UMCC_DLSI-(EPS): Paraphrases Detection Based on Semantic Distance

Héctor Dávila, Antonio Fernández Orquín, Alexander Chávez, Yoan Gutiérrez, Armando Collazo, José I. Abreu
DI, University of Matanzas
Autopista a Varadero km 3 ½
Matanzas, Cuba.
{hector.davila, tony, alexander.chavez, yoan.gutierrez, armando.collazo, jose.abreu}@umcc.cu

Andrés Montoyo, Rafael Muñoz
DLSI, University of Alicante Carretera de San Vicente S/N Alicante, Spain.
{montoyo, rafael}@dlsi.ua.es

Abstract

This paper describes the specifications and results of UMCC_DLSI-(EPS) system, which participated in the first Evaluating Phrasal Semantics of SemEval-2013. Our supervised system uses different kinds of semantic features to train a bagging classifier used to select the correct similarity option. Related to the different features we can highlight the resource WordNet used to extract semantic relations among words and the use of different algorithms to establish semantic similarities. Our system obtains promising results with a precision value around 78% for the English corpus and 71.84% for the Italian corpus.

1 Introduction

It is well known finding words similarity, even when it is lexical or semantic can improve entailment recognition and paraphrase identification; and ultimately lead to improvements in a wide range of applications in Natural Language Processing (NLP). Several areas like question answering, query expansion, information retrieval, and many others, depend on phrasal semantics (PS). PS, is concerned with how the meaning of a sentence is composed both from the meaning of the constituent words, and from extra meaning contained within the structural organization of the sentence itself (Dominey, 2005).

The aim of SemEval 2013 competition is also discovering similarity, specifically in Evaluating Phrasal Semantics (EPS). The goal of this task is to evaluate how well systems can judge the semantic similarity of a word and a short sequence of words. That is, given a set of pairs of this type; classify it on negative (if the meaning of the word is semantically different to the meaning of the sequence) or positive (if the meaning of the sequence, as a whole, is semantically close to the meaning of the word).

Based on this, we developed a system capable to detect if two phrases are semantically close.
The rest of this paper, specifically section 2 is a brief Related Work. Section 3 describes the system architecture and our run. Continuing with section 4 we describe the training phase. Following that, section 5 presents the results and discussion for our Machine Learning System. Finally we conclude and propose our future works (Section 6).

2 Related Work

There have been many WordNet-based similarity measures, among other highlights the work of researchers like (Budanitsky and Hirst, 2006; Leacock and Chodorow, 1998; Mihalcea *et al.*, 2006; Richardson *et al.*, 1994).

On the other hand, WordNet::Similarity[1] (Pedersen *et al.*, 2004) has been used by other researchers in an interesting array of domains. WordNet::Similarity implements measures of similarity and relatedness between a pair of concepts (or synsets[2]) based on the structure and content of WordNet. According to (Pedersen *et al.*, 2004), three of the six measures of similarity are based on the information content of the least

[1] http://sourceforge.net/projects/wn-similarity/
[2] A group of English words into sets of synonyms.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 93–97, Atlanta, Georgia, June 14-15, 2013.

common subsumer (LCS). These measures include res (Resnik, 1995), lin (Lin, 1998), and jcn (Jiang and Conrath, 1997).

Pursuant to Pedersen, there are three other similarity measures based on path lengths between a pair of concepts: lch (Leacock and Chodorow, 1998), wup (Wu and Palmer, 1994), and path.

Our proposal differs from those of WordNet::Similarity and other measures of similarity in the way we selected the relevant WordNet relations (see section 3.2 for detail). Unlike others, our measure assign weight to WordNet relations (any we consider relevant) depending to the place they occupy in the minimum path and the previously visited relations.

Besides these, the novelty of our approach is using the weights as a function of semantic relations in a minimal distance path and also the method we used to arrive to those weight functions or rules.

3 System Architecture and description of the run

As we can see in Figure 1 our run begin with the pre-processing of SemEval 2013's training set. Every sentence pair is tokenized, lemmatized and POS-tagged using Freeling 2.2 tool (Atserias *et al.*, 2006). Afterwards, several methods and algorithms are applied in order to extract all features for our Machine Learning System (MLS). The system trains the classifier using a model based on bagging (using JRip[3]). The training corpus has been provided by SemEval-2013 competition, in concrete by the EPS task. As a result, we obtain a trained model capable to detect if one phrase implies other. Finally, we test our system with the SemEval 2013 test set (see Table 2 with the results of our run). The following section describes the features extraction process.

