INAOE_UPV-CORE: Extracting Word Associations from Document Corpora to estimate Semantic Textual Similarity

Fernando Sánchez-Vega
Manuel Montes-y-Gómez
Luis Villaseñor-Pineda
Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y
Electrónica (INAOE), Mexico.
{fer.callotl,mmontesg,villasen}
@inaoep.mx

Paolo Rosso

Natural Language Engineering Lab., ELiRF,
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract

This paper presents three methods to evaluate the Semantic Textual Similarity (STS). The first two methods do not require labeled training data; instead, they automatically extract semantic knowledge in the form of word associations from a given reference corpus. Two kinds of word associations are considered: co-occurrence statistics and the similarity of word contexts. The third method was done in collaboration with groups from the Universities of Paris 13, Matanzas and Alicante. It uses several word similarity measures as features in order to construct an accurate prediction model for the STS.

1 Introduction

Even with the current progress of the natural language processing, evaluating the semantic text similarity is an extremely challenging task. Due to the existence of multiple semantic relations among words, the measuring of text similarity is a multifactorial and highly complex task (Turney, 2006).

Despite the difficulty of this task, it remains as one of the most attractive research topics for the NLP community. This is because the evaluation of text similarity is commonly used as an internal module in many different tasks, such as, information retrieval, question answering, document summarization, etc. (Resnik, 1999). Moreover, most of these tasks require determining the "semantic" similarity of texts showing stylistic differences or using polysemicwords (Hliaoutakis et al., 2006).

The most popular approach to evaluate the semantic similarity of words and texts consists in using the semantic knowledge expressed in ontologies (Resnik, 1999); commonly, WorldNet is used for this purpose (Fellbaum, 2005). Unfortunately, despite the great effort that has been the creation of WordNet, it is still far to cover all existing words and senses (Curran, 2003).Therefore, the semantic similarity methods that use this resource tend to reduce their applicability to a restricted domain and to a specific language.

We recognize the necessity of having and using manually-constructed semantic-knowledge sources in order to get precise assessments of the semantic similarity of texts, but, in turn, we also consider that it is possible to obtain good estimations of these similarities using less-expensive, and perhaps broader, information sources. In particular our proposal is to automatically extract the semantic knowledge from large amounts of raw data samples i.e. document corpora without labels.

In this paper we describe two different strategies to compute the semantic similarity of words from a reference corpus. The first strategy uses word co-occurrence statistics. It determines that two words are associated (in meaning) if they tend to be used together, in the same documents or contexts. The second strategy measures the similarity of words by taking into consideration second order word co-occurrences. It defines two words as associated if they are used in similar contexts (i.e., if they co-occur with similar words). The following section describes the implementation of these two strategies for our participation at the STS-SEM 2013 task, as well as their combination with the measures designed by the groups from the Universities of Matanzas, Alicante and Paris 13.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 229–233, Atlanta, Georgia, June 13-14, 2013.

2 Participation in STS-SEM2013

The Semantic Textual Similarity (STS) task consists of estimated the value of semantic similarity between two texts,D_1 and D_2 for now on.

As we mentioned previously, our participation in the STS task of SEM 2013 considered two different approaches that aimed to take advantage of the language knowledge latent in a given reference corpus. By applying simple statistics we obtained a semantic similarity measure between words, and then we used this semantic word similarity (SWS) to get a sentence level similarity estimation. We explored two alternatives for measuring the semantic similarity of words, the first one, called SWS_{ocurr} , uses the co-occurrence of words in a limited context[1],and the second, $SWS_{context}$, compares the contexts of the words using the vector model and cosine similarity to achieve this comparison. It is important to point out that using the vector space model directly, without any spatial transformation as those used by other approaches[2], we could get greater control in the selection of the features used for the extraction of knowledge from the corpus. It is also worth mentioning that we applied a stemming procedure to the sentences to be compared as well as to all documents from the reference corpus. We represented the texts D_1 and D_2 by bags of tokens, which means that our approaches did not take into account the word order.

Following we present our baseline method, then, we introduce the two proposed methods as well as a method done in collaboration with other groups. The idea of this shared-method is to enhance the estimation of the semantic textual similarity by combining different and diverse strategies for computing word similarities.

