MiniExperts: An SVM Approach for Measuring Semantic Textual Similarity

Hanna Béchara[*a], **Hernani Costa**[*b], **Shiva Taslimipoor**[a], **Rohit Gupta**[a], **Constantin Orăsan**[a], **Gloria Corpas Pastor**[b] **and Ruslan Mitkov**[a]
[a]RIILP, University of Wolverhampton, UK
[b]LEXYTRAD, University of Malaga, Spain
{hanna.bechara,hercos,shiva.taslimi,r.gupta,
c.orasan,gcorpas,r.mitkov}@{[a]wlv.ac.uk,[b]uma.es}
*These two authors contributed equally to this work.

Abstract

This paper describes the system submitted by the University of Wolverhampton and the University of Malaga for SemEval-2015 Task 2: *Semantic Textual Similarity*. The system uses a Supported Vector Machine approach based on a number of linguistically motivated features. Our system performed satisfactorily for English and obtained a mean 0.7216 Pearson correlation. However, it performed less adequately for Spanish, obtaining only a mean 0.5158.

1 Introduction

Similarity measures play an important role in a wide variety of Natural Language Processing (NLP) applications. Information Retrieval (IR), for example, relies on semantic similarity in order to determine the best result for a related query. Semantic similarity also plays a crucial role in other applications such as Paraphrasing and Translation Memory (TM). However, computing semantic similarity between sentences remains a complex and difficult task. Over the years, SemEval's shared tasks worked to fine-tune and perfect these similarity measures, and explore the nature of meaning in language.

SemEval2015's Task 2 involves computing how similar two sentences are in both English (Subtask 2a) and Spanish (Subtask 2b). In this paper we detail our submission to SemEval Task 2. We use an improved and revised version of the system presented in our SemEval 2014 submission (Gupta et al., 2014). As in Gupta et al., 2014, we employ a Machine Learning (ML) method which exploits available NLP technology, adding features inspired by deep semantics (such as parsing and paraphrasing) with distributional Similarity Measures, Conceptual Similarity Measures, Semantic Similarity Measures and Corpus Pattern Analysis[1] (CPA).

The remainder of the paper is structured as follows. Section 2 describes our approach, i.e. explains how the data was preprocessed and what features were extracted. Section 3 is divided in two section, the first one describes the ML algorithm and how it was tuned for this task (section 3.1) and the second one shows the obtained results along with a descriptive analysis of the runs based on the test and training data provided by the SemEval-2015 Task 2 (section 3.2). Finally, section 4 presents the final remarks and highlights our future plans for improving the system.

2 Approach

This section describes our approach to calculating semantic relatedness. It covers all the required preprocessing steps to extract the features themselves.

2.1 Data Preprocessing

This section presents all the tools, libraries and frameworks used to preprocess not only the test datasets but also the training datasets.

2.1.1 POS-Tagger, Lemmatiser, Stemmer

The software we used for these specific NLP tasks were: the Stanford CoreNLP[2] (Toutanova et al.,

[1]http://pdev.org.uk
[2]http://nlp.stanford.edu/software/corenlp.shtml

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 96–101,
Denver, Colorado, June 4-5, 2015.

2003) toolkit, which provides a lemmatiser, POS-Tagger, NER, parsing, and coreference; the TT4J[3] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995); and the Porter stemmer algorithm provided by the Snowball[4] library.

2.1.2 Named Entity Recogniser (NER)

The library used to identify named entities in English and Spanish was the Apache OpenNLP library[5]. For English, all the pre-trained NER models made available by the Apache OpenNLP library were used (i.e. we used models to identify dates, locations, money, organisations, percentages, persons and time). We also used all the pre-trained NER models for Spanish (in this case, we used models to identify persons, organisations, locations and miscellanea).

2.1.3 Translation Model

Since one of the features we implemented was available only for English (i.e. the Semantic Similarity Measures), we trained a Statistical Machine Translation (SMT) system to translate our Spanish dataset into English. For this purpose, we used the PB-SMT system Moses (Koehn et al., 2007), 5-gram language models with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002), the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), with refinement and phrase-extraction heuristics as described in Koehn et al., 2003. We trained this system on the Europarl Corpus (Koehn, 2005) and used Minimum Error Rate Training (MERT) (Och, 2003) for tuning on the development set.

2.1.4 Resources

Given that a number of our features depends on stopwords (see section 2.2), we compiled two lists of stopwords, one for English and another one for Spanish. Both are freely available to download[6].

We also used two lists (English and Spanish) of candidates for Multiword Expressions (MWEs) as a resource for one of the features (see section 2.2.5). These lists were extracted from the Europarl Corpus (Koehn, 2005) using the collocation modules of the NLTK package (Loper and Bird, 2002), and sorted by the degree of likelihood association between their components.