3.1 Description of the features used in the Machine Learning System

In order to detect entailment between a pair of phrases, we developed an algorithm that searches a semantic distance, according to WordNet (Miller *et al.*, 1990), between each word in the first phrase with each one in the second phrase.

We used four features which intend to measure the level of proximity between both sentences:

- The minimum distance to align the first phrase with the second (MinDist). See section 3.2 for details.
- The maximal distance to align the first phrase with the second (MaxDist).
- The average of all distances results to align the first phrase with the second one. (AverageDistance).
- The absolute relative error of all distances results to align the first phrase with the second respect to the average of them.

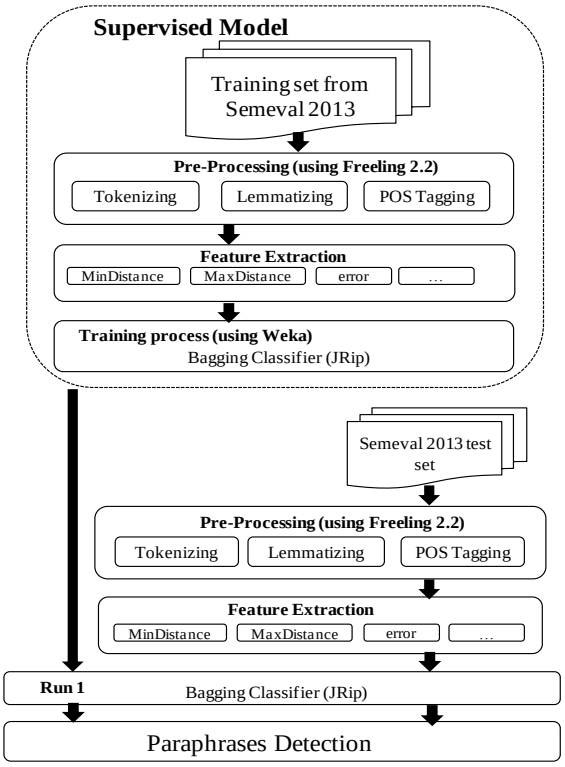

Figure 1. System Architecture.

Other features included are the most frequent relations contained in the shorted path of the minimum distance; result to align the first phrase with the second one. Following table shows the relations selected as most frequent.

A weight was added to each of them, according to the place it occupy in the shortest path between two synsets. The shortest path was calculated using Breadth -First-Search algorithm (BFS) (Cormen *et al.*, 2001).

In addition, there is one feature that takes into account any other relationship that is not previously considered.

Finally, as a result we obtain 22 features from this alignment method.

[3] JRip is an inference and rules-based learner.

Relation	Weight (*W* function)
Antonym	1000
Synonym	0
Hyponym/ Hypernym	100 if exist an antonym before, 30 if exist other relation before (except synonym, hyponym, hypernym), 5 otherwise.
Meber_Holonym/ PartHolonym	100 if exist an antonym before, 20 if exist a hyponym or a hypernym,10 otherwise.
Cause/ Entailment	100 if exist an antonym before, 2 otherwise.
Similar_To	100 if exist an antonym before, 3 otherwise.
Attribute	100 if exist an antonym before, 8 otherwise.
Also_See	100 if exist an antonym before, 10 otherwise.
Derivationaly_Related_Form	100 if exist an antonym before, 5 otherwise.
Domain_Of_Synset_Topic	100 if exist an antonym before, 13 otherwise.
Domain_Of_Synset_Usage	100 if exist an antonym before, 60 otherwise.
Member_Of_Domain_Topic	100 if exist an antonym before, 13 otherwise.
Member_Of_Domain_Usage	100 if exist an antonym before, 60 otherwise.
Other	100

Table 1. Most frequents relations with their weight.

3.2 Semantic Distance

As aforementioned, our distance depends on calculating the similarity between sentences, based on the analysis of WordNet relations, and we only took into account the most frequent ones. When searching the shortest path between two WordNet synsets, frequents relations were considered the ones extracted according to the analysis made in the training corpus, provided by SemEval-2013.