2.1 STS-baseline method

Given textsD_1 and D_2, their textual similarity is given by:

$$STS - Baseline = MIN(SIM(D_1, D_2), SIM(D_2, D_1))$$

where

$$SIM(D_i, D_j) = \frac{1}{|D_i|} \sum_{t_k \in D_i} 1(t_k \in D_j)$$

This measure is based on a direct matching of tokens. It simply counts the number of tokens from one text D_i that also exist in the other text D_j. Because STS is a symmetrical attribute, unlike Textual Entailment (Agirre et al., 2012), we designed it as a symmetric measure. We assumed that the relationship between both texts is at least equal to their smaller asymmetric similarity.

2.2 The proposed STS methods

These methods incorporate semantic knowledge extracted from a reference corpus. They aim to take advantage of the latent semantic knowledge from a large document collection. Because the extracted knowledge from the reference corpus is at word level, these methods for STS use the same basic –word matching– strategy for comparing the sentences like the baseline method. Nevertheless, they allow a soft matching between words by incorporating information about their semantic similarity.

The following formula shows the proposed modification to the *SIM* function in order to incorporate information of the semantic word similarity (SWS). This modification allowed us not only to match words with exactly the same stem but also to link different but semantically related words.

$$SIM(D_i, D_j) = \sum_{t_m \in D_i} MAX\left(\bigcup_{t_n \in D_j} SWS(t_m, t_n) \right)$$

We propose two different strategies to compute the semantic word similarity (SWS), STS_{Occur} and STS_{Contex} . The following subsections describe in detail these two strategies.

2.2.1 STS based on word co-occurrence

SWS_{Occur} uses a reference corpus to get a numerical approximation of the semantic similarity between two terms $t_i$and t_j (when these terms have not the same stem). As shown in the following formula, SWS_{occur} takes values between 0 and 1; 0 indicates that it does not exist any text sample in the corpus that contains both terms, whereas, 1 indicates that they always occur together.

[1] In the experiments we considered a window (context) formed of 15 surrounding words.

[2]Such as Latent Semantic Analysis (LSA) (Turney, 2005).

$$SWS_{occur}(t_i, t_j) = \begin{cases} t_i = t_j & 1 \\ other & \frac{\#(t_i, t_j)}{MIN(\#(t_i), \#(t_j))} \end{cases}$$

where $\#(t_i, t_j)$ is the number of times that t_i and t_j co-occur and $\#(t_i)$ and $\#(t_j)$ are the number of times that terms t_i and t_j occur in the reference corpus respectively.

2.2.2 STS based on context similarity

$SWS_{context}$ is based on the idea that two terms are semantically closer if they tend to be used in similar contexts. This measure uses the well-known vector space model and cosine similarity to compare the terms' contexts. In a first step, we created a context vector for each term, which captures all the terms that appear around it in the whole reference corpus. Then, we computed the semantic similarity of two terms by the following formula.

$$SWS_{context}(t_i, t_j) = \begin{cases} t_i = t_j & 1 \\ other & SIMCOS(\vec{T}_i, \vec{T}_j) \end{cases}$$

where the cosine similarity, *SIMCOS*, is calculated on the vectors $\vec{T}_i$ and $\vec{T}_j$ corresponding to the vector space model representation of terms t_i and t_j, as indicated in the following equation:

$$SIMCOS(\vec{T}_i, \vec{T}_j) = \frac{\sum_{k \in |V|} t_{ik} \cdot t_{jk}}{|\vec{T}_i| \cdot |\vec{T}_j|}$$

It is important to point out that *SIMCOS* is calculated on a "predefined" vocabulary of interest; the appropriate selection of this vocabulary helps to get a better representation of terms, and, consequently, a more accurate estimation of their semantic similarities.

2.3 STS based on a combination of measures

In addition to our main methods we also developed a method that combines our SWS measures with measures proposed by other two research groups, namely:

- LIPN (Laboratoire d'Informatique de Paris-Nord, Université Paris 13, France).
- UMCC_DLSI (Universidad de Matanzas Camilo Cienfuegos, Cuba, in conjuction with the Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Spain).

The main motivation for this collaboration was to investigate the relevance of using diverse strategies for computing word similarities and the effectiveness of their combination for estimating the semantic similarity of texts.

