# Chinese multi-document summarization Based on Topic Detection technology

LIU Mei-ling
Department of Computer Science and Technology, Harbin Institute of Technology
Department of Computer Science and application, NEFU
Harbin, China
mlliu@mtalab.hit.edu.cn

ZHAO Tie-jun
Department of Computer Science and Technology, Harbin Institute of Technology
Mi&tlab hit
Harbin, China
tjzhao@mtlab.hit.edu.cn

***Abstract*—This paper reports an initial study that aims to Chinese multi-document summarization. We introduce and compare different dynamic threshold model which TDT [1] (Topic Detection and Tracking), and get document sets based on topic, then focus on Chinese multi-document summarization generation. Our approaches are based on the combination TDT temporal properties and multi-document summarization. Results show that using different dynamic threshold in TDT influence output summary representation. For our future work, we will continue to study the temporal multi-document summarization base on the web.**

***Index Terms*—Topic Detection; temporal properties; dynamic threshold; multi-document summarization**

## HEADING 1 INTRODUCTION

As a new research direction of natural language processing, topic detection and tracking (Topic Detection and Tracking, TDT) is a research of information organization and utilization aimed at an event-based [2], also to cope with information overload problem in applied research. It is such technology which refers to the newswire and broadcast news sources, such as find the subject and topic relevant content in automatic news data flow information. Topic detection is essentially similar to unsupervised clustering study, but the clustering is usually based on the achievement of global information, and topic detection clustering achievement as increment mode.

Multi-Document Summarization [8] is aim at the information to form collection documents about the same subject documents. Summarization should be objective and truly reflect the contents of the original topic of the corpus, but concise than the original text. Summarization can make it possible to quickly determine whether the original content is interested, you can allow people to quickly find the document you really need, instead of wasting time reading relevant document, greatly raise the efficiency of accessing to information.

Based on the topic detection technology the Chinese multi-document summarization [9], aimed at a large number of data streams in the report. Through the topic of pre-detection processing, formation a collection document based on the theme in order to provide the same kind of multi-document summarization for the purpose of the subject content. Generate concise and exact description for main contents of the original document. Through improving the topic detection algorithm, compared generated multi-document summarization for the different dynamic threshold.

## HEADING 2 RELATED WORK

My research is the Chinese multi-document summarization based on the topic detection, the methods and algorithms used are as follows.

*1)* Topic Detection Technology

Topic detection is such an automatic method which is a part of a broader initiative called Topic Detection and Tracking (TDT), and is defined to be the task of automatically detecting new topics in the news stream and associating incoming stories with topics created so far. Topic detection is essentially similar to the unsupervised clustering except that it is done incrementally, not globally.

In the study of Topic Detection, usually pre-set threshold [3], and once set in the entire testing process will not change, this approach is unreasonable. First of all [4], as a result of the time focused on the subject, with the process of time, and the topic-related reports will be less and less, so the detection process should be gradual increase threshold; Secondly, different threshold should be taken to a different topic.

*2)* Multi-document summarization technology

This article focused on the original text-based multi-document abstracts, this type of system has the general process [10]:

*a)* According to the various forms characteristics to weight sentences;

*b)* According to the right of sentence to re-select the sentence and to remove redundant information;

*c)* Sort of the sentence to generate a coherent abstracts. Integrated sentence weighted, select technical, sort technical to realization of a Chinese [11] multi-document automatic abstracting system.

*3)* Topics detection technology based on the time information

Time information in Natural Language Processing (Natural Language Processing, NLP) field have a very important role [5], it is the base in many natural language processing tasks, such as multi-document abstracts system need to sort in chronological order in accordance with the relevant information. News reports are rich in time information, therefore, the time information in topic detection and tracking research should play a very important role. Based on the above analysis, we

978-1-4244-4607-0/09/$25.00 ©2009 IEEE

applied a dynamic threshold model based on the topic duration [6].

As follows(1):

$$Threshold(T,t) = \theta + \alpha * (Time(S) - Time(T)) \quad (1)$$

In this, *Threshold(T,t)*express threshold $t$ moment in which $T$ topic; $\theta$ is a constant, express the threshold that the topic had just been established at the time; $\alpha \propto$ is an adjustable parameter, express time information in a dynamic threshold proportion. *Time(S)* And *Time(T)* , respectively express reported time and the topic set-up time, the subject of set-up time refers to the topic of the time the seed is reported. For Experiment, Report time and the establishment time of the topic in days, based on our evaluation corpus used in the marking of time.