The distance between two synsets is calculated with the relations found; and simply it is the sum of the weights assigned to each connection.

$$MinDistP(P,Q) = MinDistS(P_X, Q_Y), \forall\ (X,Y) \quad (1)$$

$$MinDistS(X,Y) = Min(X_i, Y_j), \forall (i,j) \quad (2)$$

$$Min(X_i; Y_j) = \sum_{k=0}^{k=m} W(Rel(L[k], L[k+1])) \quad (3)$$

$$L = BFS(X_i; Y_j) \quad (4)$$

Where i and j represents the i-th and j-th sense of the word; P and Q represents words collections; P_X is the X-th word of P; Q_Y is the Y-th word of Q; $MinDistP$ obtains a value that represents a minimal semantic distance across WordNet (Miller *et al.*, 2006) resource (this resource is involved into the integrator resource, ISR-WN (Gutiérrez *et al.*, 2011a; 2010a); $MinDistS$ the minimal semantic distance between two words; Min represents the minimal semantic distance between two senses collections; L is a collection of synsets that represents the minimal path between two synsets using BFS; Rel obtains semantic relation types between two synsets; W is a functions that apply the rules described in Table 1. The maximum and average distance is calculated in a similar fashion but using the maximum and average instead of the minimum.

3.3 Semantic Alignment

First, the two sentences are pre-processed with Freeling 2.2 and the words are classified according to their parts-of-speech. Then, all senses of every word are taken and treated as a group. Distance between two groups will be the minimal distance (described in 3.1) between senses of any pair of words belonging to the group.

In the example of Figure 2, Dist=280 is selected for the pair "Balance-Culture" (minimal cost).

Following the explanation on section 3.1 we extract the features guided to measure the level of proximity between both sentences.

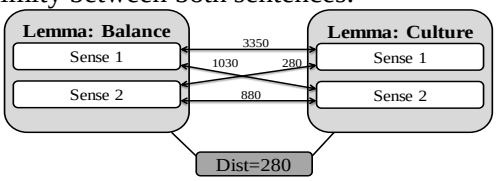

Figure 2. Distance between "Balance" and "Culture".

A maximum and average distance is calculated in a similar fashion, but using the maximum and average instead of the minimum.

4 Description of the training phase

For the training process, we used a supervised learning framework (based on Weka[4]), including all the training set (positive and negative instances) as a training corpus. We conduct several experiments in order to select the correct classifier, the best result being obtained with a model based on bagging (using JRip algorithm). Finally, we used 10-fold cross validation technique with the selected classifier, obtaining a classification value of 73.21%.

[4] http://prdownloads.sourceforge.net/weka/

5 Results and discussion

EPS task of SemEval-2013 offered many official measures to rank the systems. Some of them are the following:

- F-Measure (FM): Correct Response (CR), Instances correctly classified, True positives (TP), Instances correctly classified as positive. False Positives (FP), Instances incorrectly classified as positive, True Negatives (TN), Instances correctly classified as negative, False Negatives (FN), Instances incorrectly classified as negative.

Corpus	FM	CR	TP	FP	TN	FN
English	0.6892	2826	1198	325	1628	755
Italian	0.6396	574	245	96	329	180

Table 2. Official SemEval 2013 results.

The behavior of our system, for English and Italian corpus is shown in Table 2.

The only thing that changes to process the Italian corpus is that Freeling is used as input to identify Italian words and it returns the English WN synsets. The process continues in the same way as English.

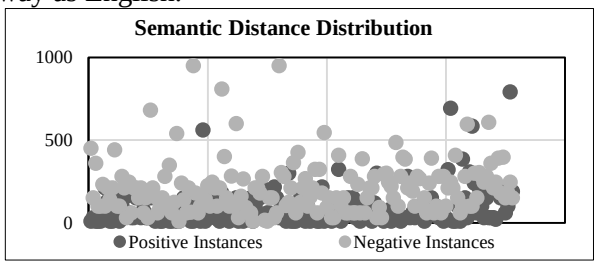

Figure 3: Semantic Distance distribution between negative and positive instances.

As shown in Table 2, our main drawback is to classify positive instances. Sometimes, the distance between positive phrases is very far. This is due to the relations found in the minimum path are very similar to the one found in other pairs of negatives instances; this can be the cause of our MLS classifies them as negatives (see Figure 3).

Figure 3 shows a distributional graphics that take a sample of 200 negative and positive instances. The graphics illustrate how close to zero value the positive instances are, while the negatives are far away from this value. However, in the approximate range between 80 and 200, we can see values of positive and negative instances positioning together. This can be the cause that our MLS misclassified some positive instances as negative.

6 Conclusion and future work

This paper introduced a new framework for EPS, which depends on the extraction of several features from WordNet relations. We have conducted the semantic features extraction in a multidimensional context using the resource ISR-WN(Gutiérrez *et al.*, 2010a).

