Exploring Unsupervised Methods to Textual Similarity

Pablo Gamallo[1] and Martín Pereira-Fariña[2]

[1] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela (Galiza)
pablo.gamallo@usc.es
[2] ARG-tech, University of Dundee (Scotland)
m.z.pereirafarina@dundee.ac.uk

Abstract. This paper presents some unsupervised methods for detecting semantic textual similarity, which are based on distributional models and dependency parsing. The systems are evaluated using the dataset realased by the ASSIN Shared Task co-located with PROPOR 2016. The results do not improve the state-of-the-art for Portuguese language.

1 Introduction

Paraphrases are defined as pairs of sentences that convey the same or almost the same information [4]. Paraphrase identification is then the task of recognizing sentences (or small textual fragments) with approximately the same meaning within a specific context. A similar task to paraphrasing is Semantic Textual Similarity (STS), which measures the degree of semantic equivalence between two chunks of texts. STS is beneficial for many NLP applications, ranging from information retrieval to plagiarism detection. There has been proposed several methods to STS: from unsupervised and resource-light approaches to supervised and resource-intensive methods.

The objective of the paper is to describe and evaluate unsupervised methods to STS based on distributional models and applied to Portuguese language. More precisely, our aim is to compare resource-light with resource-intensive (syntactic-based) unsupervised strategies for STS. Experiments will be carried out using the dataset realesed at ASSIN Shared Task (*Avaliação de Similaridade Semântica e Inferência Textual*, co-located with PROPOR 2016 [7].

The rest of the paper is organized as follows. In the next section (2), we introduce the existing STS approaches for Portuguese. Then, Section 3 describes three different unsupervised methods. In Section 4, we present and discuss the results of our experiments; and, finally, in Section 5, our main conclusions and future work are summarised.

2 Semantic Textual Similarity for Portuguese Language

STS is one of two tasks evaluated at the ASSIN (*Avaliação de Similaridade Semântica e de Inferência Textual*) [7]. The other related subtask, textual infer-

ence, is beyond the scope of the current paper. The STS task consists of assign a numerical value (from 1 to 5) to pairs of sentences, according to the degree of similarity between them: the higher the value the most similar the sentences are. Such a task was inspired by the SemEval Task 2 on Semantic Textual Similarity [1, 2]. For the shared task on STS at SemEval 2016, 119 different systems were submitted, which shows the enormous interest in this field.

Most systems at ASSIN (all except one) were based on supervised techniques. The best team [13] used linear regression to train a classifier whose features are cosine values representing the degree of similarity of each pair of sentences. Sentences are represented in two different ways: vector addition of TF-IDF values (each word in the sentence is a TF-IDF value), and vector addition of distributional values, where each word is represented as a contex vector learnt using neural network techniques Cosine similarities between these two types of repesentations are the input of the basic classifier.

The second best system (and the best one on the European Portuguese dataset) [6] trained a classifier based on regression models (namely, Kernel Ridge Regression) with a greater number of features than the previous system, including Edit distances between strings, size of the largest common substring, different similarity metrics relying on occurrences and TF-IDF values. In total, the system used more than 90 features.

The only unsupervised strategy at ASSIN is called Reciclagem and was proposed by [3], who used just similarity metrics on the basis of semantic relations extracted from external thesauri and lexical resources.

In the current paper, we will evaluate on the dataset provided by ASSIN some unsupervised strategies mainly based on distributional models.

3 Unsupervised Semantic Textual Similarity

Three different unsupervised strategies are defined: the most basic one is just based on distributional similarity and PoS tagging, while the rest of methods rely on syntactic analysis and open information extraction techniques.

3.1 Distributional Similarity

A basic strategy to compute sentence similarity consists of adding up the similarity scores of each pair of words appearing in the two compared sentences. Only lexical words (nouns, verbs and adjectives) are considered. Cosine similarity is computed by using pre-training word embeddings. The algorithm is the following: take the shorter sentence and take the first lexical word in it, then compute the cosine similarity between this word and all the lexical words in the longest sentence and sum up the similarity values in order to get the lexical relevance of the first word with regard to the longest one. Do the same for all the words of the shortest sentence and divide the final score by the total number of words of this sentence, so as to compute the average score. More formally, given a word vector $\mathbf{w_s}$ belonging to U_s, where U_s is the set of lexical word vectors of the

shortest sentence, the lexical relevance, LR, of $\mathbf{w_s}$ given the longest sentence is computed as follows:

$$LR(\mathbf{w_s}, U_l) = \sum_{\mathbf{w_i} \in U_l}^{L} Cosine(\mathbf{w_s}, \mathbf{w_i}) \tag{1}$$

where U_l is the set of lexical word vectors of the longest sentence and L the number of lexical words in that one. So, the final similarity score (DSim) for a pair U_s and U_l is the average of LR:

$$\text{DSim}(U_s, U_l) = \frac{\sum_{\mathbf{w_i} \in U_s}^{S} LR(\mathbf{w_s}, U_l)}{S} \tag{2}$$

where S is the number of lexical words in the shortest sentence. This strategy also definitely does not encode order information.

