An approach for Measuring Arabic Text Semantic Similarity

Proposal submitted to Faculty of Computers and Information - Fayoum University

In fulfillment of the requirements for the M. Sc. degree in information system

Submitted by:

Rehab Ahmed Mohamed Farouk Ahmed EL-Sayed

Supervised by:

Prof. Mohammed Helmy Khafagy
Computer Science Department,
Faculty of Computers and Information,
Fayoum University

Dr. Rasha Mohammed Badry
Information Systems Department,
Faculty of Computers and Information,
Fayoum University

ABSTRACT

Documents that are dealing with the same topic include normally many identical words. There is a need to measure the similarity between these documents. However, the problem is not a trivial task when dealing with texts that carry the same or close meaning but with different vocabularies. Semantic similarity between different texts is a challenging task. It measures the relation between texts, sentences and words to depict their degree of similarity or resemblance. Most of the existing word similarity measures are developed and used for English texts while very rare measures have been developed specifically for Arabic. The rare approaches used with Arabic languages are mainly adapted versions from those used for English texts, so we need to focus on Arabic language and work researches about semantic similarity with Arabic language. There is a need for Arabic semantic similarity in many areas such as agriculture. Agricultural is one of the important and crucial sectors since it might contribute in the national income. On the other hand, there is a lack of support for the user of the agricultural data. The research aims to develop an approach for measuring semantic similarity in Arabic sentences for agricultural data that help farmers and decision makers to provide solution to their unstructured agriculture problems. The proposed system will provide the support, information, advice for users, decision makers of the agricultural data. Many approaches used for measuring semantic similarity ,e.g. Cosine similarity ,Latent Semantic Analysis (LSA) , Vector Space Model (VSM) , etc…