UMCC_DLSI_SemSim: Multilingual System for Measuring Semantic Textual Similarity

Alexander Chávez
Héctor Dávila
DI, University of Matanzas, Cuba.
{alexander.chavez,
hector.davila}@umcc.cu

Yoan Gutiérrez
Antonio Fernández-Orquín
Andrés Montoyo
Rafael Muñoz
DLSI, University of Alicante, Spain.
{ygutierrez
montoyo,rafael}@dlsi.ua.es,
antonybr@yahoo.com

Abstract

In this paper we describe the specifications and results of UMCC_DLSI system, which was involved in Semeval-2014 addressing two subtasks of Semantic Textual Similarity (STS, Task 10, for English and Spanish), and one subtask of Cross-Level Semantic Similarity (Task 3). As a supervised system, it was provided by different types of lexical and semantic features to train a classifier which was used to decide the correct answers for distinct subtasks. These features were obtained applying the Hungarian algorithm over a semantic network to create semantic alignments among words. Regarding the Spanish subtask of Task 10 two runs were submitted, where our Run2 was the best ranked with a general correlation of 0.807. However, for English subtask our best run (Run1 of our 3 runs) reached 16th place of 38 of the official ranking, obtaining a general correlation of 0.682. In terms of Task 3, only addressing Paragraph to Sentence subtask, our best run (Run1 of 2 runs) obtained a correlation value of 0.760 reaching 3rd place of 34.

1 Introduction

Many applications of language processing rely on measures of proximity or remoteness of various kinds of linguistic units (words, meanings, sentences, documents). Thus, issues such as disambiguation of meanings, detection of lexical chains, establishing relationships between documents, clustering, etc., require accurate similarity measures.

The problem of formalizing and quantifying an intuitive notion of similarity has a long history in philosophy, psychology, artificial intelligence, and through the years has followed many different perspectives (Hirst, 2001). Recent research in the field of Computational Linguistics has emphasized the perspective of semantic relations between two lexemes in a lexical resource, or its inverse, semantic distance. The similarity of sentences is a confidence score that reflects the relationship between the meanings of two sentences. This similarity has been addressed in the literature with terminologies such as affinity, proximity, distance, difference and divergence (Jenhani, et al., 2007). The different applications of text similarity have been separated into a group of similarity tasks: between two long texts, for document classification; between a short text with a long text, for Web search; and between two short texts, for paraphrase recognition, automatic machine translation, etc. (Han, et al., 2013).

At present, the calculation of the similarity between texts has been tackled from different points of views. Some have opted for a single measure to capture all the features of texts and other models have been trained with various measures to take text features separately. In this work, we addressed the combination of several measures using a Supervised Machine Learning (SVM) approach. Moreover, we intend to introduce a new approach to calculate textual similarities using a knowledge-based system, which is based on a set of cases composed by a vector with values of several measures. We also combined both approaches.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 716–721,
Dublin, Ireland, August 23-24, 2014.

After this introduction, the rest of the paper is organized as follows. Section 2 shows the Pre-processing stage. Subsequently, in Section 3 we show the different features used in our system. In Section 4 we describe our knowledge-based system. Tasks and runs are provided in Section 5. Finally, the conclusions and further work can be found in Section 6.

2 Pre-processing

Below are listed the pre-processing steps performed by our system. In bold we emphasize some cases which were used in different tasks.

