Abstract

Frame Assignment with Active Learning

Masood Ghayoomi

Master of Science Thesis
Department of Computational Linguistics and Phonetics
Saarland University
2009

Recently natural language understanding is given a special attention, since in natural language processing techniques syntactic analysis such as part-of-speech tagging and parsing had a great progress and semantic analysis did not have such a rapid progress. In information extraction and question-answering systems semantic understanding techniques are required. Frame semantics structure analysis is one of the understanding techniques. In this type of analysis, the semantic roles of elements participated in the action would be identified. To determine the roles automatically, two steps are required: one is frame assignment, and the other one is role assignment.

What we aim to do is assigning frames with a supervised machine learning method called 'active learning'. Supervised learning method requires a huge amount of labeled data. The aim of active learning promises to maximize the performance by minimizing the human's effort to label the data.

To our end, we have selected pool-based active learning with uncertainty sampling method; and also we have chosen 14 frequent targets from FrameNet data set for our task. Random sampling which represents the distribution of frames in the corpus would be our baseline to find how effective active learning is. Since for each target there was at least one dominant frame, we faced the imbalanced problem which might have a negative impact on the classifier; so over-sampling is used to resolve this problem.

According to the results, active learning worked out for most of the targets; for some of them it was not that much effective; and for some it had a negative impact which shows active learning could not always be a help. We have discussed this issue in details.

As we know active learning is an iterative process, it should be stopped when the classifier has reached to the maximum performance. Reaching this point is so difficult; so, we have proposed a new stopping criterion which stops active learning in a near-optimum point. This stopping criterion is based on the confidence score of the classifier on the extra unlabeled data such that it uses the variance of the classifier's confidence score for a certain number of samples which are selected in each iteration.