*4)* Ratio method

In the basic model of Topic Detection, the first calculation of all the reports and the similarity of the subject already exists, then the comparison the highest similarity with the similarity threshold to be reported and getting relevant to the subject to determine the outcome. However, in the model based on dynamic threshold detection algorithm topics, each topic is the similarity threshold may not be the same. To make reports and the similarity of different comparable topic, we apply ratio method [15] to select the most similar to those reported with a topic. Based on the ratio method the base idea for choosing the most similar to the subject are as follows [7]:

*a)* Calculation the similarity of report and topic;

*b)* Calculating the ratio of similarity with the threshold of the topic, the largest ratio recorded as *MaxRatio*, the biggest ratio topic recorded as $T\max$, if *MaxRatio* greater than or equal to 1, then the reports and topics is related;

*c)* If all *MaxRatio* is less than 1, then to establishment a new topic.

*5)* Multi-document summarization baseline system

The primary function of the system [12] is to first focus on topics related document text clause, cent-word pre-processing, the application of stop words table filter out stop words and word frequency statistics; on this basis to proceed with the theme term extraction, and integration of lexical information, location information, length information to calculate the weights of the sentence; based on the value to choice the right sentence which contain large value and new information then adding the sentence to the summary until it reaches the specified length; last through the certain sorting algorithm to sort of sentence in order to generate a certain extent a logical, coherent flow of summarization.

## HEADING 3 THE EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, first of all, use the time information dynamic threshold models and the ratio method in the topic detection baseline. Then use the output of time information dynamic threshold models as the multi-documents summarization input. Finally, get the summarization for the topic detection corpus.

Experiment has analyzed the time characteristics of the topic, and applied topic duration of the dynamic threshold models and the ratio method to select the most similar to the topic report. Because of the time Threshold have a significant impact on the output of the TDT, so comparing multi-documents summarization [13] for the different time thresholds.

*1)* Experiment on corpus of Chinese topic detection

In this paper, the Chinese test corpus is news report obtained from the Internet and marked by hands, a total of 113. All the reports of Corpus in chronological order compose a text file, which marked the 12 topics.

*2)* Topic detection Evaluation Criteria

In this paper, used the normalized detection spending *CDet* to evaluate topic detection system in the TDT evaluation [14], the formula is as follows:

$$(C_{Det})_{norm} = \frac{C_{miss} \cdot P_{miss} \cdot P_{t\arg et} + C_{fa} \cdot P_{fa} \cdot P_{\neg t\arg et}}{\min(C_{miss} \cdot P_{t\arg et}, C_{fa} \cdot P_{\neg t\arg et})} \quad (2)$$

Which, $Pmiss$ is the fail to report rate for the system; $P_{t\arg et}$ is the misinformation for the system; $P_{t\arg et}$ is a new topic probability in information flow and P-target is an old topic probability to see in information flow, P-target =1 - $P_{t\arg et}$; $CMiss$ is the cost of omitting a new incident; CFA is the cost of misinformation once.

Based on the above Evaluation Criteria, has been the basic model evaluation results for Chinese topic detection in the Chinese corpus , results shown in TABEL I, the best results from the table in the topic detection system when the threshold value of 0.40 to obtain.

TABEL I. Evaluation of Chinese detection results

| threshold | PMiss | PFA | (CDet)Norm |
|---|---|---|---|
| 0.25 | 0.4816 | 0.0541 | 0.7467 |
| 0.30 | 0.1966 | 0.0693 | 0.5361 |
| 0.35 | 0.1195 | 0.0443 | 0.3367 |
| 0.40 | 0.1404 | 0.0290 | 0.2823 |
| 0.45 | 0.1962 | 0.0202 | 0.2953 |
| 0.50 | 0.2070 | 0.0170 | 0.2902 |
| 0.55 | 0.3039 | 0.0041 | 0.3238 |
| 0.60 | 0.4304 | 0.0032 | 0.4462 |

Next, using dynamic threshold model based on the time information namely $\theta + \alpha$ , $\alpha$ = 0.4, rather than the single static threshold $\alpha$ . Take different $\alpha$ values to validate the performance [15], the results shown in TABEL II.

