DERI&UPM: Pushing Corpus Based Relatedness to Similarity: Shared Task System Description

Nitish Aggarwal* **Kartik Asooja°** **Paul Buitelaar***

*Unit for Natural Language Processing
Digital Enterprise Research Institute
National University of Ireland, Galway, Ireland
firstname.lastname@deri.org

°Ontology Engineering Group
Universidad Politecnica de Madrid
Madrid, Spain
asooja@gmail.com

Abstract

In this paper, we describe our system submitted for the semantic textual similarity (STS) task at SemEval 2012. We implemented two approaches to calculate the degree of similarity between two sentences. First approach combines corpus-based semantic relatedness measure over the whole sentence with the knowledge-based semantic similarity scores obtained for the words falling under the same syntactic roles in both the sentences. We fed all these scores as features to machine learning models to obtain a single score giving the degree of similarity of the sentences. Linear Regression and Bagging models were used for this purpose. We used Explicit Semantic Analysis (ESA) as the corpus-based semantic relatedness measure. For the knowledge-based semantic similarity between words, a modified WordNet based Lin measure was used. Second approach uses a bipartite based method over the WordNet based Lin measure, without any modification. This paper shows a significant improvement in calculating the semantic similarity between sentences by the fusion of the knowledge-based similarity measure and the corpus-based relatedness measure against corpus based measure taken alone.

1 Introduction

Similarity between sentences is a central concept of text analysis, however previous studies about semantic similarities have mainly focused either on single word similarity or complete document similarity. Sentence similarity can be defined by the degree of semantic equivalence of two given sentences, where sentences are typically 10-20 words long. The role of sentence semantic similarity measures in text-related research is increasing due to potential number of applications such as document summarization, question answering, information extraction & retrieval and machine translation.

One plausible limitation of existing methods for sentence similarity is their adaptation from long text (e.g. documents) similarity methods, where word co-occurrence plays a significant role. However, sentences are too short, thats why taking syntactic role of each word with its narrow semantic meaning into account, can be highly relevant to reflect the semantic equivalence of two sentences. These narrow semantics can be reflected from any existing large lexicons [(Wu and Palmer, 1994) and (Lin, 1998)]; nevertheless, these lexicons can not provide the semantics of words which are out of lexicon (e.g. guy) or multiword expressions. These semantics can be represented by a large distributed semantic space such as Wikipedia and similarity can be reflected by relatedness of these extracted semantics. However, relatedness covers broader space than similarity, which forced us to tune the Wikipedia based relatedness with lexical structure (e.g. WordNet) based similarities driven by linguistic syntactic structure, in reflecting more sophisticated similarity of two given sentences.

In this work, we present a sentence similarity using ESA and syntactic similarities. The rest of this paper is organized as follows. Section 2 explores the related work. Section 3 describes our approaches

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 643–647,
Montréal, Canada, June 7-8, 2012.

in detail. Section 4 explains our three different submitted runs for STS task. Section 5 shows the results and finally we conclude in section 6.

2 Related Work

In recent years, there have been a variety of efforts in improving semantic similarity measures, however most of these approaches address this problem from the viewpoint of large document similarity based on word co-occurrence using string pattern or corpus statistics. Corpus based approaches such as Latent Semantic Analysis (LSA) [(Landauer et. al, 1998) and (Foltz et. al, 1998)] and ESA (Gabrilovich and Markovitch, 2007) use corpus statistics information about all words and reflect their semantics in distributional high semantic space. However, these approaches perform quite well for long texts as they use word co-occurrence and relying on the principle that words which are used in the same contexts tend to have related meanings. In case of short text similarities, syntactic role of each word with its meaning plays an important role.

There are several linguistic measures [(Achananuparp et. al, 2008) and (Islam and Inkpen, 2008)], which can account for pseudo-syntactic information by analyzing their word order using n-gram. To do this, Islam and Inkpen defined a syntactic measure, which considers the word order between two strings by computing the maximal ordered word overlapping. (Oliva et. al, 2011) present a similarity measure for sentences and short text that takes syntactic information, such as morphology and parsing tree, into account and calculate similarities between words with same syntactic role, by using WordNet.

Our work takes inspiration from existing approaches that exploit a combination of Wikipedia based relatedness with lexical structure based similarities driven by linguistic syntactic structure.

