Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 26 (2012) 260–273

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A text summarizer for Arabic*

Aqil M. Azmi [a,*], Suha Al-Thanyyan [b]

[a] Department of Computer Science, King Saud University, Riyadh, Saudi Arabia
[b] College of Computer and Information Sciences, Imam University, Riyadh, Saudi Arabia

Received 15 November 2010; received in revised form 26 November 2011; accepted 6 January 2012
Available online 17 January 2012

Abstract

Automatic text summarization is an essential tool in this era of information overloading. In this paper we present an automatic extractive Arabic text summarization system where the user can cap the size of the final summary. It is a direct system where no machine learning is involved. We use a two pass algorithm where in pass one, we produce a primary summary using Rhetorical Structure Theory (RST); this is followed by the second pass where we assign a score to each of the sentences in the primary summary. These scores will help us in generating the final summary. For the final output, sentences are selected with an objective of maximizing the overall score of the summary whose size should not exceed the user selected limit. We used ROUGE to evaluate our system generated summaries of various lengths against those done by a (human) news editorial professional. Experiments on sample texts show our system to outperform some of the existing Arabic summarization systems including those that require machine learning.

Keywords: Arabic NLP; Rhetorical Structure Theory; Automatic text summarization; 0/1-Knapsack; ROUGE

1. Introduction

Automatic text summarization is the process of abstracting large texts into a few paragraphs while preserving its information content. Hovy (2005) defines a summary as a text that is produced from one or more texts which contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). Text here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc. It is an important problem because of the rapid growth in online information. Given the current state of the art in text summarization, we believe that humans are better in this regard than machines.

Mani et al. (2002) classified text summarization algorithms into six categories. In general there are two main categories of summaries, extractive or abstractive. Extraction techniques simply copy the information (e.g. sentences) deemed most important by the system to the summary, whereas abstraction requires paraphrasing the most important sections of the source document. To the best of the authors' knowledge there is no such system in Arabic.

Unlike English which has seen a large number of systems developed to cater to it, other languages are less fortunate. Only recently did Arabic enter the scene. So far, few summarization systems have been developed for the Arabic

* This paper has been recommended for acceptance by 'Edward J Briscoe'.

* Corresponding author. Tel.: +966 1 467 6574.
E-mail addresses: aqil@ksu.edu.sa (A.M. Azmi), somay_so@yahoo.com (S. Al-Thanyyan).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.01.002

language. Most of the systems reviewed in this section generate open ended summaries, i.e. ones which have no size limit.

The Optimized Dual Classification System (Sobh et al., 2006) is an Arabic extractive text summarization system. The system requires training and uses manually labeled corpora. It integrates Bayesian and Genetic Programming (GP) classification methods. By integrating both classifiers they found that using the *union* for integration increased the recall and the summary size, while using the *intersection* for integration increased the precision and decreased the size of the summary. Roughly, precision measures how much information that the system returned is correct, while recall measures the coverage of the system. These terms are defined later (Eq. (8)).

AlSanie (2005) developed an Arabic text summarization system based on Rhetorical Structure Theory (RST). His algorithm rhetorically analyzed the Arabic texts and generated all possible RS-trees for the text, and then extracted the summary by going to level two on the generated trees. The system was evaluated by comparing human generated summaries with the result of the auto-summarization software. The system did well for small and medium sized articles.

Lakhas (Douzidia and Lapalme, 2004) is a summarization system that generates 10 word summaries of news articles. Lakhas first summarizes the original Arabic documents and then applies machine translation (MT) to translate the summary into English. This approach was very successful in generating very short (headline) summaries as compared to systems using the noisy (MT English) text at DUC 2004.

Furthermore, there are systems which summarize multilingual sets of documents. CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) in Conroy et al. (2006) is an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic/query-driven summaries for single documents or clusters of documents. CLASSY was tested on multilingual English and Arabic documents. The system did not perform well when it had to deal with English and Arabic original documents (rather than the MT translations thereof). However, it scored higher for an input of English using MT of Arabic documents (Schlesinger et al., 2008).

Farsi or Persian, a close relative of Arabic, has its share of text summarizers. Mazdak (2004) developed FarsiSum, an HTTP client/server application based on Swesum (Dalianis, 2000), a summarizer for the Swedish language, and is supposed to work on other languages in the so-called *generic mode*. The single Persian document summarizer by Karimi and Shamsfard (2006) uses five features in addition to lexical chains and graphs in their scoring module. Honarpisheh et al. (2008) multi-document text summarizer is based on singular value decomposition (SVD) and hierarchical clustering. The system takes a collection of related documents transforming them into a matrix where SVD is applied. Then hierarchical clustering is used to determine the sentences forming the summary. Parsumist (Shamsfard et al., 2009) is similar to Karimi and Shamsfard (2006) in that it uses lexical chains and graphs. The system uses statistical and heuristic methods to identify important sentences while conceptual relations are used to rank the sentences. The authors claim their algorithm outperforms FarsiSum (Mazdak, 2004) and Karimi and Shamsfard (2006). None of these authors who worked on the Farsi text summarizer addressed the applicability of their system to Arabic. Our own experience of applying FarsiSum to summarize Arabic text was unsatisfactory. While Farsi has many loan words from Arabic, the structure of the sentence is quite different. This is attributed to their different origins: Farsi is an Indo-European language whereas Arabic is a Semitic language.

An automatically generated summary that lacks cohesion will often be considered a poor final summary. The issue of a cohesive summary is a big research area in this field. One way to address this problem is to use lexical chains; but then these chains are insensitive to the non-lexical structure of texts, such as rhetorical or argumentative. Another approach would be to use RST (Mann and Thompson, 1988), a widely used discourse theory in Computational Linguistics. A comprehensive evaluation in Uzêda et al. (2008) has concluded that automatic text summarization methods which are based on RST are better than extractive summarizers and those with hybrid methods produce worse summaries.

In this paper we introduce an Arabic summarization system. There are two main components to our system, RST and a sentence scoring scheme. Our proposed system calls for generating a primary summary using RST then ranking the sentences in the primary summary using a scoring scheme and at last producing a final summary within the desired user imposed size limit.

The rest of the paper is organized as follows. Section 2 is a brief description of RST as it is an essential part of our system. In Section 3, we discuss our proposed Arabic text summarization system. System evaluation and results are covered in Section 4. The conclusion and suggestions for future research are given in Section 5.