Dissimilarity Kernels for Paraphrase Identification

Mihai Lintean (Advisor: Dr. Vasile Rus)

Paraphrasing is a text-to-text relation between two non-identical text fragments that express the same idea in different ways. Paraphrase identification is the task of deciding, given two text fragments, whether they have the same meaning. It should be noted that paraphrase identification is different from paraphrase extraction. Paraphrase extraction (Brockett and Dolan 2005) is the task of extracting fragments of texts that are in a paraphrase relation from various sources. Paraphrase identification and extraction are important tasks in a number of applications including Question Answering (Ibrahim, Katz, and Lin 2003), Natural Language Generation (Iordanskaja, Kittredge, and Polgere 1991), and Intelligent Tutoring Systems (Graesser et al. 2005).

This year we continue our search for pertinent and novel solutions to solve the paraphrase identification task. One such promising solution, which we present here, is based on lexical dissimilarity kernels. Lexical kernels in conjunction with Support Vector Machines are preferred over other learning methods, e.g. decision trees, due to their ability to handle a high number of features. Dissimilarity-based kernels emphasize dissimilarities among text fragments and therefore are appropriate for text similarity tasks characterized by high lexical overlap. We conducted experiments with our kernels on the Microsoft Research Paraphrase Corpus, a standardized data set used for assessing approaches to paraphrase identification. Our reported accuracy results are competitive and robust when compared to state-of-the-art approaches. The results were obtained using 10-fold cross-validation over the entire corpus. We also report competitive results on the test portion of the Microsoft Research Paraphrase Corpus, which is the common way, although less reliable, to report results on this corpus.

## References

Brockett, C., and Dolan, W. B. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the 3rd International Workshop on Paraphrasing*.

Ibrahim, A.; Katz B.; and Lin, J. 2003. Extracting Structural Paraphrases from Aligned Monolingual Corpora. In *Proceeding of the Second International Workshop on Paraphrasing* (ACL 2003).

Iordanskaja, L.; Kittredge, R.; and Polgere, A. 1991. Natural Language Generation in Artificial Intelligence and Computational Linguistics. *Lexical selection and paraphrase in a meaning-text generation model*, Kluwer Academic.

Graesser, A.C.; Olney, A.; Haynes, B.; and Chipman, P. 2005. *Cognitive Systems: Human Cognitive Models in Systems Design.* Erlbaum, Mahwah, NJ. chapter AutoTutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue.