Information Processing and Management 54 (2018) 903–921

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An abstractive Arabic text summarizer with user controlled granularity

Aqil M. Azmi*, Nouf I. Altmami

Department of Computer Science, College of Computer & Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

ARTICLE INFO

Keywords:
Arabic
Abstractive summary
Topic segmentation
Rhetorical Structure Theory
Manual evaluation

ABSTRACT

Automated summaries help tackle the ever growing volume of information floating around. There are two broad categories: extract and abstract. In the former we retain the more important sentences more or less in their original structure, while the latter requires a fusion of multiple sentences and/or paraphrasing. This is a more challenging task than extract summaries. In this paper, we present a novel generic abstract summarizer for a single document in Arabic language. The system starts by segmenting the input text topic wise. Then, each textual segment is extractively summarized. Finally, we apply rule-based sentence reduction technique. The RST-based extractive summarizer is an enhanced version of the system in Azmi and Al-Thanyyan (2012). By controlling the size of the extract summary of each segment we can cap the size of the final abstractive summary. Both summarizers, the enhanced extractive and the abstractive, were evaluated. We tested our enhanced extractive summarizer on the same dataset in the aforementioned paper, using the measures recall, precision and ROUGE. The results show noticeable improvement in the performance, specially the precision in shorter summaries. The abstractive summarizer was tested on a set of 150 documents, generating summaries of sizes 50%, 40%, 30% and 20% (of the original's word count). The results were assessed by two human experts who graded them out of a maximum score of 5. The average score ranged between 4.53 and 1.92 for summaries at different granularities, with shorter summaries receiving the lower score. The experimental results are encouraging and demonstrate the effectiveness of our approach.

1. Introduction

Humans, typically, are quite adept in summarizing text, often picking large chunks of text verbatim. Lin and Hovy (2003) observed that about 70% of the sentences used in manually created summaries were taken from the source with no modification. The sheer volume of online information on the World-Wide Web compels us to seriously consider automated summaries, a tool to squeeze the information to something that we can grasp and comprehend. The task of automatic summarizer is to produce an abridged text conveying the important information of the original text, and whose size is no more than half the original's and often less (Hovy, 2005; Radev, Hovy, & McKeown, 2002). The word 'text' in the previous definition should not be taken literally, and is extended to cover any other media contents as well. Text summarization is an interesting field of natural language processing (NLP). A major challenge in any summarization lies in distinguishing the most informative parts of a document from the less informative ones. Nenkova and McKeown (2011) lists ten different categories of text summarization, with extract and abstract being the more important ones. In the former, we select units (words, sentences, paragraphs, ... etc.) that contain the essence of the document to form a

* Corresponding author .
E-mail addresses: aqil@ksu.edu.sa (A.M. Azmi), nouf_ibraheem@hotmail.com (N.I. Altmami).

https://doi.org/10.1016/j.ipm.2018.06.002
Received 16 June 2017; Received in revised form 8 February 2018; Accepted 1 June 2018

summary. While in the latter, we re-articulate the most salient parts of the source document. The abstractive summary requires deep analysis of the text with the capability to synthesize a compressed version of the original sentence, or compose them in a different form. In contrast to abstract, which is a more challenging problem and requires extensive natural language processing, the field of extract is well researched. In this paper, we present a generic abstractive summarizer of a single Arabic text document.

The cohesiveness of a summary is a major research area. A summary that lacks cohesion is considered poor. The problem of cohesiveness has been addressed through lexical chains, the Latent Semantic Analysis (LSA), and the Rhetorical Structure Theory (RST). The idea of lexical chains is to capture semantic similarity between noun phrases to determine the importance of sentences (Barzilay & Elhadad, 1997). These chains are insensitive to the non-lexical structure of texts, such as rhetorical or argumentative. Moreover, this scheme relies heavily on WordNet (Fellbaum, 1998), a manually compiled thesaurus that lists the different sense of each word. This constraints the scheme effectiveness because the success is bound by the coverage of WordNet, and the sense granularity found therein. The main thrust for LSA is to be able to identify important topics in the document without using lexical resources, e.g. WordNet. LSA uses unsupervised learning to determine text semantics based on either observed co-occurrence of words (Barzilay & Elhadad, 1997) (as it was originally proposed); or co-reference resolution (Baldwin & Morton, 1998; Bergler, Witte, Khalife, Li, & Rudzicz, 2003). The work in Steinberger, Poesio, Kabadjov, and Ježek (2007) has shown that there is a potential for further improvement as the performance of co-reference system gets better. RST (Mann & Thompson, 1988), one of the leading theories in computational linguistics, is a language-independent descriptive theory of text organization that characterizes text structure using relations among the discourse or rhetorical elements that a text contains (Iruskieta, da Cunha, & Taboada, 2014). It provides a framework for describing texts and rhetorical relations among parts of a text. Mann and Thompson (1988) defined coherence as, when you can tell the function of each piece of a text with respect to all the others, using the functionals provided as RST relations. RST requires the overall structure of a text to be represented by a tree, called RS-tree (or schema). Following a comprehensive evaluation of extractive summarizers, Uzêda, Pardo, and Nunes (2010) concluded that RST based summarization are better than those employing hybrid schemes.

