Unsupervised Induction of Modern Standard Arabic Verb Classes and Alternations

Neal Snider

I exploit the resources in the Arabic Treebank (ATB) for the novel task of automatically creating lexical semantic verb classes for Modern Standard Arabic (MSA). Following the hypothesis of Levin (1993), the verbs are classified into groups that share semantic elements of meaning because exhibit similar syntactic behavior.

Verbs (in lemma form) and syntactic frames are automatically extracted from the ATB. In order to acquire an argument structure for the verbs, I only considered structure that is internal to the maximal Verb Phrase (VP) projection of the verb. However, within the VP, all sisters of the verb are excluded except for those in a close semantic relationship to the verb. This is facilitated by the fact that the ATB annotators specifically tag internal noun phrases (NP)s as subjects and objects. Moreover, I consider prepositional phrases (PPs) tagged as *closely related* since they are manually indicated by the ATB annotators as essential for the meaning of the verb. Finally, I assume clauses that are sisters of the verb are arguments of the verb.

The verbs are clustered by creating a vector for each verb whose elements represent the co-occurence of that verb lemma with each of the possible syntactic frames in the corpus. Features are also added that represent the animacy of the subject and morphological class of the verb. The verb vectors are clustered using both hard and soft clustering algorithms.

The results of the clustering experiments are compared with a gold standard set of classes, which is approximated by using the noisy English translations provided in the ATB to create Levin-like classes for MSA. The quality of the clusters is found to be sensitive to the inclusion of information about lexical heads of the constituents in the syntactic frames, as well as parameters of the clustering algorithm such as the algorithm type (hard or soft), number of clusters, and number of verbs clustered. The best set of parameters yields an F1 score of 0.501, compared to a random baseline with an F1 score of 0.37 and a LSA baseline with F1 score of 0.415.

This technique is also useful for traditional lexical semantics to help discover new lexical syntactic alternations for MSA. When a set of column entries of the verb matrix, which represent the syntactic frames, consistently co-occur with certain verbs, the columns may be hypothesized to represent a possible syntactic alternation and the verbs the class defined by that alternation. These hypothetical alternations were tested against native speaker intuitions, and three alternations have been discovered: the *bi-fiy*, the *ElaY*-CP, and the *bi*-drop alternations.