Levent Özgür

Thesis Supervisor: Assoc. Prof. Tunga Güngör

Improving Text Classification Performance with the Analysis of Lexical Dependencies and Class-Based Feature Selection

In this thesis, we present a comprehensive analysis of the feature extraction and feature selection techniques for the text classification problem in order to achieve more successful results using much smaller feature vector sizes. For feature extraction, 36 different lexical dependencies are included and analyzed independently in the feature vector as an extension to the standard bag-of-words approach. Feature selection analysis is twofold. In the first stage, pruning implementation is analyzed and optimal pruning levels are extracted with respect to dataset properties and feature variations (words, dependencies, combination of the leading dependencies). In the second stage, we compare the performance of corpus-based and class-based approaches for feature selection coverage and then, extend pruning implementation by the optimized class-based feature selection. For the final and most advanced test, we serialize the optimal use of the leading dependencies for each experimented dataset with the two stage (corpus and class-based) feature selection approach. For performance evaluation, we use the state-of-the-art measures for text classification problems: two different success score metrics and three different significance tests. With respect to these measures, the results reveal that for each extension in the methods, a corresponding significant improvement is obtained. The most advanced method combining the leading dependencies with optimal pruning levels and optimal number of class-based features mostly outperform the other methods in terms of success rates with reasonable feature sizes. To the best of our knowledge, this is the first study that makes such a detailed analysis on extracting individual dependencies and employing feature selection with two stage selection approach in text classification and more generally in text domain.

PUBLICATIONS

Journals

1. **Levent Ozgur** and Tunga Gungor, Optimization of Dependency and Pruning Usage in Text Classification, Pattern Analysis and Applications, 2011. (accepted, in press)
2. **Levent Ozgur** and Tunga Gungor, Text Classification with the Support of Pruned Dependency Patterns, Pattern Recognition Letters, Vol.31, 2010, p.1598-1607.

Conferences

1. **Levent Ozgur** and Tunga Gungor, Analysis of Stemming Alternatives and Dependency Pattern Support in Text Classification, 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), Ed. A.Gelbukh, March 2009, Mexico City – Research in Computing Science, Vol.41, 2009, p.195-206, IPN.
2. Arzucan Ozgur, **Levent Ozgur** and Tunga Gungor, TextCategorization with Class-Based and Corpus-Based Keyword Selection, 20thInternational Symposium on Computer and Information Sciences (ISCIS 2005), Eds.P.Yolum, T.Güngör, F.Gürgen

and C.Özturan, October 2005, Istanbul - LNCS (Lecture Notes in ComputerScience, Vol.3733), 2005, p.607-616, Springer-Verlag, Berlin Heidelberg.

Defense Jury Members

Assoc. Prof. Tunga Gungor	Bogazici University
Prof. Levent Akin	Bogazici University
Prof. Fikret S. Gurgen	Bogazici University
Assoc. Prof. M. Borahan Tumer	Marmara University
Dr. Suzan Uskudarli	Bogazici University

Defense Date: 03.06.2010