Advanced Methods for Sentence Semantic Similarity

Tomáš Ptáček[1]

1 Introduction

Determining the similarity between sentences is one of the crucial tasks in natural language processing (NLP). It has wide impact in many text-related research fields.

Computing sentence similarity is not a trivial task, due to the variability of natural language expressions. Techniques for detecting similarity between long texts (documents) focus on analyzing shared words but in short texts the word co-occurrence may be rare or even null. That is why sentence semantic similarities incorporate the syntactic and semantic information that can be extracted at the sentence level.

The result of this work is an evaluation of five state-of-the-art sentence similarity measures and six proposed sentence similarity measures. These sentence similarity measures are evaluated on two publicly-available sentence pair data sets. The data sets are the Microsoft Research paraphrase corpus (MSRP) and the Semantic Textual Similarity (STS) shared task.

2 Sentence Similarity Measures

Basically, sentence similarity measures compute similarity score based on the number of words shared by two sentences. The syntactic composition of the sentence or the semantic information contained in the sentence can also be used to determine the semantic similarity. For example *Phrasal Overlap Measure* is defined as the relation between length of phrases and their document frequencies.

We adjusted or combined various sentence similarity measures to achieve better results. For example we removed inverse document frequency ($idf(w)$) from *Mihalcea Semantic Similarity* (mihalcea (2006)), because in a document composed of two sentences it is mostly constant. The similarity score is computed according to equation 1.

$$sim_{EMi}(s_1, s_2) = \tanh\left(\frac{\sum_{w \in \{s_1\}} maxSim(w, s_2)}{|s_1|} + \frac{\sum_{w \in \{s_2\}} maxSim(w, s_1)}{|s_2|}\right), \quad (1)$$

where $maxSim(w, s_i)$ is the maximum semantic similarity score of w and given sentence. The semantic similarity scores are computed only between tokens in the same part of speech class because the used knowledge base (WordNet) is unable to compute semantic similarity between tokens in different part of speech classes.

3 Evaluation

We use different evaluation criteria on each data set. In the MSRP data set only two values are assigned to each sentence pair (semantically equivalent or semantically not equivalent), thus we chose accuracy as the main evaluation criteria. Rejection rate and acceptance rate are additional metrics used to evaluate the MSRP data set. For the STS data set the evaluation

[1] student of master program Computer Science and Engineering, study field Software Engineering, e-mail: tigi@students.zcu.cz

criteria was given. The STS shared task has Pearson's correlation as it's official score, thus we chose correlation as well.

3.1 Results

The STS task results are available at http://www.cs.york.ac.uk/semeval-2012/task6/. They evaluated 89 participating systems including a baseline. If we had participated in this task, we would have placed on the 70th position out of 90 participants. On table 1 the results of the STS task are shown in comparison to our best result (*Combined Mihalcea Phrasal Overlap Measure* with *Enhanced WordNet Token Similarity*).

Participant	Correlation					
	ALL	MSRpar	MSRvid	SMTeur	OnWN	SMTnews
First place	0.8239	0.6830	0.8739	0.5280	0.6641	0.4937
Our best result	0.4594	0.2330	0.4666	0.3483	0.4507	0.4844
Baseline	0.3110	0.4334	0.2996	0.4542	0.5864	0.3908

Table 1: STS task's results

The rank was computed according to correlation for STS.ALL data set. On STS.SMTnews our results are basically equivalent. In comparison to the baseline our results are better on STS.ALL, STS.MSRvid and STS.SMTnews data sets.

We computed the percentage difference of our best result in comparison to the first place and baseline. Our result for STS.SMTnews is only 1.88% inferior to the result of the winner. The STS.MSRpar, STS.SMTeur data sets are the weakness of our similarity measure otherwise we achieve at least the correlation of 0.4507. Our similarity measure is better then the baseline by 23.95% on the STS.SMTnews data set and by 55.75% on the STS.MSRvid data set. The overall result for STS.ALL data set is better than the baseline by 47.72%. That is quite good considering that our system computes results within two minutes (Intel Core i5-430M, 4GB RAM, JDK 1.6.0_20, windows 7) and doesn't involve e.g. deep syntactic parsing.

4 Conclusion

The evaluation demonstrates that the proposed semantic similarity measures are equivalent or even better than the state-of-the-art measures. Our proposed sentence similarity method (*Combined Mihalcea Phrasal Overlap Measure* with *Enhanced WordNet Token Similarity*) outperformes the baseline of the STS shared task by 47.72%. On STS.SMTnews data set our result is only 1.88% inferior to the result of the task's winner.

References

R. Mihalcea, C. Corley, and C. Strapparava, 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *Proceedings, 21th National Conference on Artificial intelligence*.

Semantic Textual Similarity (STS) shared task, 2012. URL: http://www.cs.york.ac.uk/semeval-2012/task6/, [online], cit. 2012-05-15.

W. Dolan, C. Quirk and C. Brockett, 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. *Proceedings, 20th International Conference on Computational Linguistics*.