparison is not possible since they used Google n-grams as corpus.

4 Conclusions

In this paper we have proposed a resource that can be used as gold-standard for evaluating semantic similarity and relatedness between words, which results from the manual translation into Portuguese of a well-known baseline in English. The evaluation scores were compared with similar proposals in the literature which aimed at translating the English baseline in other languages, such as French.

Automatic evaluation was also performed by comparing LSA models based on Wikipedia articles with each proposed dataset. In this experiment we observed that it might be possible to transfer semantic similarity across languages, but for Portuguese, a manual evaluation of the translated word pairs has better results. We believe that this resource in Portuguese is specially useful as gold-standard for evaluating Distributional Similarity Models, supporting the automatic evaluation of such approaches.

Similarly to Hassan and Milhalcea [7], an approach to measure semantic similarity across languages would be to use the generated datasets to tests cross-lingual similarity using Wikipedia. Unlike Hassan and Milhalcea, instead of using only the English dataset (RG65), one could use both datasets (*e.g.*, RG65 for English and PT65 for Portuguese) and the evaluation score would be the mean of both evaluation scores.

Acknowledgments. This work is partially supported by the CAPES-COFECUB Cameleon project number 707/11.

References

1. Chen, L., Chen, S.: A New Approach for Automatic Thesaurus Construction and Query Expansion for Document Retrieval. International Journal of Information and Management Sciences 18(4), 299–315 (2007)
2. Di Marco, A., Navigli, R.: Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. Computational Linguistics 39(3), 709–754 (2013)
3. Erk, K.: Vector Space Models of Word Meaning and Phrase Meaning: A Survey. Language and Linguistics Compass 6(10), 635–653 (2012)
4. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing Search in Context: The Concept Revisited. ACM Transactions on Information Systems 20(1), 116–131 (2002)
5. Granada, R.L., Vieira, R., Strube de Lima, V.L.: Evaluating co-occurrence order for automatic thesaurus construction. In: IEEE 13th International Conference on Information Reuse and Integration (IRI), pp. 474–481 (2012)
6. Harris, Z.S.: Distributional structure. Words 10(23), 146–162 (1954)
7. Hassan, S., Mihalcea, R.: Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge. In: EMNLP 2009, pp. 1192–1201. Association for Computational Linguistics, Stroudsburg (2009)
8. Iosif, E., Potamianos, A.: Similarity computation using semantic networks created from web-harvested data. Natural Language Engineering, 1–31 (2014)
9. Joubarne, C., Inkpen, D.: Comparison of Semantic Similarity for Different Languages Using the Google N-gram Corpus and Second-Order Co-occurrence Measures. In: Butz, C., Lingras, P. (eds.) Canadian AI 2011. LNCS (LNAI), vol. 6657, pp. 216–221. Springer, Heidelberg (2011)
10. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104(2), 211–240 (1997)
11. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language & Cognitive Processes 6(1), 1–28 (1991)
12. Rehurek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta (2010)
13. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Communications of the ACM 8(10), 627–633 (1965)
14. Utsumi, A.: A semantic space approach to the computational semantics of noun compounds. Natural Language Engineering 20(2), 185–234 (2014)
15. Yang, D., Powers, D.M.W.: Automatic thesaurus construction. In: 31st Australasian conference on Computer science – ACSC 2008, pp. 147–156. Australian Computer Society, Inc., Darlinghurst (2008)
16. Zhu, Z., Li, M., Chen, L., Yang, Z.: Building Comparable Corpora Based on Bilingual LDA Model. In: 51st Annual Meeting of the Association for Computational Linguistics, pp. 278–282. Association for Computational Linguistics, Sofia (2013)