TeamZ: Measuring Semantic Textual Similarity for Spanish Using an Overlap-Based Approach

Anubhav Gupta
UFR SLHS
Université de Franche-Comté
anubhav.gupta@edu.univ-fcomte.fr

Abstract

This paper presents an overlap-based approach using bag of words and the Spanish WordNet to solve the STS-Spanish subtask (STS-Es) of SemEval-2014 Task 10. Since bag of words is the most commonly used method to ascertain similarity, the performance is modest.

1 Introduction

The objective of STS-Es is to score a pair of sentences in Spanish on the scale of 0 (the two sentences are on different topics) to 4 (the two sentences are completely equivalent, as they mean the same thing) (Agirre et al., 2014). The textual similarity finds its utility in various NLP applications such as information retrieval, text categorisation, word sense disambiguation, text summarisation, topic detection, etc. (Besançon et al., 1999; Mihalcea et al., 2006; Islam and Inkpen, 2008).

The method presented in this paper calculates the similarity based on the number of words that are common in two given sentences. This approach, being simplistic, suffers from various drawbacks. Firstly, the semantically similar sentences need not have many words in common (Li et al., 2006). Secondly, even if the sentences have many words in common, the context in which they are used can be different (Sahami and Heilman, 2006). For example, based on the bag of words approach, the sentences in Table 1 would be scored the same:

No.	Spanish	English
1	Él es listo.	He is clever.
2	Él está listo.	He is ready.
3	Él está preparado.	He is prepared.

Table 1: Examples.

However, only sentences [2] and [3] mean the same.

Despite the flaws, this approach was used because of the Basic Principle of Compositionality (Zimmermann, 2011), which states that the meaning of a complex expression depends upon the meaning of its components and the manner in which they are composed. Furthermore, mainly nouns were considered in the bag of words because Spanish is an exocentric language, and nouns contain more specific, concrete semantic information than verbs (Michael Herslund, 2010; Michael Herslund, 2012).

2 Methodology

The training dataset provided for the task consisted of 65 pairs of sentences along with their corresponding similarity scores. There were two test sets: one consisted of 480 sentence pairs from a news corpus, and the other had 324 sentence pairs taken from Wikipedia.

The approach consisted of learning the scoring with the help of linear regression. Two runs were submitted as solutions. The first run used three-feature vectors, whereas the second one used four-feature vectors. The features are the Jaccard indices for the lemmas, noun lemmas, synsets, and noun subjects in each sentence pair. For both runs, the sentence pairs were parsed using the TreeTagger (Schmid, 1994). The TreeTagger was used because it provides the part-of-speech tag and lemma for each word of a sentence.

Run 1 used these features:

- The fraction of lemmas that were common between the two sentences. In other words, the number of unique lemmas common between the sentences divided by the total number of unique lemmas of the two sentences.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 633–635,
Dublin, Ireland, August 23-24, 2014.

- The fraction of noun lemmas common between the two sentences.

- The fraction of synsets common between the two sentences. For each noun, its corresponding synset[1] was extracted from the Spanish WordNet (spaWN) of the Multilingual Central Repository[2] (MCR 3.0) (Gonzalez-Agirre et al., 2012).

Run 2 employed one more feature in addition to the aforementioned, which was the fraction of synsets of noun subjects that were common for each sentence pair. The subject nouns were extracted from the sentences after parsing them with the MaltParser (Nivre et al., 2007). Since the TreeTagger PoS tagset[3] differed from the EAGLES (Expert Advisory Group on Language Engineering Standards) tagset[4] required by the MaltParser, rules were written to best translate the TreeTagger tags into EAGLES tags. However, one-to-one mapping was not possible: EAGLES tags are seven characters long and encode number and gender, whereas TreeTagger tags do not. For example, using the EAGLES tagset, the masculine singular common noun *árbol* 'tree' is tagged as NCMS000, whereas the feminine singular common noun *hoja* 'leaf' is tagged as NCFS000; TreeTagger, on the other hand, tags both as NC.

3 Results and Conclusions

Table 2 presents the performance, measured using the Pearson correlation, of the approach. **Run 1** achieved a weighted correlation of 0.66723 and ranked 15th among 22 submissions to the task.

Dataset	Run 1	Run 2
Training	0.83693	0.83773
Wikipedia (Test)	0.61020	0.60425
News (Test)	0.71654	0.70974

Table 2: Performance of the Approach.

Given that the approach relied mostly on bag of words, a modest performance was expected. The performance was also affected by the fact that the spaWN did not have synsets for most of the words. Finally, converting TreeTagger tags to those required by the MaltParser instead of using a parser which annotates with EAGLES tags may also have contributed to the relatively low **Run 2** score. However, the confidence intervals of the two runs obtained after bootstrapping overlapped. Thus, the difference between the two runs for both the datasets is not statistically significant.

[1] stored as synset offset in wei_spa-30_variant.tsv

[2] The resource can be obtained from http://grial.uab.es/descarregues.php

[3] http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/data/spanish-tagset.txt

[4] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Acknowledgements

I would like to thank Vlad Niculae, Àngels Catena and Calvin Cheng for their inputs and feedback.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. "SemEval-2014 Task 10: Multilingual Semantic Textual Similarity." In *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval-2014). Dublin, Ireland.

Romaric Besançon, Martin Rajman, and Jean-Cédric Chappelier. 1999. Textual Similarities Based on a Distributional Approach. In *Proceedings of the 10th International Workshop on Database & Expert Systems Applications.* 180–184. DEXA '99. Washington, DC, USA: IEEE Computer Society.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the Sixth International Global WordNet Conference (GWC '12).*

Aminul Islam and Diana Inkpen. 2008. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data* 2 (2): 1–25.

Michael Herslund. 2010. Predicati e sostantivi complessi. In *Language, Cognition and Identity*, eds. Irn Korzen and Emanuela Cresti. 1–9. Strumenti per La Didattica E La Ricerca. Firenze University Press.

Michael Herslund. 2012. Structures lexicales et typologie. In *Sémantique et lexicologie des langues d'Europe*, eds. Louis Begioni and Christine Bracquenier. 35–52. Rivages Linguistiques. Presses Universitaires de Rennes.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence.* 775–80. AAAI'06. Boston, Massachusetts: AAAI Press.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics.

IEEE Transactions on Knowledge and Data Engineering, 18 (8): 1138–50.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetovlas Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13 (2): 95–135.

Mehran Sahami and Timothy D. Heilman. 2006. A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets. In *Proceedings of the 15th International Conference on World Wide Web*, 377–86. WWW '06. New York, NY, USA: ACM.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Thomas Ede Zimmermann. 2011. Model-theoretic semantics. *Semantics. An International Handbook of Natural Language Meaning*. edited by Claudia Maienborn, Klaus von Heusinger, and Paul Portner. Vol. 1. Berlin, Boston: De Gruyter Mouton.