Applied Soft Computing 30 (2015) 737–747

Contents lists available at ScienceDirect

Applied Soft Computing

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

A framework for multi-document abstractive summarization based on semantic role labelling

Atif Khan[a,*], Naomie Salim[a], Yogan Jaya Kumar[b]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] Faculty of Information and Communication Technology, University Teknikal Malaysia Melaka, 76100 Melaka, Malaysia

ARTICLE INFO

Article history:
Received 18 May 2014
Received in revised form 7 September 2014
Accepted 29 January 2015
Available online 20 February 2015

Keywords:
Abstractive summary
Semantic role labeling
Semantic similarity measure
Language generation
Genetic algorithm

ABSTRACT

We propose a framework for abstractive summarization of multi-documents, which aims to select contents of summary not from the source document sentences but from the semantic representation of the source documents. In this framework, contents of the source documents are represented by predicate argument structures by employing semantic role labeling. Content selection for summary is made by ranking the predicate argument structures based on optimized features, and using language generation for generating sentences from predicate argument structures. Our proposed framework differs from other abstractive summarization approaches in a few aspects. First, it employs semantic role labeling for semantic representation of text. Secondly, it analyzes the source text semantically by utilizing semantic similarity measure in order to cluster semantically similar predicate argument structures across the text; and finally it ranks the predicate argument structures based on features weighted by genetic algorithm (GA). Experiment of this study is carried out using DUC-2002, a standard corpus for text summarization. Results indicate that the proposed approach performs better than other summarization systems.

1. Introduction

The information on Web is growing at exponential pace. In the current era of information overload, multi-document summarization is an essential tool that creates a condensed summary while preserving the important contents of the source documents. The automatic multi-document summarization of text is a major task in the field of natural language processing (NLP) and has gained more consideration in recent years [1]. One of the problems of information overload is that many documents share similar topics, which creates both difficulties and opportunities for natural language systems. On one hand, the similar information conveyed by several different documents, causes difficulties for the end users, as they have to read the same information repeatedly. On the other side, such redundancy can be used to identify accurate and significant information for applications such as summarization and question answering. Thus, summaries that synthesize common information across many text documents would be useful for users and reduce their time for finding the key information in the text documents. Such a summary would significantly help users interested in single event described in many news documents [1]. In this paper, we propose a framework that will automatically fuse similar information across multiple documents and use language generation to produce a concise abstractive summary.

Two approaches are employed to multi-document summarization: extractive and abstractive. Most of the studies have focused on extractive summarization, using techniques of sentence extraction [2], statistical analysis [3], discourse structures and various machine learning techniques [4]. On other hand, abstractive summarization is a challenging area and dream of researchers [5], because it requires deeper analysis of the text and has the capability to synthesize a compressed version of the original sentence or may compose a novel sentence not present in the original source. The goal of abstractive summarization is to improve the focus of summary, reduce its redundancy and keeps a good compression rate [6].

Past literature shows that there have been a few research efforts made toward abstractive summarization. Many researchers have tried to generate abstractive summaries using various methods. These abstractive methods can be grouped into two categories: Linguistic (Syntactic) based approach and Semantic based approach. Linguistic (Syntactic) based approach employs syntactic parser to analyze and represent the text syntactically. Usually, in this approach, verbs and nouns identified by syntactic parser are used for text representation and further processed to generate the

[*] Corresponding author. Tel.: +60 1126799325.
E-mail addresses: atifkhan@icp.edu.pk (A. Khan), naomie@utm.my (N. Salim), yogan@utem.edu.my (Y. Jaya Kumar).

http://dx.doi.org/10.1016/j.asoc.2015.01.070
1568-4946/© 2015 Elsevier B.V. All rights reserved.

abstractive summary. On other hand, semantic based approach aims to produce abstractive summary from semantic representation of document text. Different semantic representations of text used in the literature are ontology based and template based representation. Titov and Klementiev [7] made a distinction between syntactic and semantic representation of sentence. They addressed that syntactic analysis is far away from representing the meaning of sentences. In particular, syntactic analysis does not define who did what to whom (and how, when. . .).

All the linguistic based approaches [1,6,8,9] proposed for the abstractive summarization rely on the syntactic representation of source document. These approaches employ syntactic parser to represent the source text syntactically. The major limitations of these approaches is the lack of semantic representation of source text. Since abstractive summarization requires deep analysis of text, therefore, semantic representation of source text will be a more suitable representation.

