Corpus-based and Knowledge-based Measures of Text Semantic Similarity

Rada Mihalcea and Courtney Corley, Carlo Strapparava

Name: FNU, Arpana Hosabettu, **NetId** : fa97

Introduction:

The paper presents a method for measuring the semantic similarity of short texts, using corpus-based and knowledge-based measures of similarity as previous work in this field has been focused on the large text or individual words. The authors, carefully select various models from corpus and knowledge based measures, explain their approach followed by analysis of individual model and the overall results of their approach. The paper also walks over examples to give the reader the sense of its applications and correctness.

Strong Points:

The author have conducted a very detailed study of the field and provided relevant details as required for the purpose. They have carefully chosen various models with corresponding features . The model has been described very clearly in the paper. The two text comparison for word-word similarities by applying of specificity weight and averaging is a clever approach. Also where the knowledge based does not apply the model applies lexical match measure as a fallback which produces better results. A walk through example provides a very good insight to the reader about the model and the relevancy of the model.

The paper included the individual results providing more scope for analysis and further improvement in the model and on the selection of the relevant metrics for other related work. The paper also provides personal observation on each metric where it produces good result and where it fails and reasoning for the same. The authors have produced clear results and examples in providing a very rational evaluation their model. The author have also discussed the extent of reusability or applicability of these methods to the current application.

Drawbacks:

The drawbacks of the paper is in not defining how the model performed when combining all the divergent models. The author has followed a simple averaging to determine the combined results. It could have been incorporated in their model in a sophisticated way. The author suggested that these different model show the same behavior for small texts. Also the individual results raises a question as to whether those many knowledge based metrics were in fact necessary for the models efficiency. The author makes a claim as to the error rate reduction was 13%, this was not justified in the evaluation of the their results.

Conclusion:

In conclusion, the authors have been quite successful in describing their proposed model. The details of the approach and computational model used, the study and explanation of various metrics as applicable to their model is insightful. This model presented can be further extended by incorporating more sophisticated approaches and factoring other details of semantics like including a functional words which produces semantics through syntax involved in real world. Also, more investigation can be performed on the cases where this approach fails and how the model could be applied to various other corpus of paraphrases could be possible extensions.