Expert Systems with Applications 40 (2013) 4346–4352

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Rhetorics-based multi-document summarization ☆

John Atkinson *, Ricardo Munoz

Department of Computer Sciences, Universidad de Concepcion, Chile

ARTICLE INFO

Keywords:
Multi-document summarization
Semantic analysis
Statistical language models
Discourse Models
Machine Learning
Rhetorical roles

ABSTRACT

In this paper, a new multi-document summarization framework which combines rhetorical roles and corpus-based semantic analysis is proposed. The approach is able to capture the semantic and rhetorical relationships between sentences so as to combine them to produce coherent summaries. Experiments were conducted on datasets extracted from web-based news using standard evaluation methods. Results show the promise of our proposed model as compared to state-of-the-art approaches.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Multi-document summarization is the process of generating a generic summary by reducing documents in size while retaining the main characteristics of the original documents. Since one of the problems of data overload is caused by the fact that many documents share similar topics, automatic multi-document summarization has became very popular in recent years. With the explosive increase of documents on the web, there are various summarization applications. For example, the informative snippets generation in web search can assist users in further exploring, and in a question/answer system, a question-based summary is often required to provide information asked in the question. Another example is short summaries for news groups in news services, which can facilitate users to better understand the news articles in the group.

There are key issues for multi-document summarization which must be addressed. Firstly, the information contained in different documents often overlaps with each other, hence it is necessary to find an effective way to merge the documents while recognizing and removing redundancy. In order to avoid repetition, humans tend to use different words to describe the same person, the same topic as a story goes on. Thus simple word-matching types of similarity such as *cosine* cannot capture the content similarity. In addition, the sparseness of words between similar concepts make the similarity metric uneven. Another issue is identifying important differences between documents and covering the informative content as much as possible. Current document summarization methods usually involve natural language processing and Machine Learning Techniques, such as unsupervised learning (i.e., clustering), classification, etc. Yet, extracting this kind of key information (i.e., relevant sentences) from multiple documents pose strong challenges in terms of referring expression issues and so, the coherence of the finally generated summary so that this can be easily understood. For example, discourse referent (i.e., pronouns) existing in multiple documents are naturally expressing different entities, so if we are extracting relevant but sometimes, isolated sentences from different documents, how can we preserve this coherence in terms of the entities the summary is referring to? Kou, Takao, and Isamu (2006) and Jurafsky and Martin (2008). For this, Machine Learning Techniques and discourse-level processing methods should be considered to address the major problems. Some approaches apply *Natural-Language Processing* (NLP) to generate coherent texts, whereas others use information retrieval techniques to extract relevant information from full documents but missing discourse information that allows a summary to be understood Saravanan and Ravindran (2010).

In general, summarization methods are effective for coherent documents having certain given structure and genre such as scientific articles, legal documents, news, etc. Saravanan and Ravindran (2010). However, when using multimedia and informal information such as that available on the web (i.e., webpages), traditional multi-document summarization methods are not effective enough as the approaches rely on explicit syntactical semantical markers. This kind of document usually contain data or metadata describing points in which its structure is not fully coherent. Hence, there is no current approaches to generate full and coherent summaries from multiple websites dAcierno, Moscato, Persia, Picariello, and Penta (2010).

Accordingly, in this work a novel approach to multi-document summarization from the web which combines discourse-level knowledge and corpus-based semantic analysis is proposed. Our mail claim is that our approach using rhetorical knowledge may generate better quality summaries than state-of-the-art techniques. Thus, this paper is organized as follows: Section 2 discusses

☆ This research is sponsored by the technology project **FONDEF**, no. D08I1155: "Semantic Identification and Automatic Synthesis of Educational Materials for Specialised Domains".

* Corresponding author. Tel.: +56 41 2204305.
E-mail address: atkinson@inf.udec.cl (J. Atkinson).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.01.017

the fundamentals and state-of-the-rt methods for multi-document summarization, Section 3 describes the proposed rhetorics-based model for summarizing multiple web documents, Section 4 discussed the main experiments carried out and the obtained results to assess the method with documents extracted from multiple sources (newspapers on the web), and finally, Section 5 highlights the main conclusions of this research.

2. Related work

Basically, the core of the text summarization methods relies on the appropriate selection of relevant sentences from a full document so as to build a summary. Generally, it uses information containing certain linguistic roles Jurafsky and Martin (2008). In order to extract and/or infer basic discourse-level knowledge from texts such as rhetorical functions, referring expressions, etc., there are two groups of basic techniques:

(1) *Tree search algorithms*: this kind of approach builds syntactical representations for sentences, which are connected by discourse markers. The overall tree becomes an implicit discourse representation model from which further inferences can be made. The aim of the algorithm is to resolve references to entities existing in the discourse tree.
(2) *Centering based techniques*: unlike the previous approach, this kind of method build an explicit discourse model. The strategy uses adjacent utterances from a text so as to identify the main entity/focus. It then creates an ordered list of referencing entities. Highest ranked entities of the second utterance is usually assumed to be the central topic of the list. It allows the method to find the relationship among utterances and the main focus.

