CS 388 Natural Language Processing

Project Proposal

Due : April $7^{th}, 2014$

Madhura Parikh
mparikh@cs.utexas.edu

Avani Gupta
avanigupta@utexas.edu

1 Problem statetment

We intend to work on the task of sentence similarity - i.e. given a pair of English sentences, we would like to score on a scale of $(0-5)$ how similar the two sentences are, with 5 meaning semantically equivalent and 0 meaning no similarity. Specifically we shall be working on the Semantic Textual Similarity (STS) task - which was one of the 6 tasks in Sem Eval 2012 and also the core task in *SEM 2013. Unlike binary decision, wherein pairwise sentences are considered to be either 'similar' or 'not similar', this type of graded equivalence notion is likely to have much more utility, when applied to other NLP tasks like Machine Translation, Information Retrieval and Question Answering. One major motivation of choosing this for our class project is that it will mean interacting with several NLP areas we learnt in class - lemmatization, POS-tagging, parsing, WSD, etc to extract different measures of lexical, syntactic and semantic similarity.

2 Related Work

One major advantage of working on the STS task is that there have been 80 odd submissions for the official Sem Eval task in both 2012-13, meaning that we can draw from these experiences to find out which approaches and features worked the best and how we can leverage off them to get improved results. A majority of the teams have used a supervised approach. For lexical similarity most teams have used the edit distance and other metrics of string similarity such the Needleman-Wunsch distance, the Smith-Waterman distance, etc. Most of the teams have also worked with the lemmatized sentences rather than considering the actual words. Most teams have also used POS-tags as a feature and some have also used chunking and chunk-overlap amongst the two sentences as well as n-gram and skip-gram overlap as features. Another often used feature is based on the dependency parse of the sentence, which enables comparing the S-V-O structure of sentences. For the n-gram modeling, several approaches have used not just the training data but also external corpora such as the Google n-gram 1T corpus, Wikipedia, etc. Another simple but highly useful metric was to find the *number overlap* , i.e if the numbers/dates/percentages that appear in one sentence match with those in the other.

A much harder aspect of the problem is judging the semantic similarity of the two sentences. Most approaches first picked out the content words of each sentence and then used WordNet to find the similarity between the words pairwise, aggregating them in different ways to come up with a final similarity score. Interestingly while some teams used the least common ancestor or the shortest path metric, there really was no well defined metric that could be deemed helpful for the task - which is something we could look at. The source used for similarity also played a very important role and one team mentions that using the Roget's thesaurus gave better results than WordNet. The BLEU measure was also a frequently used metric. LSA and Random Indexing were also commonly used as were several different distributional vector models. WSD was performed using the Lesk algorithm in conjunction with WordNet for most cases. A few teams also used NER for judging similarity, while SRL was used just by a couple of teams.

A few other interesting approaches used LDA based models to assign topics to a sentence, and used IR engines like Lucene as well as TF-IDF for comparing similarity.

Most of the teams used SVR for predicting the final outcome and a couple of teams used an ensemble-based approach such as boosting.

3 Our Contribution

While we will be able to come up with a more concrete approach only once we start working with the data, there are a few contributions we hope to add. First of all we would like to use the features that were judged the best and most useful by a majority of the teams, using that as our baseline model. As improvisations on the baseline model, we suggest the following:

- Pinpoint which metric should be preferred when comparing similarity of word-pairs by traversing WordNet.
- While teams have used dependency parse of a sentence, we would also like to use the syntactic Parse tree, and compute similarity using tree kernels
- Another interesting approach that is frequently used in query engines is *shingling* to determine how similar a query is to the target documents. We will incorporate this in our feature set.
- Finally we would like to experiment which features work best for different classifiers and train these classifiers using different feature sets, using the ensemble as our final system.

4 Data

We shall be using the testing and training data available from the Sem Eval 2012 and 2013 tasks for our project.

5 Evaluation

Just like the official Sem Eval STS task, we shall use the Pearson correlation coefficient as our primary evaluation metric.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-agirre, and Weiwei Guo. sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*, 2013.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics, 2012.