On other hand, a few semantic based approaches have also been proposed for abstractive summarization and are briefly discussed as follows. A multi-document summarization system, GISTEXTER, presented in [10] exploits template based method to produce abstractive summary from multiple newswire/newspaper documents depending on the output of the information extraction (IE) system. This approach use template for topic representation of document. The major limitation observed in this approach was that linguistic patterns and extraction rules for template slots were manually created by humans, which is time consuming. Moreover, this method could not handle or capture the information about similarities and differences across multiple documents.

A fuzzy ontology based approach [11] was proposed for Chinese news summarization to model uncertain information and hence can better describe the domain knowledge. This approach has several limitations. First, domain ontology and Chinese dictionary has to be defined by a domain expert which is time consuming. Secondly, this approach is limited to Chinese news, and might not be applicable to English news.

The methodology proposed in [12] generates short and well-written abstractive summaries from clusters of news articles on same event using abstraction schemes. The abstraction scheme uses a rule based information extraction module, content selection heuristics and one or more patterns for sentence generation. The drawback of the methodology was that information extraction (IE) rules and generation patterns were written by hand, which was extremely time consuming.

A framework proposed by [13] generates abstractive summary from a semantic model of a multimodal document. The semantic model is constructed using knowledge representation based on objects (concepts) organized by ontology. The important concepts represented by semantic model are rated based on information density metric and expressed as sentences using the available phrasings stored for each concept in a semantic model. The limitation of this framework was that it relies on manually built ontology, which is time consuming. Secondly, it is manually evaluated by humans i.e. human judges were asked to assess the quality of system summary by comparing it with other summary generated by traditional extraction methods.

The abstractive approach presented by [14] summarizes a document by creating a Rich Semantic Graph (RSG) for the source document. Rich semantic graph is an ontology based representation i.e. graph nodes are the instances of ontology noun and verb classes. Like other approaches, the limitation of this approach was that it also relies on manually built ontology, which is time consuming.

The major limitation of all the semantic based approaches for abstractive summarization is that they are mostly dependent on human expert to construct domain ontology and rules; which is a drawback for an automatic summarization system. Our work, in contrast, aims to treat this limitation by using semantic role labeling (SRL) technique to build semantic representation from the document text automatically.

SRL has been widely applied in text content analysis tasks such as text retrieval [15], information extraction [16], text categorization [17] and sentiment analysis [18]. In the area of text summarization, [19] introduced a work that combined semantic role labeling with general statistic method (GSM) to determine important sentences for single document extractive summary. At first, they employed SRL and semantic similarity measure to compute the sentence similarity score. Secondly, the general statistic method was used to computes the sentence score based on features, without taking into account their weights. Finally, the sentence scores obtained from both methods are combined to assign the overall score to each sentence in the document and the top ranked sentences are extracted according to 20% compression rate. However, our work is different from [19] in the following manner. We focus on multi-document abstractive summarization while [19] focus on single document extractive summarization. We employ SRL in second phase of the framework, for semantic representation of text in the document collection. On other hand, [19] employed SRL and semantic similarity measure to compute sentence semantic similarity score. To the best of our knowledge, semantic role labeling (SRL) technique, which exploits semantic role parser, has not been employed for the semantic representation of text in multi-document abstractive summarization.

Therefore, this study proposes a framework that will employ SRL for semantic representation of text in order to generate a good abstractive summary. The framework for multi-document abstractive summarization presented in this study, is different from previous abstractive summarization approaches in a few aspects. First, it employs semantic role labeling to extract predicate argument structure (semantic representation) from the contents of input documents. Secondly, it analyzes the text semantically by utilizing semantic similarity measure in order to cluster semantically similar predicate argument structures across the text; and finally it ranks the predicate argument structures based on the features weighted and optimized by genetic algorithm; since text features are sensitive to the quality of the generated summary.

The rest of this paper is organized as follows: Section 2 outlines the proposed framework. The evaluation of the framework is given in Section 3. Finally we end with conclusion in Section 4.

2. Framework for multi-document abstractive summarization

2.1. Overview of approach

The framework of our proposed approach is illustrated in Fig. 1. Given a document collection that need to be summarized, first of all, we split the document collection into sentences in such a way that each sentence is preceded by its corresponding document number and sentence position number. Next, SENNA semantic role labeler [20] is employed to extract predicate argument structure from each sentence in the document collection. In semantic similarity matrix computation phase (as discussed in Section 2.3), the similarities of predicate argument structures (PASs) are computed by comparing them pair wise based on Jiang's semantic similarity measure [21] and edit distance algorithm. Once the similarity matrix for PASs is obtained, we perform agglomerative hierarchical clustering (HAC) algorithm based on average linkage method to cluster semantically similar predicate argument structures. The number of clusters will depend on compression rate of summary. Section 2.4 will describe this phase in detail. The PASs in each cluster are scored based on features, weighted and optimized by genetic algorithm and the top