Expert Systems with Applications 36 (2009) 7764–7772

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A new sentence similarity measure and sentence based extractive technique for automatic text summarization

Ramiz M. Aliguliyev *

Institute of Information Technology of National Academy of Sciences of Azerbaijan, 9, F.Agayev str., AZ1141 Baku, Azerbaijan

ARTICLE INFO

Keywords:
Similarity measure
Text mining
Sentence clustering
Summarization
Evolution algorithm
Sentence extractive technique

ABSTRACT

The technology of automatic document summarization is maturing and may provide a solution to the information overload problem. Nowadays, document summarization plays an important role in information retrieval. With a large volume of documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. Document summarization is a process of automatically creating a compressed version of a given document that provides useful information to users, and multi-document summarization is to produce a summary delivering the majority of information content from a set of documents about an explicit or implicit main topic. In our study we focus on sentence based extractive document summarization. We propose the generic document summarization method which is based on sentence clustering. The proposed approach is a continue sentence-clustering based extractive summarization methods, proposed in Alguliev [Alguliev, R. M., Aliguliyev, R. M., Bagirov, A. M. (2005). Global optimization in the summarization of text documents. *Automatic Control and Computer Sciences 39*, 42–47], Aliguliyev [Aliguliyev, R. M. (2006). A novel partitioning-based clustering method and generic document summarization. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI–IAT 2006 Workshops) (WI–IATW'06)*, 18–22 December (pp. 626–629) Hong Kong, China], Alguliev and Alyguliev [Alguliev, R. M., Alyguliev, R. M. (2007). Summarization of text-based documents with a determination of latent topical sections and information-rich sentences. *Automatic Control and Computer Sciences 41*, 132–140] Aliguliyev, [Aliguliyev, R. M. (2007). Automatic document summarization by sentence extraction. *Journal of Computational Technologies 12*, 5–15.]. The purpose of present paper to show, that summarization result not only depends on optimized function, and also depends on a similarity measure. The experimental results on an open benchmark datasets from DUC01 and DUC02 show that our proposed approach can improve the performance compared to sate-of-the-art summarization approaches.

1. Introduction

The technology of automatic document summarization is maturing and may provide a solution to the information overload problem (Hahn & Mani, 2000; Mani & Maybury, 1999). Nowadays, document summarization plays an important role in information retrieval (IR). With a large volume of documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents (Gong & Liu, 2001). Text summarization is the process of automatically creating a compressed version of a given text that provides useful information to users, and multi-document summarization is to produce a summary delivering the majority of information content from a set of documents about an explicit or implicit main topic (Wan, 2008). Authors of the paper (Radev, Hovy, & McKeown, 2002) provide the following definition for a summary:"A summary can be loosely defined as a text that is produced from one or more texts that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. Text here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc. The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its informativeness. Luckily, information content in a document appears in bursts, and one can therefore distinguish between more and less informative segments. Identifying the informative segments at the expense of the rest is the main challenge in summarization". Jones (2007) assumes a tripartite processing model distinguishing three stages: source text interpretation to obtain a source representation, source representation transformation to summary representation, and summary text generation from the summary representation.

* Fax: +994 12 439 61 21.
E-mail address: a.ramiz@science.az

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.11.022

A variety of document summarization methods have been developed recently. The paper (Jones, 2007) reviews research on automatic summarizing over the last decade. This paper reviews salient notions and developments, and seeks to assess the state-of-the-art for this challenging natural language processing (NLP) task. The review shows that some useful summarizing for various purposes can already be done but also, not surprisingly, that there is a huge amount more to do.

Sentence based extractive summarization techniques are commonly used in automatic summarization to produce extractive summaries. Systems for extractive summarization are typically based on technique for sentence extraction, and attempt to identify the set of sentences that are most important for the overall understanding of a given document. In paper Salton, Singhal, Mitra, and Buckley (1997) proposed paragraph extraction from a document based on intra-document links between paragraphs. It yields a text relationship map (TRM) from intra-links, which indicate that the linked texts are semantically related. It proposes four strategies from the TRM: bushy path, depth-first path, segmented bushy path, augmented segmented bushy path. An improved version of this approach proposed in paper (Alguliev & Aliguliyev, 2005).

In our study we focus on sentence based extractive summarization. We propose the generic document summarization method which is based on sentence-clustering. The proposed approach is a continue sentence-clustering based extractive summarization methods, proposed in Alguliev, Aliguliyev, and Bagirov (2005), Aliguliyev (2006), Alguliev and Alyguliev (2007), Aliguliyev (2007). The purpose of present paper to show, that summarization result not only depends on optimized function, and also depends on a similarity measure. The experimental results on an open benchmark datasets from DUC01 and DUC02 (http://duc.nist.gov) show that our proposed approach can improve the performance compared to sate-of-the-art summarization approaches.

