Journal of King Saud University – Computer and Information Sciences (2014) **26**, 450–461

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization

Houda Oufaida [a,*], **Omar Nouali** [b], **Philippe Blache** [c]

[a] *Ecole Nationale Supérieure d'Informatique (ESI), Algiers, Algeria*
[b] *Research Center on Scientific and Technical Information (CERIST), Algiers, Algeria*
[c] *Aix Marseille Université, CNRS, LPL UMR 7309, 13604 Aix en Provence, France*

Available online 28 September 2014

KEYWORDS

Arabic text summarization;
Sentence extraction;
mRMR;
Minimum redundancy;
Maximum relevance

Abstract Automatic text summarization aims to produce summaries for one or more texts using machine techniques. In this paper, we propose a novel statistical summarization system for Arabic texts. Our system uses a clustering algorithm and an adapted discriminant analysis method: mRMR (minimum redundancy and maximum relevance) to score terms. Through mRMR analysis, terms are ranked according to their discriminant and coverage power. Second, we propose a novel sentence extraction algorithm which selects sentences with top ranked terms and maximum diversity. Our system uses minimal language-dependant processing: sentence splitting, tokenization and root extraction. Experimental results on EASC and TAC 2011 MultiLingual datasets showed that our proposed approach is competitive to the state of the art systems.

1. Introduction

Automatic summarization has received considerable attention in the past several years. Because it is a relatively old field (Luhn, 1958; Edmundson, 1969), the rapid growth of available documents in digital format was like "a breath of fresh air" to the field and has highlighted the importance of developing specific tools to find relevant information. Arabic documents are no exception. Indeed, Arabic content on the Internet has undergone a constant expansion: Arabic websites were ranked eighth at 3% [1] in April 2013, and there were more than 255 thousand Arabic Wikipedia articles in December 2013. Moreover, Arabic is the fifth most spoken language in the world,[2] and Arabic Internet users were ranked seventh at 3% in May 2011.

Recently, new tasks and challenges arose such as multi-document, multilingual and guided and updated summaries and gave a new boost to the automatic summarization field. Cross-lingual and multilingual summarizations received

* Corresponding author.
E-mail addresses: h_oufaida@esi.dz (H. Oufaida), onouali@mail.cerist.dz (O. Nouali), blache@lpl-aix.fr (P. Blache).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

[1] http://en.wikipedia.org/w/index.php?title=Languages_used_on_the_Internet&oldid=581612018.
[2] http://www.ethnologue.com/statistics/size.

http://dx.doi.org/10.1016/j.jksuci.2014.06.008

considerable attention and made some interesting multilingual datasets available.

The goal of text summarization is to produce a condensed version of one or more texts using computer techniques. This will help the reader to decide if a document contains needed information with minimum effort and time loss. Thus, a good summarizer should find key information and omit redundant information. For example, if we perform a search on a particular subject, such as "Arabic text summarization", we will find a multitude of documents. Among them, some are very interesting, and others are less relevant. Sorting these documents is a tedious task, which will take significant time. Thus, tools such as single- and multi-document summarization systems are quite useful.

Early summarization systems used natural language processing (NLP)-based techniques in a bid to *understand* the source text and *generate* new sentences to form an *abstract*: paraphrasing identification and information fusion, for example Barzilay and McKeown (2005). In a few other cases, NLP techniques were used to identify salient sentences, such as the use of rhetorical analysis RST (Marcu, 1998). However, these techniques are not yet mature; they still require heavy NLP processing often based on limited and language-dependent resources. Moreover, considerable work remains for Arabic NLP to reach the actual level of English NLP tools, for example: sentence splitting, tokenization, part-of-speech tagging, named entity recognition, and anaphora resolution. These basic NLP tasks have relatively acceptable performance for English and were used in many state of the art summarization systems. However, Arabic NLP systems are still in an early stage. Thus, developing effective Arabic summarization systems based on heavy NLP techniques is not yet possible.