According to a study in (www.internetworldstats.com/stats7.htm), Arabic is the fastest-growing language on the web with a growth of 6592.5% in the last fifteen years (2000–2015) for the number of Internet users. This tremendous growth of Arabic contents on the web, resulted in a prolific work on extractive summaries for the Arabic language. In a recent survey of Arabic summarizers, Al-Saleh and Menai (2016) concluded there exists no Arabic summarization system that can generate abstractive summaries.

Munot and Govilkar (2014) stated that abstract summaries might contain words not explicitly present in the original. This spells an ambitious scenario, not feasible with the current state of art of natural language generation (NLG) technology. On the other hand, a more realistic view was reflected by Allahyari et al. (2017). They claim, there is no complete abstractive summarizer system today; and often, existing abstractive summarizers rely on an extractive preprocessing component. The output, we may add, is passed on to NLG component to modify, combine and compress the extracted text to produce the abstract of the text. The framework of our summarizer is very similar. In this work we devise a system to automatically generate abstractive summaries of single Arabic text documents where the user can cap their size. The four phase generic abstractive summarizer are: topic segmentation, headline generation, extractive summarization and sentence reduction techniques. The extractive summarizer we use is an enhanced version of an earlier work (Azmi & Al-Thanyyan, 2012), a hybrid summarizer that combines RST and sentence scoring. By controlling the size of the extractive summaries we can control the overall size of the abstractive summary. We evaluate both the extractive and abstractive summarizers. The first experiment compares the performance of our enhanced extractive summarizer and the one it is based on. For performance evaluation we use the standard measures of recall, precision, F-measure and ROUGE. Overall there is an improvement in the performance, and it is more apparent in shorter summaries. The second experiment was to assess the abstractive summarizer. As there are no known standard measures, or metrics that automatically evaluate the summaries, so we sought the help of two linguist experts to read and judge the abstractive summaries of different sizes (in terms of word count). Each expert was free to devise his/her own assessment methodology, and they both graded out of 5. Testing on 150 documents, the score ranged between 4.53 (for summaries sized 50% of original's) and 1.92 (summaries sized 20% of original). This is the average of both experts. These results are very promising and help demonstrate the effectiveness of our system.

This paper is organized as follows. Section 2 overviews some of the related works. Challenges facing Arabic text summarization are covered in Section 3. In Section 4, we cover the implementation detail of the summarizer. The results of evaluating the summarizer are in Section 5. The last section concludes the paper with some directions for future work.

2. Related work

Though the field of automatic text summarization is over half a century old (Luhn, 1958), there are many problems awaiting effective solutions. One open problem is an automatic scheme to evaluate summaries that does not require reference summaries, and is more consistent than human evaluation (Saggion & Poibeau, 2013). The work on summarizing Arabic texts started appearing during the last decade, so it is fairly small compared to the body of literature on other languages such as English, but it has been growing steadily (El-Haj, Kruschwitz, & Fox, 2011). Generally, most of the efforts were directed towards extractive summarization, covering single and multi-documents. That leaves us with few works in the area of abstractive summarization of Arabic text. We will briefly go over some of the recent works on extractively summarizing Arabic text and then move to the abstractive summarization.

Azmi and Al-Thanyyan (2012) devised a hybrid Arabic text summarizer which combined RST and sentence scoring. The RST was used to generate the primary summary, while sentence scoring was used to guide the generation of a final summary that is within the user's predetermined size of the summary. The authors evaluated their summaries using precision (P), recall (R), F-measure (F_1) and ROUGE (Lin, 2004). The system performed extremely well for medium sized summaries but not so well for shorter summaries. Because