World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:10, 2015

Detecting Paraphrases in Arabic Text

Authors : Amal Alshahrani, Allan Ramsay

Abstract : Paraphrasing is one of the important tasks in natural language processing; i.e. alternative ways to express the same concept by using different words or phrases. Paraphrases can be used in many natural language applications, such as Information Retrieval, Machine Translation, Question Answering, Text Summarization, or Information Extraction. To obtain pairs of sentences that are paraphrases we create a system that automatically extracts paraphrases from a corpus, which is built from different sources of news article since these are likely to contain paraphrases when they report the same event on the same day. There are existing simple standard approaches (e.g. TF-IDF vector space, cosine similarity) and alignment technique (e.g. Dynamic Time Warping (DTW)) for extracting paraphrase which have been applied to the English. However, the performance of these approaches could be affected when they are applied to another language, for instance Arabic language, due to the presence of phenomena which are not present in English, such as Free Word Order, Zero copula, and Pro-dropping. These phenomena will affect the performance of these algorithms. Thus, if we can analysis how the existing algorithms for English fail for Arabic then we can find a solution for Arabic. The results are promising.

Keywords : natural language processing, TF-IDF, cosine similarity, dynamic time warping (DTW)
Conference Title : ICSIR 2015 : International Conference on Semantic Information Retrieval
Conference Location : London, United Kingdom
Conference Dates : October 23-24, 2015