- All brackets were removed.
- The abbreviations were expanded to their respective meanings. It was applied using a list of the most common abbreviations in English, with 819 and Spanish with 473. Phrases like "The G8" and "The Group of Eight" are detected as identical.
- Deletion of hyphen to identify words forms. For example, "well-studied" was replaced by "well studied". Example taken from line 13 of MSRpar corpus in test set of Semeval STS 2012 (Agirre, et al., 2012).
- The sentences were tokenized and POS-tagged using Freeling 3.0 (Padró and Stanilovsky, 2012).
- All contractions were expanded. For example: *n't*, *'m*and *'s*. In the case of *'s* was replaced with "is" or "of", "Tom's bad" to "Tom is bad" and "Tom's child" by "Child of Tom". (**Only for English tasks**).
- Punctuation marks were removed from the tokens except for the decimal point in numbers.
- Stop words were removed. We used a list of the most common stop words. (28 for English and 48 for Spanish).
- The words were mapped to the most common sense of WordNet 3.0. (**Only for Spanish task**).
- A syntactic tree was built for every sentence using Freeling 3.0.

3 Features Extraction

Measures of semantic similarity have been traditionally used between words or concepts, and much less between text segments, (i.e. two or more words). The emphasis on word to word similarity is probably due to the availability of resources that specifically encode relations between words or concepts (e.g. WordNet) (Mihalcea, et al., 2006). Following we describe the similarity measures used in this approach.

3.1 Semantic Similarity of Words

A relatively large number of word to word similarity metrics have previously been proposed in the literature, ranging from distance-oriented measures computed on semantic networks, to metrics based on models of distributional similarity learned from large text collections (Mihalcea, et al., 2006).

3.2 Corpus-based Measures

Corpus-based measures of word semantic similarity try to identify the degree of similarity between words using information exclusively derived from large corpora (Mihalcea, et al., 2006). We considered one metric named Latent Semantic Analysis (LSA) (Landauer, et al., 1998).

Latent Semantic Analysis: The Latent semantic analysis (LSA) is a corpus/document based measure proposed by Landauer in 1998. In LSA term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by singular value decomposition (SVD) on the term-by-document matrix *T* representing the corpus (Mihalcea, et al., 2006). There is a variation of LSA called HAL (*Hyperspace Analog to Language*) (Burgess, et al., 1998) that is based on the co-occurrence of words in a common context. The variation consists of counting the number of occurrences in that two words appear at n^1 distance (called windows).

For the co-occurrence matrix of words we took as core the UMBC WebBase corpus[2] (Han, et al., 2013), which is derived from the Stanford WebBase project[3]. For the calculation of HAL measure we used the Cosine Similarity between the vectors for each pair of words.

[1] The windows is the number of intermediate words between two words.

[2] Dataset of high quality English paragraphs containing over three billion words and it is available in http://ebiquity.umbc.edu/resource/html/id/351

[3] Stanford WebBase 2001. http://bit.ly/WebBase.

3.3 Knowledge-based Measures

There are many measures developed to quantify the degree of semantic relation between two words senses using semantic network information. For example:

Leacock & Chodorow Similarity: The Leacock & Chodorow (LC) similarity is determined as follows:

$$Sim_{lc} = -\log\left(\frac{length}{2*D}\right) \quad (1)$$

Where *length* is the length of the shortest path between senses using node-counting and *D* is the maximum depth of the taxonomy (Leacock and Chodorow, 1998)

Wu and Palmer: The Wu and Palmer similarity metric (Wup) measures the depth of two given senses in the WordNet taxonomy, and the depth of the least common subsumer (LCS), and combine them into a similarity score (Wu and Palmer, 1994):

$$Sim_{Wup} = \frac{2*depth(LCS)}{depth(sense_1)+depth(sense_2)} \quad (2)$$

Resnik: The Resnik similarity (Res) returns the information content (IC) of the LCS of two senses:

$$Sim_{Res} = IC(LCS) \quad (3)$$

Where IC is defined as:

$$IC(c) = -\log P(c) \quad (4)$$

And *P(c)* is the probability of encountering an instance of sense *c* in a large corpus (Resnik, 1995) (Mihalcea, et al., 2006).