Furthermore, information retrieval methods such as *terms distribution* algorithms or *tf-idf* methods have also been explored to extract relevant sentences Saravanan and Ravindran (2010). Here, word distribution related probabilistic data are used as term weight to decide on relevant sentences, by assuming that good word indicators represent good sentences, hence the finally generated summary become very related to the main theme of the document. A popular method based on this principle is called *K-Mixture* and produces fair results in terms of selecting good relevant sentences from a full document Saravanan and Ravindran (2010). The method computes the probability that a word i occurs k times in a document as follows:

$$P_i(k) = (1-r)\delta_{k,0} + \frac{r(s)^k}{s+1(s+1)^k} \quad (1)$$

where $r = t/s$, $s = tx2^{IDF} - 1 = (cf_i - df_i)/df_i$, $t = cf_i/N$ and $IDF = log_2 N/df_i$. In addition, $\delta_{k,0} = 1$ iff $k = 0$, and 0 otherwise, cf_i is the word frequency in a collection of documents, df_i is the documents frequency containing a word and N is the number of documents. From this, sentences obtaining best weighted terms will be relevant to build up the summary. However, extracting relevant sentences is not sufficient to create a coherent summary as many unrelated text may be generated so it will become very difficult to understand the text. It is mainly due to that utterances in a summary must follow some logical order so as to provide a readable discourse. Hence, NLP techniques are required in addition to term distribution based methods. Some discourse-level summarization methods identify rhetorical roles connecting sentences Saravanan and Ravindran (2010). These act as linguistic functions that sentences should fulfill to connect adjacent utterances. Thus, rhetorical roles will depend on the text's structure and domain (i.e., cause-effect relationships, consequence relationships, etc). Once rhetorical roles are identified, the intention of a text may be understood and so key utterance fulfilling certain roles can be extract to create the summary without losing coherence. For this, a summary's structure can be defined as the logically related set of relevant sentences connected by their discourse roles Jurafsky and Martin (2008) and Saravanan and Ravindran (2010). In simple words, this kind of approach can usually be divided into two steps:

(1) Assigning a relevance weight to each sentence within the original text so that those having the highest values become candidate sentences for the final summary.
(2) Recognizing rhetorical roles for previously selected sentences.

A text's argumentative structure can be captured by finding relationship between its rhetorical roles, which are usually seen as a set of 'tags' representing regularities of the intentions of the document's author. For example, for scientific articles, authors use rhetorical roles to refer to the text's aims and the scientific background stated in the document (i.e., *AIM* and *BACKGROUND*, respectively). At the same time, these roles connect adjacent utterances of the text, hence utterances related by the role *AIM* should have higher preference than those related by the role *BACKGROUND*, in terms of ordering in the final summary. Thus, once sentences having specific roles are identified, they can logically put into the summary so as to produce a coherent text.

For texts having a fixed structure and identified coherence relations, a tree-like discourse model can be built. This *Rhetorical Structure Tree* (RS-tree), represents units and relationships implicitly stated in the original text. In order to produce an RS-tree, a text must be segmented into *Elementary Discourse Units* (EDUs). Recognizing the limits of the EDUs, discourse *markers* matching syntactical information and punctuation indicators, can be applied by applying a discourse parser. In general, the task of identifying rhetorical roles can be seen as a text segmentation problem in which sentences boundaries must be detected based on their rhetorical functions Saravanan and Ravindran (2010). It can usually be carried out by using techniques such as classification rule induction Saravanan and Ravindran (2010), Hidden Markov Models (HMMs), Maximum Entropy Models (MEMMs), and *Conditional Random Fields* (CRFs). Rule-based approaches generate a set of rules that can be applied to a set of documents. Each rule represents the mapping of sentences into rhetorical roles. The method then learns rules from a rhetorically-annotated corpus of texts, and it then applies rules to the best matched sentences. In addition, chaining relations (i.e., co-occurrence of roles with sentences) are also verified for each iteration so that additional roles can be introduced. A more effective recent approach for identifying roles uses CRF Saravanan and Ravindran (2010) to tag roles in a sequence of input sentences. The model defines a linear chain containing a sequence of tags and a conditional probability for each of them. Given a sequence of sentences $S = (s_1, \ldots, s_w)$, the conditional probability is calculated as:

$$P_C(L/S) = \frac{1}{Z_s} \exp\left[\sum_{t=1}^{w}\sum_{a} C_a f_a(l_{t-1}, l_t, s)\right] \quad (2)$$

where Z_s is a normalization factor, $f_a(l_{t-1}, l_t, s)$ is a characteristic function, and C_a is a learnt weighting. Characteristic functions represent the model's prediction variables which may depend on the existence of key terms in a sentence, and they are defined as a pair $a = (v, l)$, where v is a binary feature of s_t and l_t is an output state as described in the following equation.

$$f_{(v,l)}(l_t, s_t) = \begin{cases} 1 & \text{If } v(s_t) = 1 \wedge l_t = l. \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

Download English Version:

https://daneshyari.com/en/article/382664

Download Persian Version:

https://daneshyari.com/article/382664

Daneshyari.com