The proposed method used a set of measures provided by each one of the groups. These measures were employed as features to obtained a prediction model for the STS. Table 1 summarizes the used measures. For the generation and fitting of the model we used three approaches: linear regression, a Gaussian process and a multilayer neural network.

Description	Team	#	Mean Rank	Best Rank
Based on IR measures	LIPN	2	2.0	1
Based on distance on WordNet	LIPN	2	8.5	2
STS-Context	INAOE-UPV	1	4.0	4
Complexity of the sentences	INAOE-UPV	34	27.8	5
STS-Occur	INAOE-UPV	1	7.0	7
Based on the alignment of particulars POS.	UMCC_DLSI	12	40.9	18
n-gram overlap	LIPN	1	20.0	20
Based on Edit distance	UMCC_DLSI	4	42.6	27
Syntactic dependencies overlap	LIPN	1	29.0	29
Levenshtein's distance	LIPN	1	42.0	42
Named entity overlap	LIPN	1	57.0	57

Table 1. General description of the features used by the shared method. The second column indicates the source team for each group of features; the third column indicates the number of used features from each group; the last two columns show the information gain rank of each group of features over the training set.

3 Implementation considerations

The extraction of knowledge for the computation of the *SWS* was performed over the Reuters-21578 collection. This collection was selected because it is a well-known corpus and also because it includes documents covering a wide range of topics.

Due to time and space restrictions we could not consider all the vocabulary from the reference corpus; the vocabulary selection was conducted by taking the best 20,000 words according to the tran-

sition point method (Pinto et al., 2006). This method selects the terms associated to the main topics of the corpus, which presumably contain more information for estimating the semantic similarity of words. We also preserved the vocabulary from the evaluation samples, provided they also occur in the reference corpus. The size of the vocabulary used in the experiments and the size of the corpus and test set vocabularies are shown in Table 2.

Experiment's Vocabulary	Selected Vocabulary	Ref. Corpus Vocabulary	Evaluation Vocabulary
26724	20000	31213	11491

Table 2. Number of different stems from each of the considered vocabularies

4 Evaluation and Results

The methods proposed by our group do not require to be trained, i.e., they do not require tagged data, only a reference corpus, therefore, it was possible to evaluate them on the whole training set available this year. Table 3 shows their results on this set.

Method	Correlation
STS-Baseline	0.455
STS-Occur	0.500
STS-Contex	0.511

Table 3. Correlation values of the proposed methods and our baseline method with human judgments.

Results in Table 3 show that the use of the co-occurrence information improves the correlation with human judgments. It also shows that the use of context information further improves the results. One surprising finding was the competitive performance of our baseline method; it is considerably better than the previous year's baseline result (0.31).

In order to evaluate the method done in collaboration with LIPN and UMCC_DLSI, we carried out several experiments using the features provided by each group independently and in conjunction with the others. The experiments were performed over the whole training set by means of two-fold cross-validation. The individual and global results are shown in Table 4.

As shown in Table 4, the result corresponding to the combination of all features clearly outperformed the results obtained by using each team´s features independently. Moreover, the best combination of features, containing selected features from the three teams, obtained a correlation value very close to last year's winner result.

Featured by Group	Perdition Model	Correlation
LIPN	Gaussian Process	0.587
LIPN	Lineal Regression	0.701
LIPN	Multilayer-NN	0.756
UMCC_DLSI	Gaussian Process	0.388
UMCC_DLSI	Lineal Regression	0.388
UMCC_DLSI	Multilayer-NN	0.382
INAOE-UPV	Gaussian Process	0.670
INAOE-UPV	Lineal Regression	0.674
INAOE-UPV	Multilayer-NN	0.550
ALL	Gaussian Process	0.770
ALL	Lineal Regression	0.777
ALL	Multilayer-NN	0.633
SELECTED-SET	Multilayer-NN	0.808
LAST YEAR´S WINNER	Simple log-linear regression	0.823

Table 4. Results obtained by the different subsets of features, from the different participating groups.

