Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 10–18

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

Automatic Arabic Summarization: A survey of methodologies and systems

Lamees Mahmoud Al Qassem[a*], Di Wang[a], Zaid Al Mahmoud[a], Hassan Barada[b], Ahmad Al-Rubaie[a], Nawaf I. Almoosa[a]

[a]*EBTIC, Khalifa University, Abu Dhabi 127788, UAE*
[b]*KU, Khalifa University, Abu Dhabi 127788, UAE*

Abstract

Text summarization has been a field of intensive research over the last 50 years, especially for commonly-used and relatively simple-grammar languages such as English. Moreover, the unprecedented growth in the amount of online information available in many languages to users and businesses, including news articles and social media, has made it difficult and time consuming for users to identify and consume sought after content. Hence, an automatic text summarization for various languages to generate accurate and relevant summaries from the huge amount of information available is essential nowadays. Techniques and methodologies for Arabic text summarization are still immature due to the inherent complexity of the Arabic language in terms of both structure and morphology. This paper describes the main challenges for Arabic text summarization and surveys the various methodologies and systems in the literature. This survey would be a good basis for the design of an Arabic automatic text summarization that combines the various "good" features of the existing systems and dismiss the "not-so-good" features.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Automatic Text Summarization; Arabic Summarization systems; Arabic Natural Language Processing

* Corresponding author. Tel.: +9-712-401-8000; fax: +9-712-447-2442.
E-mail address: 100037108@kustar.ac.ae

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.088

1. Introduction

Due to the exponentially increasing amount of available content online over the past two decades, having automatic text summarization systems to extract and generate key information from the vast amount of available text is highly needed. With the maturity of internet techniques and availability of mobile devices, one can save much time by simply reading or viewing summarized information such as news, articles, etc. rather than reading the whole manuscripts [1].

Techniques for automatic text summarization for widely-used and relatively simple-grammar languages such as English are mature. However, not much work has been done for Arabic summarization [2] due to the complexity of the language in terms of both structure and morphology. On the other hand, Arabic summarization systems are needed nowadays. There are more than 300 million Arabic speakers in the world and Arabic is an official language in the United Nations [2] and 22 other countries [3].

Text summarization is to extract and generate the key information in a concise expression from long document(s) by using various techniques [4]. From the viewpoint of sources, summaries can be generated from either a single or multiple document(s), for a single-language (monolingual) or multiple languages (multilingual) text [4]. From the viewpoint of methodologies, text summarization can be either extractive or abstractive. Extractive summarization selects a subset of important sentences in the original text to form the summary [2]. Abstractive summarization, on the other hand, tries to comprehend the semantics of the original documents beyond the sentence level and re-builds the summarized information in a concise and understandable way. Abstractive summarization is usually a challenging task, especially for complex languages, and not much work has been done in this area. Current active research still focuses on extractive text summarization. Furthermore, from the viewpoint of the user, summaries of text can be either done in a generic way on available document(s) or driven by a user query [2].

This paper describes the main challenges for Arabic text summarization and surveys the various methodologies and systems in the literature. The challenges are discussed in Section 2. The different approaches, methodologies, and systems are described in Section 3. Section 4 is a conclusion.

2. Challenges in Arabic summarization

The main challenge in Arabic text summarization is in the complexity of the Arabic language itself: 1) the meaning of a text is highly dependent on the context; 2) there are more inherent variations within Arabic than any other language; 3) the diacritics are usually absent in the texts of news articles and any online content.

The Arabic language has 28 characters and each character's shape changes based on its position; e.g., the letter "ح" has three different shapes: "حـ" at the beginning of the word; "ـحـ" in the middle; and "ـح" at the end. Arabic has two types of vowels: long vowels represented by letters and short vowels appearing as diacritical marks [2]. Hence, the size of the Arabic alphabet can be extended to be ninety [3]. Furthermore, all words in Arabic are derived from a list of roots with 3 or 4 constants. Morphological analysis is hard because the language is derivational and inflectional [2]. The broken plurals are also considered as a challenge. Broken plurals do not resemble the singular form [5] as in English. All of these challenges make the methods used for text summarization in other languages such as English not appropriate for Arabic summarization. More work on the morphological aspect of the Arabic language and on the pre-processing of the text is needed in order to generate accurate summaries.

Another challenge for Arabic summarization is the evaluation process. Summarization evaluation can be done manually, automatically, or semi-automatically [2]. However, there are no gold standard summaries for Arabic. The lack of Arabic benchmark corpora, lexicons and machine-readable dictionaries make automatic evaluation for Arabic summarization more difficult. In the literature, different Arabic summarization systems apply different corpus and different evaluation criteria for their own evaluation. Without unified benchmark corpus and criteria, the results reported from existing systems can only be a hint for overall performance comparison.