TWO-STAGE SENTENCE SELECTION APPROACH FOR
MULTI-DOCUMENT SUMMARIZATION

Zhang Shu    Zhao Tiejun    Zheng Dequan    Zhao Hua
(Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract Compared with the traditional method of adding sentences to get summary in multi-document summarization, a two-stage sentence selection approach based on deleting sentences in a candidate sentence set to generate summary is proposed, which has two stages, the acquisition of a candidate sentence set and the optimum selection of sentence. At the first stage, the candidate sentence set is obtained by redundancy-based sentence selection approach. At the second stage, optimum selection of sentences is proposed to delete sentences in the candidate sentence set according to its contribution to the whole set until getting the appointed summary length. With a test corpus, the ROUGE value of summaries gotten by the proposed approach proves its validity, compared with the traditional method of sentence selection. The influence of the token chosen in the two-stage sentence selection approach on the quality of the generated summaries is analyzed.

Key words Two-stage; Sentence selection approach; Multi-document summarization

CLC index TP391

DOI 10.1007/s11767-007-0181-2

I. Introduction

In recent years, automatic summarization has drawn a lot of interest in the natural language processing and information retrieval communities. The Document Understanding Conference (DUC)[3] aimed to compile standard training and test collections that enable researchers to participate in large-scale experiments and evaluations on single and multiple document summarization.

Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user and task[2]. One part of the text summarization, multi-document summarization has been developed in recent years, which focused on summarizing important information from multiple documents. There are some representative methods, such as the HMM-based[5], the topic signature-based[6], the cluster-based[5] and the centroid-based[6] multi-document summarizations.

Sentence extraction based multi-document summarization has two main stages, one is the sentence weighting and ranking, the other is the sentence selection. At present, the approaches of sentence selection mainly adopt the strategy of adding sentences to summary step by step until getting the appointed summary length, according to the sentence’s weight and its redundancy to the already selected sentences. This approach has three problems: (1) Sentence selection is based on the selected sentences at some extent, so the summary may be far away from the main topic of the document set with the improper first sentence. (2) The length of the summary is restricted, so depending deeply on the sentence weight, which results in the sentences with a little lower weight having no chance to be added into summary. (3) The process of adding sentences to the summary step by step leads to the sentence selection being a choice at the current state of already selected sentences but not the whole summary’s sentences.

In this letter, we propose a two-stage sentence selection approach to deal with the above problems. At the first stage, we obtain the candidate summary sentence set with n times as long as the appointed summary length by the traditional redundancy-based sentence selection method. At the
second stage, we delete the sentence with the least contribution to the candidate summary sentence set until getting the appointed summary length. Compared with the traditional method of adding sentences to get summary, our approach is to delete sentences in a set of candidate sentences. There are two advantages. One is to increase the probability of sentences selected to the summary with \( n \) times as long as the appointed length. The other is to delete the sentences according to their contribution to the whole candidate sentence set but not the already selected sentences. With the test corpus used by DUC 2004, experiment results show its validity.

The structure of the letter is as follows. Section II describes the traditional redundancy-based sentence selection method. Section III presents the two-stage sentence selection approach. Section IV gives sentences ranking adopted by this letter. Section V shows the experiments and analyzes the results. Section VI gives the conclusions.

II. Redundancy-based Sentence Selection Approach

The traditional sentence selection method is to add sentences with high weight one by one to get the appointed length of summary. In multiple documents set, the important information is repeated more than once, and a lot of the same or similar sentences appeared in the documents. So sentence selection should not only consider the weight of the sentence but also measure its redundancy with the already selected sentences in order to contain more information in the restricted summary length.

The redundancy-based sentence selection approach is one of the traditional sentence selection approaches. After ranking sentences according to their weights, it selects the highest ranked sentence to the summary at the first step. As long as the summary is under the appointed summary length, the next highest ranked sentence is added to the summary if its redundancy with the already selected sentences is lower than the threshold \( \alpha \). The redundancy is measured as follows:

\[
\text{Redundancy}(\text{Sent}_i, S) = \frac{|\text{Token}(\text{Sent}_i) \cap \text{Token}(S)|}{\text{Token}(\text{Sent}_i)}
\]

(1)

where \( \text{Token}(\text{Sent}_i) \) and \( \text{Token}(S) \) are the set of tokens included by \( \text{Sent}_i \) and set \( S \) respectively. The topic word is taken as the tokens, which have been acquired in Section IV. If the value of \( \text{Redundancy}(\text{Sent}_i, S) \) is high, then it indicates that the sentence contains too many common tokens with the already selected sentences, and it is meaningless to add this sentence to the summary.

III. Two-stage Sentence Selection Approach

In order to let more sentences have the chance to be added to the summary and consider a sentence’s contribution to the summary from the global view, a two-stage sentence selection approach is proposed. This approach has two stages: The acquisition of a candidate sentence set \( S \) and the optimum selection of sentences. At the first stage, we use the redundancy-based sentence selection approach to obtain a candidate sentence set \( S \), which is \( n \) times as long as the appointed length of the summary, in order to contain more important sentences. At the second stage, we implement the optimum sentence selection by deleting the sentences in the set \( S \) with the least contribution to the set \( S \) in each circulation until getting the restricted length of the summary. The character of the optimum sentence selection is that it is a process of deleting the sentences in the set \( S \), rather than adding sentences to the summary traditionally. And it measures the sentence’s contribution through the whole set \( S \), but not the only already selected sentences at that time, so it changes the view angle from the local to the global.

The sentence’s contribution function used in the optimum sentence selection is aimed to delete the sentence with the least loss of information for the whole candidate sentence set \( S \). So we define the following equation to find the sentence with the least new tokens to be added to the set \( S \).

\[
\text{Sent}_s = \underset{\text{Sent}_i \in S}{\arg \min \{|\text{Token}(\text{Sent}_i) - \text{Token}(\text{Sent}_i) \cap \text{Token}(S \setminus \text{Sent}_i)|\}}
\]

(2)

where \( \text{Token}(\text{Sent}_i) \) and \( \text{Token}(S \setminus \text{Sent}_i) \) are the set of tokens included by \( \text{Sent}_i \) and the set \( S \) eliminating \( \text{Sent}_i \), respectively. We choose word and topic word as the token in the experiments.