The rest of this paper is organized as follows: Section 2 introduces related works. The proposed sentence-clustering based approach for generic single-document summarization is presented in Section 3. The differential evolution algorithm for optimization procedure is given in Section 4. The extractive technique is represented in Section 5. The experiments and results are given in Section 6. Lastly, we conclude our paper in Section 7.

2. Related work

Generally speaking, the methods can be either extractive summarization or abstractive summarization. Extractive summarization involves assigning salience scores to some units (e.g. sentences, paragraphs) of the document and extracting the sentences with highest scores, while abstraction summarization (e.g. http://www1.cs.columbia.edu/nlp/newsblaster/) usually needs information fusion, sentence compression and reformulation (Mani & Maybury, 1999; Wan, 2008).

Sentence extraction summarization systems take as input a collection of sentences (one or more documents) and select some subset for output into a summary. This is best treated as a sentence ranking problem, which allows for varying thresholds to meet varying summary length requirements. Most commonly, such ranking approaches use some kind of similarity or centrality metric to rank sentences for inclusion in the summary – see, for example, Alguliev and Aliguliyev (2005), Alguliev et al. (2005), Aliguliyev (2006), Alguliev and Alyguliev (2007), Erkan and Radev (2004), Aliguliyev (2007), Fisher and Roark (2006), Radev, Jing, Stys, and Tam (2004), Salton, Singhal, Mitra and Buckley, 1997.

The centroid-based method (Erkan & Radev, 2004; Radev et al., 2004) is one of the most popular extractive summarization methods. MEAD (http://www.summarization.com/mead/) is an implementation of the centroid-based method for either single- or multi-document summarizing. It is based on sentence extraction. For each sentence in a cluster of related documents, MEAD computes three features and uses a linear combination of the three to determine what sentences are most salient. The three features used are centroid score, position, and overlap with first sentence (which may happen to be the title of a document). For single-documents or (given) clusters it computes centroid topic characterizations using tf–idf-type data. It ranks candidate summary sentences by combining sentence scores against centroid, text position value, and tf–idf title/lead overlap. Sentence selection is constrained by a summary length threshold, and redundant new sentences avoided by checking cosine similarity against prior ones (Zajic, Dorr, Lin, & Schwartz, 2007).

In the past, extractive summarizers have been mostly based on scoring sentences in the source document. In paper (Shen, Sun, Li, Yang, & Chen, 2007) each document is considered as a sequence of sentences and the objective of extractive summarization is to label the sentences in the sequence with 1 and 0, where a label of 1 indicates that a sentence is a summary sentence while 0 denotes a non-summary sentence. To accomplish this task, is applied conditional random field, which is a state-of-the-art sequence labeling method (Lafferty, McCallum, & Pereira, 2001). In paper Wan, Yang, and Xiao (2007) proposed a novel extractive approach based on manifold–ranking of sentences to query-based multi-document summarization. The proposed approach first employs the manifold–ranking process to compute the manifold–ranking score for each sentence that denotes the biased information-richness of the sentence, and then uses greedy algorithm to penalize the sentences with highest overall scores, which are deemed both informative and novel, and highly biased to the given query.

The summarization techniques can be classified into two groups: supervised techniques that rely on pre-existing document-summary pairs, and unsupervised techniques, based on properties and heuristics derived from the text. Supervised extractive summarization techniques treat the summarization task as a two-class classification problem at the sentence level, where the summary sentences are positive samples while the non-summary sentences are negative samples. After representing each sentence by a vector of features, the classification function can be trained in two different manners (Mihalcea & Ceylan, 2007). One is in a discriminative way with well-known algorithms such as support vector machine (SVM) (Yeh, Ke, Yang, & Meng, 2005). Many unsupervised methods have been developed for document summarization by exploiting different features and relationships of the sentences – see, for example Alguliev and Aliguliyev (2005), Alguliev et al. (2005), Aliguliyev (2006), Alguliev and Alyguliev (2007), Aliguliyev (2007), Erkan and Radev (2004), Radev et al. (2004) and the references therein.

On the other hand, summarization task can also be categorized as either generic or query-based. A query-based summary presents the information that is most relevant to the given queries (Dunlavy, O'Leary, Conroy, & Schlesinger, 2007; Fisher & Roark, 2006; Li, Sun, Kit, & Webster, 2007; Wan, 2008) while a generic summary gives an overall sense of the document's content (Alguliev & Aliguliyev, 2005; Alguliev et al., 2005; Aliguliyev, 2006; Alguliev & Alyguliev, 2007; Aliguliyev, 2007; Dunlavy et al., 2007; Gong & Liu, 2001; Jones, 2007; Li et al., 2007; Salton et al., 1997; Wan, 2008). The QCS system (Query, Cluster, and Summarize) (Dunlavy et al., 2007) performs the following tasks in response to a query: retrieves relevant documents; separates the retrieved documents into clusters by topic, and creates a summary for each cluster. QCS is a tool for document retrieval that presents results in a format so that a user can quickly identify a set of documents of interest. In paper McDonald and Chen (2006) are developed a generic, a query-based, and a hybrid summarizer, each with