Lin: The Lin similarity builds on Resnik's measure and adds a normalization factor consisting of the information content of the two inputs senses (Lin, 1998):

$$Sim_{Lin} = \frac{2*IC(LCS)}{IC(sense)+IC(sense_2)} \quad (5)$$

Jiang & Conrath: The Jiang and Conrath similarity (JC) is defined as follows (Jiang and Conrath, 1997):

$$Sim_{jc} = \frac{1}{IC(sense_1)+IC(sense_2)-2*IC(LCS)} \quad (6)$$

PathLen: The PathLen similarity (Len) involves the path lengths between two senses in the taxonomy (Pedersen, et al., 2004).

$$Sim_{Path} = -\log pathlen(sense_1, sense_1) \quad (7)$$

Where $pathlen(sense_1, sense_1)$ is the number of edges in the shortest path between $sense_1$and $sense_2$.

Word Similarity: In order to calculate the similarity between two words (WS) we used the following expression:

$$WS(w1,w2) = \max_{\substack{s1 \in senses(w1) \\ s2 \in senses(w2)}} sim(s1,s2) \quad (8)$$

Where $sim(s1,s2)$ is one of the similarity metrics at sense level previously described.

3.4 Lexical Features

We used a well-known lexical attributes similarity measures based on distances between strings. Dice-Similarity, Euclidean-Distance, Jaccard-Similarity, Jaro-Winkler, Levenstein Distance, Overlap-Coefficient, QGrams Distance, Smith-Waterman, Smith-Waterman-Gotoh, Smith-Waterman-Gotoh-Windowed-Affine.

These metrics have been obtained from an API (Application Program Interface) SimMetrics library v1.5 for.NET[4] 2.0.

3.5 Word Similarity Models

With the purpose of calculating the similarity between two words, we developed two models involving the previous word similarity metrics. These were defined as:

Max Word Similarity: The Max Word Similarity (MaxSim) is defined as follows:

$$MaxSim(w1,w2) = \begin{cases} 1 & if QGDistance(w1,w2) = 1 \\ Max\left(Sim_{Hal}(w1,w2), Sim_{Wup}(w1,w2)\right) & \end{cases} \quad (9)$$

Where $QGDistance(w1,w2)$ is the QGram-Distance between *w1* and *w2*.

Statistics **and Weight Ratio:** For calculating the weight ratio in this measure of similarity was used WordNet 3.0 and it was defined in (10):

$$StaWeiRat\ (w1,w2) = \frac{\left(Sim_{Hal}(w1,w2) + \left(\frac{1}{WeiRat(w1,w2)}\right)\right)}{2} \quad (10)$$

[4] Copyright (c) 2006 by Chris Parkinson, available in http://sourceforge.net/projects/simmetrics/

Where $WeiRat(w1, w2)$ takes a value based on the type of relationship between *w1* and *w2*. The possible values are defined in Table 1.

Value	Relation between $w1$ and $w2$
10	Antonym.
1	Synonym.
2	Direct Hypernym, Similar_To or Derivationally Related Form.
3	Two-links indirect Hypernym, Similar_To or Derivationally Related Form.
3	One word is often found in the gloss of the other.
9	Otherwise.

Table 1: Values of Weight Ratio.

3.6 Sentence Alignment

In the recognition of texts' similarities, several methods of lexical alignment have been used and can be appreciated by different point of views (Brockett, 2007) (Dagan, et al., 2005). Glickman (2006) used the measurement of the overleap grade between bags of words as a form of sentence alignment. Rada et al. (2006) made reference to an all-for-all alignment, leaving open the possibility when the same word of a sentence is aligned with several sentences. For this task, we used the Hungarian assignment algorithm as a way to align two sentences (Kuhn, 1955). Using that, the alignment cost between the sentences was reduced. To increase the semantic possibilities we used all word similarity metrics (including the two word similarity models) as a function cost.

3.7 N-Grams Alignment

Using the Max Word Similarity model, we calculated three features based on 2-gram, 3-gram and 4-gram alignment with the Hungarian algorithm.