3.2 Basic Proposition Extraction

DSim only considers lexical variations by identifying semantic relations at the word level without considering word order and syntactic dependencies. In order to consider such phenomena, we apply the previous similarity strategy (DSim) to the basic propositions extracted from the sentences instead of applying them directly on the whole sentences. Basic propositions are subject-verb-object relations identified by means of Open Information Extraction (OIE) techniques [5, 9]. Each sentence may contain several basic propositions. So proposition-based similarity (BPROP) is just computed on the basis of words contained in the extracted propositions.

3.3 Argument Structure

The last strategy is very similar to BPROP, but instead of extracting all possible subject-verb-object relations, the goal is to identify the main argument structure of each sentence. We consider that the main argumentative structure of each sentence is constituted by the root and its direct dependent arguments. So argument-based similarity (ARGSTR) is computed on the basis of lexical words contained in the skeleton structure extracted from the compared sentences.

4 Expermiments

To evaluate how well the strategies defined in the previous section are suited to capture STS, we test them on the datasets provided by ASSIN shared task [7]. Experiments were carried out with several publicly available pre-trained semantic models, namely the syntax-based and trasparent distributional models reported in [8].

The test dataset was processed with different modules of the multilingual and open source suite, LinguaKit [10].[3] More precisely, in order to implement all the strategies described above, we used the PoS tagger module [12], the dependency-based parser provided by LinguaKit [11], which was required for ARGSTR strategy, and the OIE-based relation extractor module [9], required for BPROP.

Table 1. Scores (Pearson correlation) achieved by our three systems and the unsupervised strategy (*Reciclagem*) submitted at ASSIN shared task.

Systems	European PT	Brazilian PT	Total
DSim	**0.54**	0.56	0.53
ARGSTR	0.27	0.22	0.24
BPROP	0.29	0.24	0.26
Reciclagem	0.53	**0.59**	**0.54**

Table 1 shows the scores, in terms of Pearson correlation, of the three strategies defined above (DSim, ARGSTR, and BPROP) by using three STS lists of sentences' pairs: European Portuguese, Brazilian Portuguese, and the union of both lists (Total). Each pair of sentence is assigned a value between 1 to 5, so that the greater the value, the greater the similarity between the two sentences. A system is evaluated by measuring the correlation between the annotated values and those provided by the system. The Table also shows in the last row the scores reached by the only unsupervised system, *Reciclagem*, submitted to ASSIN Shared task. The best scores are reached by the most basic approaches: DSIM and Reciclagem. Both approaches just rely on lemmatization and pre-existing semantic resources: corpus-based distributional models (DSim) and external thesauri (Reciclagem). By contrast, the two strategies based on syntactic analysis and open information extraction (ARGSTR and BPROP) return very disappointing score values. Although an in-depth error analysis is needed, a surface analysis of the results suggests that the syntactic errors made by the analyzer are determinant.

5 Conclusions

Different unsupervised strategies for semantic textual similarity have been tested and evaluated. The baseline method relying on just counting shared words and similar ones clearly outperform more complex techniques enriched with syntactic analysis and basic proposition extraction. In future work, we will analyze in detail the type of errors made by the syntactic-based techniques in order proposes new unsupervised strategies for STS. We will also test these techniques with datasets oriented towards other types of tasks than the STS, for instance tasks aimed

[3] `https://github.com/citiususc/Linguakit`

at identifying paragraphs or rephrases, which would be still more sensitive to syntactic information.

Acknowledgments

This work has received financial support from project TelePares (MINECO, ref:FFI2014-51978-C2-1-R), and the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

1. Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *SemEval@NAACL-HLT*, pages 252–263. The Association for Computer Linguistics, 2015.
2. Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, 2016.
3. Ana Oliveira Alves, Ricardo Rodrigues, and Hugo Gonçalo Oliveira. ASAPP: alinhamento semântico automático de palavras aplicado ao português. *Linguamática*, 8(2):43–58, 2016.
4. I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135 – 187, 2010.
5. Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open Information Extraction: the Second Generation. In *International Joint Conference on Artificial Intelligence*, pages 3–10. AAAI Press, 2011.
6. Pedro Fialho, Ricardo Marques, Bruno Martins, Luísa Coheur, and Paulo Quaresma. Inesc-id@assin: Medição de similaridade semântica e reconhecimento de inferência textual. *Linguamática*, 8(2):33–42, 2016.
7. Erick Rocha Fonseca, Leandro Borges dos Santos, Marcelo Criscuolo, and Sandra Maria Aluísio. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13, 2016.
8. Pablo Gamallo. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 51(3):727–743, 2017.
9. Pablo Gamallo and Marcos García. Multilingual Open Information Extraction. In *Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11*, pages 711–722. Springer, 2015.
10. Pablo Gamallo and Marcos Garcia. Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática*, 9(1), 2017.

11. Pablo Gamallo and Marcos Garcia. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management*, Available online 5 June 2018, 2018.
12. Marcos Garcia and Pablo Gamallo. Yet another suite of multilingual NLP tools. In *Languages, Applications and Technologies*, volume 563 of *Communications in Computer and Information Science*, pages 65–75, Switzerland, 2015. Springer. Revised Selected Papers of the Symposium on Languages, Applications and Technologies (SLATE 2015).
13. Nathan Siegle Hartmann. Solo queue at ASSIN: combinando abordagens tradicionais e emergentes. *Linguamática*, 8(2):59–64, 2016.