Our semantic distance provides an appealing approach for dealing with phrasal detection based on WordNet relation. Our team reached the sixth position of ten runs for English corpus, with a small difference of 0.07 points compared to the best results with respect to accuracy parameter.

Despite the problems caused by poorly selected positive instances, our distance (labeled as Our) obtained very similar results to those obtained by the best team (labeled as First[5]), which indicates that our work is well underway (see Table 3 for details).

Team	accuracy	recall	precision
First	0.802611	0.751664	0.836944128
Our	0.723502	0.613415	0.786605384

Table 3. Comparative results (English corpus).

It is important to remark that our system has been the only competitor to evaluate Italian texts. It has been possible due to our system include Freeling in the preprocessing stage.

Our future work will aim to resolve instances misclassified by our algorithm. In addition, we will introduce lexical substitutions (synonyms) to expand the corpus, we will also apply conceptual semantic similarity using relevant semantic trees (Gutiérrez *et al.*, 2010b; Gutiérrez *et al.*, 2011b).

Acknowledgments

This research work has been partially funded by the Spanish Government through the project TEXT-MESS 2.0 (TIN2009-13391-C04), "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03) and "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224); and by the Valencian Government through the project PROMETEO (PROMETEO/2009/199).

References

Atserias, J.; B. Casas; E. Comelles; M. González; L. Padró and M. Padró. FreeLing 1.3: Syntactic and

[5] christian_wartena. Team HsH.

semantic services in an open-source NLP library. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), 2006. 48-55 p.

Budanitsky, A. and G. Hirst Evaluating wordnet-based measures of lexical semantic relatedness Computational Linguistics, 2006, 32(1): 13-47.

Cormen, T. H.; C. E. Leiserson; R. L. Rivest and C. Stein. Introduction to algorithms. MIT press, 2001. 0262032937.

Dominey, P. F. Aspects of descriptive, referential, and information structure in phrasal semantics: A construction-based model Interaction Studies, 2005, 6(2): 287-310.

Gutiérrez, Y.; A. Fernández; A. Montoyo and S. Vázquez. Integration of semantic resources based on WordNet. XXVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Universidad Politécnica de Valencia, Valencia, SEPLN 2010, 2010a. 161-168 p. 1135-5948.

Gutiérrez, Y.; A. Fernández; A. Montoyo and S. Vázquez. UMCC-DLSI: Integrative resource for disambiguation task. Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, Association for Computational Linguistics, 2010b. 427-432 p.

Gutiérrez, Y.; A. Fernández; A. Montoyo and S. Vázquez Enriching the Integration of Semantic Resources based on WordNet Procesamiento del Lenguaje Natural, 2011a, 47: 249-257.

Gutiérrez, Y.; S. Vázquez and A. Montoyo. Improving WSD using ISR-WN with Relevant Semantic Trees and SemCor Senses Frequency. Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, RANLP 2011 Organising Committee, 2011b. 233--239 p.

Jiang, J. J. and D. W. Conrath Semantic similarity based on corpus statistics and lexical taxonomy arXiv preprint cmp-lg/9709008, 1997.

Leacock, C. and M. Chodorow Combining local context and WordNet similarity for word sense identification WordNet: An electronic lexical database, 1998, 49(2): 265-283.

Lin, D. An information-theoretic definition of similarity. Proceedings of the 15th international conference on Machine Learning, San Francisco, 1998. 296-304 p.

Mihalcea, R.; C. Corley and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the national conference on artificial intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006. 775 p.

Miller, G. A.; R. Beckwith; C. Fellbaum; D. Gross and K. Miller Introduction to WordNet: An On-line Lexical Database International Journal of Lexicography, 3(4):235-244., 1990.

Miller, G. A.; C. Fellbaum; R. Tengi; P. Wakefield; H. Langone and B. R. Haskell. WordNet a lexical database for the English language. Cognitive Science Laboratory Princeton University 2006.

Pedersen, T.; S. Patwardhan and J. Michelizzi. WordNet:: Similarity: measuring the relatedness of concepts. Demonstration Papers at HLT-NAACL 2004, Association for Computational Linguistics, 2004. 38-41 p.

Resnik, P. Using information content to evaluate semantic similarity in a taxonomy arXiv preprint cmp-lg/9511007, 1995.

Richardson, R.; A. F. Smeaton and J. Murphy. Using WordNet as a knowledge base for measuring semantic similarity between words, Technical Report Working Paper CA-1294, School of Computer Applications, Dublin City University, 1994.

Wu, Z. and M. Palmer. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1994. 133-138 p.