3 Methodology

We implemented two approaches for the STS task [(Agirre et. al, 2012)]. First approach is a fusion of corpus-based semantic relatedness and knowledge-based semantic similarity measures. The core of this combination is the corpus-based measure because the combination includes the corpus-based semantic relatedness score over the whole sentences and the knowledge-based semantic similarity scores for the words falling under the same syntactic roles in both the sentences. Machine learning models are trained by taking all these scores as different features. For the submission, we used Linear regression and Bagging models. Also, the equation obtained after training the linear regression model shows more weightage to the score obtained by the corpus-based relatedness measure as this is the only score (feature), which reflects the semantic relatedness/similarity score over the full sentences, out of all the considered features for the model. We used ESA as the corpus based semantic relatedness measure and modified WordNet-based Lin measure as the knowledge-based similarity. The WordNet-based Lin relatedness measure was modified to reflect better the similarity between the words. For the knowledge-based similarity, currently we considered only the words lying in the three major syntactic role categories i.e. subjects, actions and the objects. We see the first approach as the corpus-based measure ESA tuned with the knowledge-based measure. Thus, it is referred as TunedESA later in the paper.

Our second approach is based on the bipartite method over the WordNet based semantic relatedness measures. WordNet-based Lin measure (without any modification) was used for calculating the relatedness scores for all the possible corresponding pair of words appearing in both the sentences. Then, the similarity/relatedness score for the sentences is calculated by perceiving the problem as the computation of a maximum total matching weight of a bipartite graph having the words as nodes and the relatedness scores as the weight of the edges between the nodes. To solve this, we used Hungarian method. Later, we refer this method as WordNet-Bipartite.

3.1 TunedESA

In this approach, the ESA based relatedness score for the full sentences is combined with the modified WordNet-based Lin similarity scores calculated for the words falling under the corresponding syntactic role category in both the sentences.

	ALL	Rank-ALL	ALLnrm	RankNrm	Mean	RankMean
Baseline	0.3110	87	0.6732	85	0.4356	70
Run1	0.5777	52	0.8158	20	0.5466	52
Run2	0.5833	51	0.8183	17	0.5683	42
Run3	0.4911	67	0.7696	57	0.5377	53

Table 1: Overall Rank and Pearson Correlation of all runs

	MSRpar	MSRvid	SMTeuro	OnWN	SMTnews
Baseline	0.4334	0.2996	0.4542	0.5864	0.3908
ESA*	0.2778	0.8178	0.3914	0.6541	0.4366
Run1	0.3675	**0.8427**	0.3534	0.6030	0.4430
Run2	0.3720	0.8330	0.4238	**0.6513**	**0.4489**
Run3	**0.5320**	0.6874	**0.4514**	0.5827	0.2818

Table 2: Pearson Correlation of all runs with all five STS test datasets

TunedESA could be summarized as these four basic steps:

- Calculate the ESA relatedness score between the sentences.
- Find the words corresponding to the linguistic syntactical categories like subject, action and object of both the sentences.
- Calculate the semantic similarity between the words falling in the corresponding subjects, actions and objects in both the sentences using modified WordNet-based measure Lin.
- Combine these four scores for ESA, Subject, Action and Object to get the final similarity score on the basis of an already learned machine learning model with the training data.

ESA is a promising technique to find the relatedness between documents. The texts which need to be compared are represented as high dimensional vectors containing the TF-IDF weight between the term and the Wikipedia article. The semantic relatedness measure is calculated by taking the cosine measure between these vectors. In this implementation of ESA [1], the score was calculated by considering the full sentence at a time for making the Wikipedia article vector while in the standard ESA, vectors are made for each word of the text followed by the addition of all these vectors to represent the final vector for the text/sentence. It was done just to reduce the time complexity.

To calculate the lexical similarity between the words, we implemented WordNet-based semantic relatedness measure Lin. This score was modified to reflect a better similarity between the words. In the current system, basic linguistic syntactic categories i.e. subjects, actions and objects were used. For instance, below is a sentences pair from the training MSRvid dataset with the gold standard score and the syntactic roles.

Sentence 1: A man is playing a guitar.
Subject: Man, Action: play, Object: guitar

Sentence 2: A man is playing a flute.
Subject: Man, Action: play, Object: flute

Gold Standard Score (0-5): 2.2

As the modification, the scores given by Lin measure were used only for the cases where subsumption relation or hypernymy/hyponymy exists

[1] ESA* considering full sentence at a time to make the vector i.e. different from standard ESA

between the words. This modification was done only for the words falling under the category of subjects and objects.