4 Knowledge-based System

For similarity calculation between two phrases, we developed a knowledge-based system using SemEval-2012, SemEval-2013 and SemEval-2014 training corpus (Task 10 and Task 1 for the last one). For each training pair of phrases we obtained a vector with all measures explained above. Having it, we estimated the similarity value between two new phrases by applying the Euclidian distance between the new vector (made with the sentence pair we want to estimate the similarity value) and each vector in the training corpus. Then, the value of the instance with minor Euclidian Distance was assigned to the new pair of phrases.

5 Tasks and runs

Our system participated in Sentence to Phrase subtask of Task 3: "Cross-Level Semantic Similarity" (Jurgens, et al., 2014) and in two subtasks of Task 10: "Multilingual Semantic Textual Similarity" of SemEval-2014. It is important to remark that our system, using SVM approach, did not participate in Task 1: "Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment", due to deadline issues. We compared our system results with the final ranking of Task 1 and we could have reached the 6th place of the ranking for Relatedness Subtask with a 0.781 of correlation coefficient, and the 9th place for Entailment Subtask with an accuracy of 77.41%.

	Task 10 Sp		Task 10 En			Task 3 Sentence to Phrase	
Features/Runs	1	2	1	2	3	1	2
PathLenAlign	x		x	x		x	x
ResAlign	x		x	x		x	x
LcAlign	x		x	x		x	x
WupAlign	x		x	x		x	x
Res	x		x	x		x	x
Lc	x		x	x		x	x
DiceSimilarity	x	x	x	x		x	x
EuclideanDistance	x	x	x	x		x	x
JaccardSimilarity	x	x	x	x		x	x
JaroWinkler	x	x	x	x		x	x
Levenstein	x	x	x	x		x	x
Overlap-Coefficient	x	x	x	x		x	x
QGramsDistance	x	x	x	x		x	x
SmithWaterman	x	x	x	x		x	x
SmithWatermanGotoh	x	x	x	x		x	x
SmithWatermanGotoh-WindowedAffine	x	x	x	x		x	x
BiGramAlingHungMax	x		x	x		x	x
TriGramAlingHungMax	x		x	x		x	x
TetraGramAlingHungMax	x		x	x		x	x
WordAlingHungStatWeigthRatio	x		x	x		x	x
SentenceLengthPhrase1	x		x	x		x	x
SentenceLengthPhrase2	x		x	x		x	x

Table 2: Features and runs. Spanish (Sp) and English (En).

In Table 2 is important to remark the following situations:

- In Task 10 Spanish (two runs), we used the training corpus of Task 10 English.

- In Run2 of Task 10 English, the similarity score was replaced for the knowledge-based system value if Euclidean Distance of the most similar case was less than 0.30.
- Run3 of Task 10 English was a knowledge-based system.
- In Run1 of Sentence to Phrase of Task 3, we trained the SVM model using only the training corpus of this task.
- In Run2 of Sentence to Phrase of Task 3, we trained the SVM model using the training corpus of this task and the training corpus of Task 10 English.

6 Conclusion

In this paper we introduced a new framework for recognizing Semantic Textual Similarity, involving feature extraction for SVM model and a knowledge-based system. We analyzed different ways to estimate textual similarities applying this framework. We can see in Table 3 that all runs obtained encouraging results. Our best run was first position of the ranking for task 10 (Spanish) and other important positions were reached in the others subtasks. According to our participation, we used a SVM which works with features extracted from six different strategies: String-Based Similarity Measures, Semantic Similarity Measures, Lexical-Semantic Alignment, Statistical Similarity Measures and Semantic Alignment. Finally, we can conclude that our system achieved important results and it is able to be applied on different scenarios, such as task 10, task 3.1 and task 1. See Table 3 and the beginning of Section 5.

Subtask	Run	SemEval-2014 Position
Task 10-Spanish	Run1	4
	Run2	1
Task 10-English	Run1	16
	Run2	18
	Run3	33
Task-3	Run1	3
	Run2	16

Table 3: SemEval-2014 results.

