ABSTRACT

Semantic Argument Classification is the process of analyzing the sentence to investigate the pattern of WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW, from a structured text data. Research on the classification of semantic arguments requires data that has been labeled semantically in large numbers, which is called corpus. In the preliminary research, two types of corpus have been built, namely FrameNet and Propbank, both are from news domain or news genre. Because building a corpus is costly and time-consuming, recently many studies have used FrameNet and Propbank corpus as training data to conduct semantic argument classification research on new domains without the need to build a new corpus for those new domains.

This thesis will perform a research related to semantic argument classification on a new domain that is Quran English Translation by utilizing Propbank corpus as training data. The Quran English translation is a translation of the original Arabic Quran. Hence the composition of grammar and the sentence structure, English-Quran is still influenced by the original languages, namely Arabic. In its original language, Arabic, the holy Quran has a significant difference from the newswire domain, being closer to poetic language, more creative linguistic expression, and has many variations of vocabulary and sentence structure.

Previous studies have proven that there is a significant decrease in performance when classifying semantic arguments on different domain between the training and the testing data. The main problem is when there is a new argument that found in the testing data but not found in the training data. To recognize this new argument in training data, one solution is by extending the argument features in the training data to accommodate the new features of the new argument. This thesis proposes four new features to improve the baseline system performance.

By using SVM Linear, the experiment has proven that the performance of semantic argument classification on Quran data using Propbank Corpus as training data can be improved by augmenting the proposed features to the baseline system with some combination option. When tested on auto labeled data, the augmentation of PTO+SP features to the baseline system can improve the accuracy by 1.25% and improve F-1 score by 1.30%. When tested on hand-labeled data, the augmentation of combination PO+PTO features to the baseline system can improve the accuracy by 0.47% and improve F-1 score by 0.40%.

Keywords: semantic argument classification, semantic role labeling, shallow semantic parsing, classification algorithm, Support Vector Machine classifier.