3.2 WordNet Bipartite

WordNet-based semantic relatedness measure was used for the second approach.

Following steps are performed :

- Each sentence is tokenized to obtain the words.
- Semantic relatedness between every possible pair of words in both the sentences is calculated using WordNet-based measure e.g. Lin.
- Using the scores obtained in the second step, the semantic similarity/relatedness between the sentences is calculated by transforming the problem as that of computing the maximum total matching weight of a bipartite graph, which can be done by using Hungarian method.

4 System Description

We submitted three runs in the semantic textual similarity task. The first two runs are based on the first approach i.e. TunedESA and they differ only in the machine learning algorithm used for obtaining the final similarity score based on all the considered scores/features.

ESA was implemented on the current Wikipedia dump. WordNet based relatedness measure Lin was modified to give a better semantic similarity degree. Stanford Core-NLP library was used for obtaining the words with their syntactic roles. All the required scores/feature i.e. ESA based relatedness for the complete sentences and modified WordNet-based Lin similarity scores were calculated for the corresponding words lying in the same syntactic categories. Bagging and Linear Regression models were built using the training data for the first and second runs respectively. Based on the category of the test dataset, model was trained on the corresponding training dataset.

For the surprise test datasets, we trained our model with the training dataset of the MSRvid data based on the fact that we obtained good results with this category. Then the built models were used for calculating the similarity scores for the test data.

For the third run, WordNet Bipartite method was used to calculate the similarity scores. It didn't require any training.

5 Results and Discussion

All above described runs are evaluated on STS test dataset. Table 1 shows the overall results[2] of our three runs against the baseline system which follows the bag of words approach. Table 2 shows the Pearson correlation on different test datasets for all the three runs. It provides a comparison between corpus based relatedness measure ESA and our system TunedESA (Run 1 & Run 2).

The results show significant improvement against ESA. Although, it can be seen that the baseline results are even better than of the ESA in the cases of MSRpar and SMTeuro. It may be because this implementation of ESA is not the standard one.

6 Conclusion

We presented a method to calculate the degree of sentence similarity based on tuning the corpus based relatedness measure with the knowledge-based similarity measure over the syntactic roles. The results show a definite improvement by the fusion. As future work, we plan to improve the syntactic role handling and considering more syntactical categories. Also, experimentation[3] with standard ESA and other semantic similarity/relatedness measures needs to be performed.

Acknowledgments

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

[2] results can also be found at `http://www.cs.york.ac.uk/semeval-2012/task6/index.php?id=results-update` with the name **nitish_aggarwal**

[3] We plan to provide the further results and information at `http://www.itssimilar.com/`

References

Achananuparp Palakorn and Xiaohua Hu and Xiajiong Shen 2008 The Evaluation of Sentence Similarity Measures, In: DaWaK. pp. 305-316

Agirre Eneko , Cer Daniel, Diab Mona and Gonzalez-Agirre Aitor 2012 SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Foltz P. W., Kintsch W. and Landauer T. K. 1998. *In: journal of the Discourse Processes. pp. 285-307*, The measurement of textual Coherence with Latent Semantic Analysis,

Gabrilovich Evgeniy and Markovitch Shaul 2007 Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, In: Proceedings of The Twentieth International Joint Conference for Artificial Intelligence. pp. 1606–1611,

Islam, Aminul and Inkpen, Diana 2008 Semantic text similarity using corpus-based word similarity and string similarity, In: journal of ACM Trans. Knowl. Discov. Data. pp. 10:1–10:25

Landauer Thomas K. ,Foltz Peter W. and Laham Darrell 1998. An Introduction to Latent Semantic Analysis, *In: Journal of the Discourse Processes. pp. 259-284*,

Lin Dekang 1998 Proceeding of the 15th International Conference on Machine Learning. pp. 296–304 An information-theoretic definition of similarity

Oliva, Jesús and Serrano, José Ignacio and del Castillo, María Dolores and Iglesias, Ángel April, 2011 SyMSS: A syntax-based measure for short-text semantic similarity In: journal of Data Knowledge Engineering. pp. 390–405

Wu, Zhibiao and Palmer, Martha 1994 Verbs semantics and lexical selection, In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics,