Expert Systems With Applications 65 (2016) 68–86

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization

Hilário Oliveira[a,*], Rafael Ferreira[a,b], Rinaldo Lima[a,b], Rafael Dueire Lins[a,b], Fred Freitas[a], Marcelo Riss[c], Steven J. Simske[d]

[a] Federal University of Pernambuco, Recife, Brazil
[b] Federal Rural University of Pernambuco, Recife, Brazil
[c] HP Brazil, Porto Alegre, Brazil
[d] HP Labs., Fort Collins, CO 80528, USA

ARTICLE INFO

Article history:
Received 22 April 2016
Revised 20 June 2016
Accepted 4 August 2016
Available online 9 August 2016

Keywords:
Automatic text summarization
Extractive-based approaches
Sentence scoring techniques
Single-document summarization
Multi-document summarization

ABSTRACT

The volume of text data has been growing exponentially in the last years, mainly due to the Internet. Automatic Text Summarization has emerged as an alternative to help users find relevant information in the content of one or more documents. This paper presents a comparative analysis of eighteen shallow sentence scoring techniques to compute the importance of a sentence in the context of extractive single- and multi-document summarization. Several experiments were made to assess the performance of such techniques individually and applying different combination strategies. The most traditional benchmark on the news domain demonstrates the feasibility of combining such techniques, in most cases outperforming the results obtained by isolated techniques. Combinations that perform competitively with the state-of-the-art systems were found.

1. Introduction

Alvin Toffler in 1970 coined the expression “information overload”, when he predicted that the exponential growth in the amount of information being produced would eventually cause people problems (Toffler, 1970). Such a scenario is the current reality. The Web, for example, allows creating, sharing, and accessing a vast amount of digital information, particularly textual documents such as news articles, online books, blogs, emails, scientific papers, tweets, among others. Despite the development of web search engines, sieving useful information from such massive volume of data is still a hard task, unfeasible to be performed manually. In such a context, there is a constant interest in tools capable of retrieving, classifying, and summarizing such information in an efficient manner.

In this scenario, Automatic Text Summarization (ATS) arises as a possible feasible solution to reduce users’ time in identifying the most relevant information from a single document or a collection of text documents. ATS can be defined as the process of creating automatically a condensed version (summary) from a single- or multi-documents, while keeping their key information (Nenkova & McKeown, 2012). According to the Cambridge dictionary[1], a *summary* can be defined as “a short description that gives the main facts or ideas about something”. Based on this definition, an ATS system should deal with two fundamental issues (Saggion & Poibeau, 2013) **(i)** How to select the most relevant information, and **(ii)** Expressing the selected information in a compact way.

In general, ATS approaches have been classified in two major subfields *extractive* and *abstractive* (Lloret & Palomar, 2012). Extractive methods select the most relevant sentences in a document and use them to create the summary. Extractive summaries, due to selecting sentences *verbatim* from the original text, often present problems such as lack of coherence (Christensen, Soderland, Bansal, & Mausam, 2014), e.g., broken coreferences. On the other hand, *abstractive* approaches (Banerjee, Mitra, & Sugiyama, 2015; Khan, Salim, & Kumar, 2015) focus on selecting the most salient information fragments of a document and expressing them in a new form using operations such as sentence compression (Zajic, Dorr, Lin, & Schwartz, 2007) and merging (Filippova, 2010). Abstractive methods require a deep Natural Language Processing (NLP) analysis such as semantic representation and natural language generation. There are different methods to deal with abstractive summariza-

* Corresponding author.
E-mail addresses: htao@cin.ufpe.br (H. Oliveira), rflm@cin.ufpe.br (R. Ferreira), rjl4@cin.ufpe.br (R. Lima), rdl@cin.ufpe.br (R.D. Lins), fred@cin.ufpe.br (F. Freitas), marcelo.riss@hp.com (M. Riss), steven.simske@hp.com (S.J. Simske).

[1] http://dictionary.cambridge.org/us/

http://dx.doi.org/10.1016/j.eswa.2016.08.030
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

tion such as semantic graph based method (Khan et al., 2015; Liu, Flanigan, Thomson, Sadeh, & Smith, 2015), multimodal semantic model (Greenbacker, 2011), among others. However, those methods are usually not completely automatic, as they require resources previously built, and demand a high computational effort. Due to these facts, extractive methods are more widely investigated today.

This work focuses on investigating extractive-based methods. Usually, this kind of method is performed in three steps (Nenkova & McKeown, 2012): **(i)** creation of an intermediate representation; **(ii)** computation of sentence salience (importance) scoring; and **(iii)** summary generation. Text documents are in an unstructured form; thus, it is necessary to pre-process these documents and represent them in a structured fashion. The first step usually involves some NLP tasks such as dividing the text into paragraphs, sentences, tokens, stopword removal, stemming, among others. Strategies to represent the main topic discussed in the document are also performed. Such strategies may compute the frequency or co-occurrence of words, sentence lengths and location into the document, presence of cue phrases, among others. The second step tries to estimate which sentences are the most relevant, based on the representation previously created. For each sentence a score is created, as a measure of its relevance. Finally, in the third step, the top-ranked sentences are selected to create the final summary. One of the most challenging issues in this step is to avoid redundancy, i.e., sentences with overlapping information in the summary.

Several extractive summarization techniques have been proposed and evaluated to estimate the relevance of a sentence. The techniques range from simple heuristics such as sentence position, sentence similarity with the document title, and statistical-based methods such as word frequency and co-occurrence. More sophisticated approaches such as clustering-based methods (Wan & Yang, 2008), graph-based methods (Mihalcea & Tarau, 2004), combinatorial optimization-based methods such as Integer Linear Programming (ILP) (Gillick & Favre, 2009; Li, Liu, & Zhao, 2015), supervised machine-learning approaches (Fattah, 2014), hierarchical approaches (Christensen et al., 2014), methods based on information extraction (Binh Tran, 2013), event-based summarization (Glavaš & Šnajder, 2014; Marujo et al., 2015), and semantic analysis (Baralis, Cagliero, Jabeen, Fiori, & Shah, 2013) have also been investigated.

This paper aims to investigate the performance of several shallow sentence salience scoring techniques widely used and referenced in the literature in the context of single- and multi-document summarization on the news domain. Different strategies to combine the individual scores of the techniques seeking to outperform the results obtained are also analyzed. The focus is in shallow sentence scoring techniques, i.e., heuristics or methods that are simple to implement and do not require massive computational effort to be computed. Experiments used the CNN corpus and the traditional DUC 2001–2004 datasets on both single- and multi-document summarization tasks. The results demonstrate that the performance of the features investigated and the combinations identified in terms of the most commonly used ROUGE evaluation measures (Lin, 2004) are feasible to identify the main gist of the documents, achieving comparable results against the state-of-the-art summarizers.

The main contributions of this paper are:

- Investigating several shallow sentence scoring techniques and ensemble strategies considering single- and multi-document summarization tasks in the most used datasets on the news domain.
- Showing that combining shallow sentence scoring techniques leads to an improvement in the performance of the summarization tasks based on the traditional ROUGE scores, in both single- and multi-document summarization tasks.
- Identifying combinations that perform competitively against several state-of-the-art systems on various benchmark datasets.

The remaining of this paper is organized as follows. Section 2 briefly presents the related works that assess several sentence salience scoring techniques. Section 3 introduces the summarization process adopted and the sentence salience scoring methods investigated in this work. Section 4 presents the results of the performed experiments. Finally, Section 5 presents the conclusions and draws lines for further work.

2. Related work

This section focuses on presenting the works that conducted studies either to compare the performance of the different sentence scoring techniques or the strategies to combine them in the context of extractive document summarization. The reader interested in an overview of ATS techniques may refer to the recent surveys in the field (Gambhir & Gupta, 2016; Lloret & Palomar, 2012; Nenkova & McKeown, 2012; Saggion & Poibeau, 2013; Torres-Moreno, 2014).

Meena and Gopalani (2014) investigated seven linear combinations using nine different sentence scoring techniques: Term Frequency - Inverse Document Frequency (TF-IDF), word co-occurrence, sentence centrality, sentence location, named entities frequency, the presence of positive and negative keywords, Textrank, and proper nouns frequency. The experiments used only ten documents of the Document Understanding Conferences (DUC) 2002 corpus[2]. The authors compared the performance of the combinations using the traditional ROUGE toolkit (Lin, 2004), which is extensively used to evaluate ATS systems. In a later work, Meena, Deolia, and Gopalani (2015) also investigated all possible linear combinations of six sentence scoring techniques. The authors assessed each combination using ten documents of the DUC 2002 dataset.

Ferreira et al. (2013) conducted an extensive assessment of seventeen sentence salience scoring techniques such as word frequency, TF-IDF, sentence centrality, sentence position, among others. The authors investigated the performance of these techniques individually using three different corpora on news, blog, and scientific paper domains. The authors complemented that study in another paper (Ferreira et al., 2014) analyzing the performance of ten proposed linear combinations using the seventeen features previously investigated on the three cited corpora. In both studies, each scoring technique and the proposed combinations were compared using the ROUGE toolkit and by counting the overlap of the sentences chosen by the methods in the automatically generated summaries and their *gold standards*, the extractive summaries created by experts using a computer-assisted methodology.

Other works addressed the sentence salience extraction task as a classification problem. They investigated the performance of many sentence scoring techniques as input features to Machine Learning (ML) algorithms. In such an approach, the problem consists of creating a classification model that estimates if a sentence should be included in the summary or not. Neto, Freitas, and Kaestner (2002) assessed thirteen sentence scoring techniques such as sentence length, sentence position, similarity to the title, among others, as input features to two well-known ML classification algorithms C4.5 (Quinlan, 1992) and Naive Bayes (John & Langley, 1995). Leite and Rino (2008) investigated the performance of several features based on both linguistic and statistical information, and complex networks to ATS using different ML algorithms. Fattah (2014) investigated eight shallow

[2] http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html.