4.1 Officials Runs

For the official runs (refer to Table 5) we submitted the results corresponding to the STS_{Occur} and $STS_{Context}$ methods. We also submitted a result from the method done in collaboration with LIPN and UMCC_DLSI. Due to time restrictions we were not able to submit the results from our best configuration; we submitted the results for the linear regression model using all the features (second best result from Table 4).Table 5 shows the results in the four evaluation sub-collections; Headlines comes from news headlines, OnWN and FNWN contain pair senses definitions from WordNet and other resources, finally, SMT are translations from automatic machine translations and from the reference human translations.

As shown in Table 5, the performances of the two proposed methods by our group were very close. We hypothesize that this result could be caused by the use of a larger vocabulary for the computation of co-occurrence statistics than for the calculation of the context similarities. We had to use a smaller vocabulary for the later because its higher computational cost.

Finally, Table 5 also shows that the method done in collaboration with the other groups ob-

tained our best results, confirming that using more information about the semantic similarity of words allows improving the estimation of the semantic similarity of texts. The advantage of this approach over the two proposed methods was especially clear on the OnWN and FNWN datasets, which were created upon WordNet information. Somehow this result was predictable since several measures from this "share-method" use WordNet information to compute the semantic similarity of words. However, this pattern was not the same for the other two (WordNet unrelated) datasets. In these other two collections, the average performance of our two proposed methods, without using any expensive and manually constructed resource, improved by 4% the results from the share-method.

Method	Headlines	OnWN	FNWN	SMT	MEAN
STS-Occur	0.639	0.324	0.271	0.349	0.433
STS-Contex	0.639	0.326	0.266	0.345	0.431
Collaboration	0.646	0.629	0.409	0.304	0.508

Table 4. Correlation values from our official runs over the four sub-datasets.

5 Conclusions

The main conclusion of this experiment is that it is possible to extract useful knowledge from raw corpora for evaluating the semantic similarity of texts. Other important conclusion is that the combination of methods (or word semantic similarity measures) helps improving the accuracy of STS. As future work we plan to carry out a detailed analysis of the used measures, with the aim of determining their complementariness and a better way for combining them. We also plan to evaluate the impact of the size and vocabulary richness of the reference corpus on the accuracy of the proposed STS methods.

Acknowledgments

This work was done under partial support of CONACyT project Grants: 134186, and Scholarship 224483. This work is the result of the collaboration in the framework of the WIQEI IRSES project (Grant No. 269180) within the FP 7 Marie Curie. The work of the last author was in the framework the DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems. We also thank the teams from the Universities of Paris 13, Matanzas and Alicante for their willingness to collaborate with us in this evalaution exercise.

References

AngelosHliaoutakis, GiannisVarelas, EpimeneidisVoutsakis, Euripides G. M. Petrakis, EvangelosMilios, 2006, *Information Retrieval by Semantic Similarity*, Intern. Journal on Semantic Web and Information Systems: Special Issue of Multimedia Semantics (IJSWIS), 3(3): 55–73.

Carmen Banea, Samer Hassan, Michael Mohler and RadaMihalcea, 2012, *UNT: A Supervised Synergistic Approach to Semantic Text Similarity*, SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Montreal, Vol. 2: 635-642.

Christiane Fellbaum,2005, *WordNet and wordnets*, Encyclopedia of Language and Linguistics, Second Ed., Oxford, Elsevier: 665-670.

David Pinto, Hector Jiménez H. and Paolo Rosso. *Clustering abstracts of scientific texts using the Transition Point technique*, Proc. 7th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2006, Springer-Verlag, LNCS(3878): 536-546.

EnekoAgirre, Daniel Cer, Mona Diab and Aitor Gonzalez-Agirre, *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity*. SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval2012), Montreal, Vol. 2: 386-393.

James Richard Curran, 2003, *Doctoral Thesis: From Distributional to Semantic Similarity*, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Peter D. Turney, 2005, Measuring semantic similarity by latent relational analysis, IJCAI'05 Proceedings of the 19th international joint conference on Artificial intelligence, Edinburgh, Scotland: 1136-1141

Peter D. Turney, 2006, *Similarity of Semantic Relations*, Computational Linguistics, Vol. 32, No. 3: 379-416.

Philip Resnik, 1999, *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*, Journal of Artificial Intelligence Research, Vol. 11: 95-130.