Arab Academy for Science,
Technology & Maritime Transport

ABSTRACT

ARABIC WORD SENSE DISAMBIGUATION

By Mohamed Mosleh El-gamml

Chairperson of the Supervisory Committee: Professor M. Waleed Fakhr
Professor Mohsen Rashwan

Word sense disambiguation is a core problem in many tasks related to language processing and was recognized at the beginning of the scientific interest in machine translation and artificial intelligence. In this thesis, we introduce the possibilities of using the Support Vector Machine (SVM) classifier to solve the Word Sense Disambiguation problem in a supervised manner after using the Levenshtein Distance algorithm to measure the matching distance between words through the usage of the lexical samples of five Arabic words. The performance of the proposed technique is compared to supervised and unsupervised machine learning algorithms, namely the Naïve Bayes Classifier (NBC) and Latent Semantic Analysis (LSA) with K-means clustering, representing the baseline and state-of-the-art algorithms for WSD. In contrast, the SVM decision function is fully determined by a small subset of the training data, called support vectors. Therefore, it is desirable to remove from the training set the data that is irrelevant to the final decision function. In this thesis we propose a new method that selects a subset of data for SVM training. Using real-world dataset, we compare the effectiveness of the proposed data selection strategies in terms of their ability to reduce the training set size while maintaining the generalization performance of the SVM classifier results.