2.2 Extracted Features

This section details the features that our system uses to measure the semantic textual similarity between two sentences. The system uses the same features for both Subtask 2a and Subtask 2b. In addition to the baseline features used in Gupta et al., 2014, we introduced a set of Distributional, Semantic and Conceptual Similarity Measures, as well as a feature reflecting MWEs across sentences.

2.2.1 Baseline Features

The system is built on the baseline system developed for SemEval2014, which consists of 13 features explained in detail in Gupta et al., 2014. The code which implements these features can be found on GitHub[7].

2.2.2 Distributional Similarity Measures

Information Retrieval (IR) (Singhal, 2001) is the task of locating specific information within a collection of documents or other natural language resources according to some request (Salton and Buckley, 1988; Costa et al., 2010; Costa et al., 2011). Among IR methods, we can find a large number of statistical approaches based on the occurrence of words in documents or sentences. Following Harris' distributional hypothesis (Harris, 1970), which assumes that similar words tend to occur in similar contexts, these methods are suitable, for instance, to find similar sentences based on the words they contain or to compute the similarity of words based on their co-occurrence. To that end, we can assume that the amount of information contained in a sentence could be evaluated by summing the amount of information contained in the sentence words. Moreover, the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988). Bearing this in mind, we used two independent IR measures, the Spearman's Rank Correlation Coefficient (SCC) and the χ^2 to compute the similarity between two sentences

[3] https://code.google.com/p/tt4j
[4] http://snowball.tartarus.org
[5] http://opennlp.apache.org
[6] https://github.com/hpcosta/stopwords
[7] https://github.com/rohitguptacs/wlvsimilarity

written in the same language (cf. Kilgarriff, 2001). Both measures are particularly useful for this task because they are independent of text size (mostly because both measures use a list of the common entities), and they are language-independent. In detail, for every pair of sentence (English and Spanish), we used the lemmas to extract the list of common terms to compute both measures.

2.2.3 Conceptual Similarity Measures

This feature aims to find the conceptual similarity between two sentences written in the same language. In order to calculate the conceptual similarity, we took advantage of the BabelNet[8] (Navigli and Paolo Ponzetto, 2012) multilingual semantic network. As BabelNet organises lexical information in a semantic conceptual way, we created a conceptual sentence for all input pair of sentences (English and Spanish). More precisely, for every pair of sentence we only extracted lemmatised nouns, verbs, adjectives and adverbs. Then, a conceptual term list was built by extracting all the occurrences of the term in the conceptual network (i.e. BabelNet). As a result, we got a "conceptual representation" of both sentences, each of them containing a set of conceptual term lists. Next, for every term in the "conceptual_sentence_1", we counted the number of co-occurrences in the conceptual term lists in the "conceptual_sentence_2". In other words, we intersected the terms in sentence 1 with all the conceptual term lists in sentence 2. After computing all the co-occurrences, we used these values to calculate the Jaccard' (Jaccard, 1901), Lin' (Lin, 1998) and PMI' (Turney, 2001) scores.

2.2.4 Semantic Similarity Measures

This feature takes advantage of the Align, Disambiguate and Walk (ADW)[9] library (Pilehvar et al., 2013), a WordNet-based approach for measuring semantic similarity of arbitrary pairs of lexical items. It is important to mention that this feature is the only one that only works for English, which explains why we have a translation model (see section 2.1.3). In other words, when we are dealing with Spanish text, we use the trained model to translate from Spanish to English.

As the ADW library permits us to measure the semantic similarity between two raw English sentences, either by using disambiguation or not, we used both options to calculate all the comparison methods made available by the library, i.e. WeightedOverlap, Cosine, Jaccard, KLDivergence and JensenShannon divergence.

2.2.5 Multiword Expressions

Multiword Expressions (MWEs) are meaningful lexical units whose distinct idiosyncratic properties call for special treatment within a computational system. Non-compositionality is one of the properties of MWEs. The degree of association between the components of a MWE has been proved to be a promising approach to find out how much they are non-compostional and therefore how probable they are acceptable MWEs (Ramisch et al., 2010). The more non-compositional a MWE is, the more important is not to treat its components separately for NLP purposes, including processing semantic similarities.

For the purpose of our experiments, we focused on two more common types of MWEs in English and Spanish: `verb noun` combinations and `verb particle` constructions. Whenever a `verb+noun` or a `verb+particle` combination occurs in our sentence pair, we search a prepared list MWEs, sorted according to their likelihood measures of association. The degree of association of these combinations served as a feature in our ML system.

3 Predicting Through Machine Learning

In this section, we outline the ML model trained on the extracted features to compute a relatedness score between two sentences. It details the tools and parameters used to build a support vector regressor, which we used to predict a number between 0 and 5, denoting a degree of semantic similarity.

3.1 Model Description

We used a Support Vector Machine (SVM) in order to compute semantic relatedness for both subtasks.

[8] `http://babelnet.org`

[9] `http://lcl.uniroma1.it/adw`

We used LibSVM[10], a library for SVMs developed by Chang and Lin, 2011.

