A Chinese Automatic Text Summarization system for mobile devices

Lei Yu, Mengge Liu, Fuji Ren, Shingo Kuroiwa

Faculty of Engineering, The University of Tokushima
2-1 Minamijosanjima, Tokushima 770-8506, Japan
School of Information Engineering, Beijing University
of Posts and Telecommunications, Beijing 100876, China
{yulei, liumengge, ren, kuroiwa}@is.tokushima-u.ac.jp

Abstract. A large amount of on-line information and lengthiness information can't fit for the mobile devices. In order to save this problem, we propose a method which collects original news text from on-line information and extracts summary sentences from them automatically. On this basis, we adopt WML(Wireless Markup Language) to build a news website for mobile devices browsing through the news summary. The system is mainly made up by Automatic News Collection and Auto Text Summarization. Our experimental results proved the effectiveness of the means.

Keyword: World Wide Web; Automatic Text Summarization; Automatic News Collection

1 Introduction

As the news websites grow rapidly, the on-line news becomes large and repeatedly, even some news websites quote the contents of other news websites[1]. It is more and more difficult for the reader to skim over websites and get a quick idea of their content. Therefore, based on reporting intensity and updating speed, we select several Chinese news Home sites as the resource of news collection.

In the last few decades, researchers have brought up a number of automatic summarization methods[2][3][4][5][6]. Basically these methods fall into two categories. One is called human-imitation approach or generative approach, which constructs the summary based on the structure and semantics of the source text. The other group of approaches tries to extract important sentences based on information acquired from the surface clues of the text. This approach is termed the extraction approach. Therefore, at the current stage, it is better to use statistical information in automatic extraction. The method proposed in this paper also extracts important sentences based on textual statistical information. But our method not only uses statistical information but also employs text structural features to improve the performance of important sentences extraction.

2 Basic idea of Automatic News Collection

The basic approach of Automatic News Collection is to analyze target websites and find out the rule of the articles text in HTML page. We have built the experiment news collection system named ACnews to automatically collect news on the proposed method. In this section, we describe the main process for extracting news content from a website.

(1) Download a news HTML page from destination websites.
(2) Analyze the HTML page to find out the rule.
(3) Collect news HTML pages from destination websites.
(4) Extract information about news author and news date.
(5) Extract news text from HTML pages.
(6) Delete ad and ad link from original news text.

(7) Save the information obtained by fourth step and fifth step in database.

At present, ACnews is only for Chinese articles collection. Of course our method can be applied to other language articles easily, just by additional support of the multinational language code.

3 Basic idea of Automatic Text Summarization

Our basic idea is to extract sentences based on the statistical information and structural features. H. P. Edmundson[7][8] has proposed four methods to decide importance of sentences in his article. On this basis, according to observation to the Chinese summaries, we have confirmed that the most important sentences are extracted according to the occurrence frequency of words, the length of the sentences, and the position of the sentences. The method uses the following two assumptions:

- The assumption made by Luhn that sentences closely related to the topic of the text occur frequently in the same text.
- The assumption that generally the first sentence and the longer sentences in articles summarize the main content of the articles.

Therefore in our method, while judging the importance of sentences, we only use noun phrases or noun clauses as the clues to determine candidate sentences to be extracted. At the same time, we judge the importance of words based on classical approach *TF • IDF*. The second assumption says that important sentence often appears at the beginning of paragraphs and sentences with strong generality are longer than the others. Therefore, our calculation of importance of a sentence also depends on these two components.

3.1 Computing word importance

We use the following formula to compute a word importance.

$$W_i = TF \times IDF = n \times \log(\frac{M}{m}) \tag{1}$$

Where

- *Wi* : the importance of *i*th word,
- *n* : the number of occurrence of the *i*th word in *i*th *sentence,*.
- *M* : the total number of sentences in the text,
- *m* : the number of the other sentences contained *i*th word.

3.2 Computing a sentence importance

According to the first assumption and the second assumption we have confirmed the following formula to compute the weight of sentences:

$$l_i = \alpha \frac{\sum_{w \in S_i} tf(w) \bullet idf(w)}{\max_{S_j \in C} \left\{ \sum_{w \in S_j} tf(w) \bullet idf(w) \right\}} + \beta \frac{Length_{S_i}}{\max_{S_j \in C} \left\{ Length_{S_j} \right\}} + \gamma \, position_{S_i} \tag{2}$$

Where

- l_i : the importance of *i*th sentence,
- C : the set of all sentences in the article,
- w : the word in the articles,
- $Lenghth_{si}$: the length of ith sentence,
- $Position_{si}$: value given according to sentence position,
- $\alpha+\beta+\gamma=1$: In the test parameter that we used is 0.3, 0.3, 0.4.