As further work, we plan to analyze the main differences between task 10 for Spanish and English in order to homogenise both system's results.

Acknowledgments

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects, "Tratamiento inteligente de la información para la ayuda a la toma de decisiones" (GRE12-44), ATTOS (TIN2012-38536-C03-03), LEGOLANG (TIN2012-31224), SAM (FP7-611312), FIRST (FP7-287607) and ACOMP/2013/067.

Reference

Eneko Agirre, Mona Diab, Daniel Cer and Aitor Gonzalez-Agirre, 2012. *SemEval 2012 Task 6: A Pilot on Semantic Textual Similarity.*. s.l., First Join Conference on Lexical and Computational Semantic (*SEM), Montréal, Canada. 2012., pp. 385-393.

Chris Brockett, 2007. *Aligning the RTE 2006 Corpus.* Microsoft Research, p. 14.

Curt Burgess, Kay Livesay and Kevin Lund, 1998. *Explorations in Context Space: Words, Sentences, Discourse.* Discourse Processes, Issue 25, pp. 211 - 257.

Ido Dagan, Oren Glickman and Bernardo Magnini, 2005. *The PASCAL Recognising Textual Entailment Challenge.* En: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment.

Oren Glickman, Ido Dagan and Moshe Koppel, 2006. *A Lexical Alignment Model for Probabilistic Textual Entailment.* In: Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. Southampton, UK: Springer-Verlag, pp. 287--298.

Lushan Han et al., 2013. *UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems.* s.l., s.n.

Alexander B. Hirst and Graeme, 2001. *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures.*

Ilyes Jenhani, Nahla Ben Amor and Zi Elouedi, 2007. *Information Affinity: A New Similarity Measure for Possibilistic Uncertain Information.* En: Symbolic and Quantitative Approaches to Reasoning with Uncertainty. s.l.:Springer Berlin Heidelberg, pp. 840-852.

Jay Jiang and David Conrath, 1997. *Semantic similarity based on corpus statistics and lexical taxonomy.* s.l., Proceedings of the International Conference on Research in Computational Linguistics.

David Jurgens, Mohammad Taher and Roberto Navigli, 2014. *SemEval-2014 Task 3: Cross-*

Level Semantic Similarity. Dublin, Ireland, In Proceedings of the 8th International Workshop on Semantic Evaluation., pp. 23-24.

Harold W. Kuhn, 1955. *The Hungarian Method for the assignment problem.* Naval Research Logistics Quarterly.

Thomas K. Landauer, Peter W. Foltz and Darrell Laham, 1998. *Introduction to latent semantic analysis.* Discourse Processes, Issue 25, pp. 259-284.

Claudia Leacock and Martin Chodorow, 1998. *Combining local context and WordNet sense similarity for word sense identification.* s.l.:s.n.

Lin Dekang, 1998. *An information-theoretic definition of similarity.* s.l., Proceedings of the International Conf. on Machine Learning.

Rada Mihalcea, Courtney Corley and Carlo Strapparava, 2006. *Corpus-based and knowledge-based measures of text semantic similarity.* In: IN AAAI'06. s.l.:21st National Conference on Artificial Intelligence, pp. 775--780.

Luís Padró and Evgeny Stanilovsky, 2012. *FreeLing 3.0: Towards Wider Multilinguality.* Istanbul, Turkey, Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.

Ted Pedersen, Siddharth Patwardhan and Jason Michelizzi, 2004. *WordNet::Similarity - Measuring the Relatedness of Concepts.* American Association for Artificial Intelligence.

Philip Resnik, 1995. *Using information content to evaluate semantic similarity.* s.l., Proceedings of the 14th International Joint Conference on Artificial Intelligence.

Zhibiao Wu and Martha Palmer, 1994. *Verb semantics and lexical selection.*