We built a regression model which estimates a continuous score between 0 and 5 for each sentence pair. The values of C and γ have been optimised through a grid-search which uses a 5-fold cross-validation method, and all systems use an RBF kernel.

The system for Subtask 2a (English) is trained on a combination of training and trial data provided by the 2012, 2013 and 2014 SemEval tasks. We used these datasets to form a training set of 9750 sentence pairs combining the different domains covered by the STS task: image description (image), news headlines (headlines), student answers paired with reference answers (answers-students), answers to questions posted in stach exchange forums (answers-forum), English discussion forum data exhibiting committed belief (belief). However, the training set for Subtask 2b (Spanish) was much smaller, at only 804 sentence pairs collected by combining previous datasets from the Newswire and Wikipedia domains.

3.2 Results and Analysis

The task required the submission of 3 different runs for each task. The runs for the Subtask 2a (English) were identical except for some parameter differences for the SVM training. Our system performed adequately, with our primary run achieving a mean Pearson Correlation of 0.7216.

However, the runs for Subtask 2b (Spanish) were trained on different training sets. Run-1 and Run-2 are trained on the 804 Spanish sentence-pairs. The Spanish set's Run-3, however, is trained on the much larger English training set. For this purpose, we needed to translate the Spanish test set into English in order to use the Semantic Similarity language-dependent features (see sections 2.1.3 and 2.2.4). This system did not outperform the basic Spanish model used in Run-1 and Run-2, despite the much larger training set. Our Spanish system did not yield a satisfactory performance, achieving a Pearson Correlation score of only 0.5158. This could be part due to the smaller training set in Spanish, and the imperfect translations into English which consequently influenced the performance of the language-dependent features. The detailed results for both tasks are given in Table 1 and 2.

	Run-1	**Run-2**	**Run-3**
answers-forums	0.6781	0.6454	0.6179
answers-students	0.7304	0.7093	0.6977
belief	0.6294	0.5165	0.3236
headlines	0.6912	0.6084	0.5775
images	0.8109	0.7999	0.7954
mean	***0.7216***	***0.6746***	***0.6353***
rank (out of 74)	***33***	***45***	***55***

Table 1: Task 2a – Pearson Correlation for English.

	Run-1	**Run-2**	**Run-3**
wikipedia	0.5239	0.4671	0.4402
newswire	0.5076	0.5437	0.5524
mean	***0.5158***	***0.5054***	***0.4963***
rank (out of 17)	***9***	***10***	***11***

Table 2: Task 2b – Pearson Correlation for Spanish.

4 Conclusion and Future Work

We have presented an efficient approach to calculate semantic relatedness for both English and Spanish sentence pairs. We used the same feature set for both tasks, even though it meant translating the Spanish sentences into English before extracting one of the features (i.e. the Semantic Similarity). The system did not performed well for Spanish as it ranked 9 out of 17, with a 0.5158 average Person correlation over two test sets (0.1747 correlation points less than the best submitted run). On the other hand, it performed reasonably well for English, where the system's best result ranked 33 among 74 submitted runs with 0.7216 Pearson correlation over five test sets (only 0.0799 correlation points less than the best submitted run).

In the future we plan to extract the conceptual description provided by the BabelNet network in order to match it with the conceptual terms. We have not done that for now because we need to treat these descriptions as sentences, which requires filtering out the noise produced by them.

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Acknowledgements

Hanna Béchara, Hernani Costa and Rohit Gupta are supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); and the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017).

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. In *15th Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal. Springer.

Rohit Gupta, Hanna Bechara, Ismail El Maarouf, and Constantin Orasan. 2014. UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In *8th Int. Workshop on Semantic Evaluation (SemEval'14)*, pages 785–789, Dublin, Ireland. ACL and Dublin City University.

Zelig Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Adam Kilgarriff. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Conf. of the North American Chapter of the ACL on Human Language Technology - Volume 1*, NAACL'03, pages 48–54. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *15th Int. Conf. on Machine Learning*, ICML'98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP'02, pages 62–69. ACL.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Franz Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *41st Annual Meeting on ACL - Volume 1*, ACL'03, pages 160–167. ACL.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *51st Annual Meeting of the ACL - Volume 1*, pages 1341–1351, Sofia, Bulgaria. ACL.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword Expressions in the Wild?: The Mwetoolkit Comes in Handy. In *23rd Int. Conf. on Computational Linguistics: Demonstrations*, COLING'10, pages 57–60. ACL.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer*

Society Technical Committee on Data Engineering, 24(4):35–42.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *7^{th} Int. Conf. on Spoken Language Processing*, ICSLP'02, pages 901–904.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAAC 2003*, pages 252–259, Edmonton, Canada. ACL.

Peter D. Turney. 2001. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *12^{th} European Conf. on Machine Learning*, EMCL'01, pages 491–502, London, UK. Springer.