HELWAN
UNIVERSITY

Faculty of Computers and Information

Information Systems Department

Bilingual Text Summarization

Dissertation for Partial Fulfillment Requirements for the
Doctor of Philosophy (Ph.D)
In Information Systems

Submitted to:
Information Systems Department
Faculty of Computers and Information
Helwan University

Prepared By:

Rasha Mohammed Badry Sayed

B.Sc.,Computers and Information,Information systems, May 2003
M.Sc.,Computers and Information,Information systems, Jan 2008

Supervised By:

Prof. Ahmed Sharaf Eldin Ahmed
Professor,
Information Systems Department
Faculty of Computers and Information
Helwan University

Dr. Doaa Saad Elzanfaly
Associate Professor,
Information Systems Department
Faculty of Computers and Information
Helwan University

2015

Abstract

BILINGUAL TEXT SUMMARIZATION

With the revolution of the information age and the evolution of the World Wide Web, computer users face a lot of lengthy text. Users need to read all the available documents to determine the most relevant ones. In addition, they may prefer to have a shorter version of the text. This is a big problem. It is very difficult and a lengthy process for human to manually summarize large documents of text. Automatic text summarization solves this problem.

Nowadays, Text summarization (TS) is among the most attractive research areas. TS is used to provide a shorter version of the original text while keeping the overall meaning. There are various methods that aim to find out well-formed summaries. One of the most commonly used methods is the Latent Semantic Analysis (LSA). LSA is the one that is adopted in this research.

Hundreds of millions of Arabic native speakers and other groups are interested in the Arabic language. Unfortunately there are very few researches covering this area. This is the reason why in this study documents written in Arabic/English languages are addressed. In this research, a generic extractive Arabic/English text summarization system based on LSA (semantic analysis)

and classical method (sentence features) is proposed, designed, implemented, and evaluated.

The proposed system is experimentally tested on a collection of sixty three documents. Three types of documents were considered: Arabic only, English only, bilingual Arabic/English. We use standard set of documents used for testing purposes and available on the web. For evaluating our system, we used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores. ROUGE is a method to measure the quality of the summary using a manually done summary as an ideal and reference summary. Moreover, we compared our summarization system with seven commercially summarization systems.

There is a significant improvement in the summarization system done by the proposed system compared to the seven commercially available text summarization systems for documents written in English only. A similar improvement is also noticed when comparing our system to the four commercially available text summarization systems which support Arabic in the case of documents written in Arabic only and in the case of bilingual Arabic/English documents.