Recently, statistical techniques have proved their performance and gained more ground. In this paper, we propose an Arabic statistical summarization method, which uses light language-dependent information. Our method extracts relevant sentences from single and multiple Arabic documents by maintaining minimum redundancy and maximum relevance. To achieve this, we first proceed to document preprocessing: sentence splitting, tokenization, stop words removal and root extraction. Second, we build a [Sentences $\times$ Terms] matrix, where each entry corresponds to the term's weight in the sentence. Third, we build a sentence-to-sentence similarity/distance matrix and perform clustering to put similar sentences in the same cluster. Fourth, we apply an adapted discriminant analysis method: minimum redundancy maximum relevance (mRMR) (Peng et al., 2005) to select most relevant terms from input Arabic document/documents with minimum redundancy. Then, a score is assigned to each sentence based on the new mRMR weights of its terms. Finally, n sentences are selected to construct the output summary, n depending on the required summary size.

This paper is organized as follows: we first introduce a brief review of related work in the field in Section 2. Second, Section 3 describes the original mRMR method, and Sections 4 and 5 present our mRMR adaptation to the Arabic summarization task. Section 6 details our experiments in both single- and multi-document summarizations. Finally, we present our paper's conclusions and several interesting perspectives in Section 7.

2. Previous work

Identifying relevant sentences is the key element for extractive summarization systems. Statistical techniques assign a score to each sentence. Computing this score varies from the use of positional- and frequency-based information to the use of topic signatures, abstractive terms and sentence recommendation.

Compared to English document summarization, very few works have been performed for Arabic document summarization. To the best of our knowledge, Douzidia and Lapalme (2004) was the first work in the Arabic summarization field. It uses classical sentence scoring features: sentence position, terms frequency, title words (Luhn, 1958) and cue words (Edmundson, 1969): for example, "تجدر الإشارة إلى" or "we underline" and "وبناء على ما سبق" or "in conclusion" are used to capture sentences in which the author has emphasized. Douzidia and Lapalme (2004) used a weighted linear combination of these features (which is often the case) to score sentences (1). The system uses character level normalization, a light lemmatization (simple prefix and suffix removal) and a rule-based sentence compression component to reduce several indirect discourse parts, such as name substitution.

$$Sc = \alpha_1 Sc_{lead} + \alpha_2 Sc_{title} + \alpha_3 Sc_{cue} + \alpha_4 Sc_{tf.idf} \quad (1)$$

Sobh et al. (2006) used additional features: the sentence and paragraph length, the sum of sentence cosine similarity values with the rest of sentences and some POS-based features: number of infinitives, verbs, *Marfo'at*, and *identified* words and whether the sentence includes a digit or not. Next, the authors apply three classifiers: two basic classifiers (Bayesian, genetic programing) and a combination of these two as a dual one. Among this work's interesting conclusions is the fact that on the basis of an evaluation on 213 articles from "Al Ahram" newspaper, features were classified into three categories: strong, weak and intermediate. Strong features were: the sentence's term weight, length and similarity sum.

Schlesinger et al. (2008) used a rule-based sentence splitter and six-gram tokenization to process Arabic texts. The authors outlined the lack of resources to accomplish these two tasks. Motivated by Douzidia and Lapalme's (2004) good evaluation results, authors use original Arabic texts rather than English machine translated texts, a unigram language model and signature terms to score sentences. Once top ranked sentences are extracted, the system replaced Arabic sentences with the corresponding machine translated (MT) sentences.

El-Haj et al. (2011a) proposed two Arabic text summarization systems: AQBTSS, a query-based Arabic single-document summarizer, and ACBTSS, a concept-based summarizer. The first, AQBTSS, attempts to fit the generated summary to a specific Arabic user's query while the second, ACBTSS, attempts to fit it against a bag-of-words representation of a certain concept. The two systems use a vector space model to score sentences. Interestingly, for the concept-based summarizer, the author dressed a list of 10 concepts with the corresponding most frequent terms. This was defined on the basis of a 10,250 Arabic newspaper articles corpus with approximately 850 documents per concept. On the basis of this work, the KALIMAT 1.0[3] corpus was recently released; it is a free and publically available dataset of 20,291 newspaper articles which

[3] http://sourceforge.net/projects/kalimat/.