Abstract

This paper reports on compiling a large Arabic corpus of the Holy Qur’an script, annotated with anaphoric relation and other anaphoric information, providing multi-dimensional feature vector rich with most of basic anaphoric information needed in statistical anaphora resolution systems. About 24,653 personal pronouns are tagged with their antecedents and other anaphoric information like distance between the anaphor and its antecedent in terms of verses, words, and segments, gender, number, person, and other information which can be used to implement the feature vector of a statistical anaphora resolution system. In addition, it describes the compilation of a bank of sentence patterns consisting of 481 antecedent patterns; each pattern represents particular part-of-speech tag corresponding to its antecedent phrase. The aim is to provide a valuable resource that enables future research in Arabic anaphora resolution, and help in future work in analyzing Quran script. Also, it will be a valuable resource that can be used for training and testing anaphora resolution systems, and evaluating.


Index Terms

Computer Science                      Artificial Intelligence

Keywords

Anaphora resolution, Arabic language, Corpus, Quran, Pronominal anaphora.