4 Experiment and evaluation

Basically, there are two major two types of summarization evaluation methods: intrinsic evaluation and extrinsic evaluation[9][10]. Intrinsic evaluation compares automatically generated summaries with gold standard (ideal summaries) to measure the quality of these summaries straightforwardly. Extrinsic evaluation measures the performance of automatically generated summaries in a particular task (e.g., classification).

In our test, an intrinsic evaluation method is presented. The following three systems are employed to do the same experiment.

- **ATSS-SI:** A system we built using both the statistical information and the structural information.
- **ATSS-SS:** A system we built using only statistical information.

Evaluation experiments are discussed as follows. We use ACnews to collect 30 Chinese news articles that are various genres randomly. Then we use these 30 Chinese news texts to construct the testing corpus. For each text, three students constructed manually and independently “ideal” summaries in two proportions: 10% and 20%, which are the rates of the summary length to the original text length. Then the summaries generated by using our summary system are compared with the ideal summaries extracted by human. For each text, the precision and recall are computed to evaluate the quality of the summary. They are defined as follows:

$$precision = \frac{|S_t \cap S_m|}{|S_m|} \qquad (3)$$

$$recall = \frac{|S_m \cap S_c|}{|S_c|} \qquad (4)$$

Where

- Sm : the set of summary sentences produced by the system,
- St : the union set of the 3 sets of summary sentences manually extracted by 3 graduate students, $St = S1 \cup S2 \cup S3$,
- Sc : the intersection set of that, $Sc = S1 \cap S2 \cap S3$.

We compare the summaries generated by ATSS-SI with the summaries generated by ATSS-SS. The results are shown in Table 1:

Table 1. The result of summary evaluation

Summary Rate		ATSS-SI	ATSS-SS
10%	*precision*	0.77	0.55
	recall	0.78	0.58
20%	*precision*	0.74	0.51
	recall	0.76	0.53

Table 1 shows the results of the presented system ATSS-SI is better than ATSS-SS. The result clearly demonstrates that the approaches used in our system are reasonable.

5 Conclusion

In this paper, we described an approach which uses both Automatic News Collection and Automatic Text Summarization. It is a feasible application for the users who often use mobile devices such as PDA and mobile telephone to get a quick idea about on-line news. Compared with web crawlers, our Automatic News Collection system (AC-news) can extract original news text quickly and accurately. And in the Automatic Text Summarization, we proposed a method which extract important sentence based on statistical information and the structural information of the text. The evaluation result stated shows the method is reasonable. However the summary generated based on information acquired from the surface clues of the text is mechanical. Constructing the summary based on the structure and semantics of the source text has become more and more popular recently. We are working on a better method to make generated summary better understood by humans.

Reference

1. Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30 (1998) 107–117
2. Alam, H., Kumar, A., Nakamura, M., Rahman, F., Tarnikova, Y., Wilcox, C.: Structured and unstructured document summarization: Design of a commercial summarizer using lexical chains. In: 7th International Conference on Document Analysis and Recognition. Volume 2., Edinburgh, Scotland, UK (2003) 1147–1150
3. Salton, G., Singhai, A., Mitra, M., Buckly, C.: Automatic text structuring and summarization [A] . In advances in Automatic Text Summarization [C], Eds. I.Mani and M. T.Maybury. The MIT Press. (1999) 62 - 70
4. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer [M]. Addison2Wesley, (1989)
5. H. P. Edmundson, R. E. Wyllys.: Automatic Abstracting and Indexing—Surveyand Recommendations. Communications of the ACM., 4(5) (1961) 226-234
6. F. J. Ren.: AUTOMATIC ABSTRACTING IMPORTANT SENTENCES. International Journal of Information Technology & Decision Making Vol. 4, No. 1 (2005) 141-152
7. H. P. Luhn.: The automatic creation of literature abstrats, IBM Journal of Research and Development 2-2 (1958) 159-165.
8. H. P. Edmundson.: New Methods in Automatic Extraction[A]. In Advances in Automatic Text Summarization[C]. (1998) 23-42
9. S. Jones and J. Galliers.: Evaluating Natural Language Processing Systems: an Analysis and Review. Springer, New York, (1996)
10. I. Mani. M. Maybury.: Advances in Automatic Text Summarization, MIT Press. ISBN 0-262-13359-8 (1999)