TABEL II. the results of time information model of dynamic threshold

| $\alpha$ | PMiss | PFA | (CDet)Norm |
|---|---|---|---|
| 0.001 | 0.2893 | 0.0021 | 0.2997 |
| 0.002 | 0.2299 | 0.0017 | 0.2381 |
| 0.003 | 0.2469 | 0.0014 | 0.2539 |
| 0.004 | 0.2451 | 0.0011 | 0.2505 |
| 0.005 | 0.3056 | 0.0009 | 0.3101 |

Experiment results show that the topic dynamic threshold model based on duration with the use of the ratio achieved good results.

As we can see that the dynamic threshold model based on time information achieved very good results. However, with the value increased, the topic threshold will become increasingly large, so that many reports can not be detected. Therefore, this paper should take proper values.

*3)* Criteria Multi-document summarization based on Topic detection

Experiment on the value equal to 0.001, 0.002, 0.003, 0.004, 0.005done a comparative, results shown in TABLE IV. After the expert evaluation, visual comparison can be seen that the proportion of the parameter changes in the time information in the dynamic threshold, the generated multi-document summarization is difference in the quality.

TABEL IV.Different value VS summarization length

| $\alpha$ | length |
|---|---|
| 0.001 | 307 |
| 0.002 | 322 |
| 0.003 | 299 |
| 0.004 | 87 |
| 0.005 | 87 |

As a result of different language evaluation corpus, this experiment not to compare the same results at home and abroad about the same study. However, the results contrast with their own more or less can be seen that this method of generating multi-document summarization achieved some satisfactory results.

The experimental results show that in the dynamic threshold model based on the topic duration, different thresholds have different results. Such as value equal to 0.001 and 0.003 have done a comparative experiment, through the visual comparison, we considered that threshold equal to 0.0003 have the best digest quality. When the time is too small or contain too much information, the digest result can only be the general.

With the increased value, the threshold of the subject will become increasingly large and many reports can not be detected. Also the information that summarization contained increased, with less fluency in reading and more information coverage of major. When the value increases to a certain extent, the summarization is not readable, and the quality of documents is not good. So corpus should pay propriety attention to the information in time, that multi-document summarization can cover with better comprehensive integration information.

By comparing the generated multi-document summarization, we can integrate topic detection technology and multi-document summarization technology. We can balanced the TDT output through adjust the value, so that generate better quality multi-document summarization.

In this paper, multi-document summarization is as to an important part of the topic detection and output incident report. Multi-document summarization also plays an important role on the customization and information integration and so on. Chinese multi-document summarization based on the topic Detection technology, combined with the timing and dynamic of the TDT. Because of the network contain enormous information, so we can format the document collection for the same theme, according to the topic of clustering in the massive data streams. Then generate multi-document summarization for the subject of inquiry in accordance with the relevant user, to get the latest summary.

## HEADING 4 CONCLUSION AND FUTURE WORKS

Multi-document summarization is related to other information processing technologies such as information retrieval, information extraction, topic detection and tracking, as well as single-document summarization and so on. Therefore, multi-document summarization will deeply have a huge impact and boost the natural language processing technology. In particular, Chinese multi-document summarization is developed speediness yet. So summarization technology obtains a great advancement on large-scale data processing and feasibility.

This paper introduces the topic detection system and multi-document summarization base model system. Test is to consider to combining TDT and multi-document summarization technology. Time is an important feature of the subject, and it not only more apply on the topic detection, but also generate good result for multi-document summarization technology. The next step will be continue to identify and track the data which up-to-date news reports, verify the correlation between the follow-up data and the topic has been in existence. Update the topic contents of the report, and then update the multi-document summarization system.

## HEADING 5 ACKNOWLEDGMENT

This paper is supported by National 863 project emphases item (No. 2006AA010108), and Natural Science Foundation of China (No. 60736044).

## REFERENCES

[1] James Allan, Jaime Carbonell. Topic Detection and Tracking Pilot Study: Final Report. In : Proceeding of the DARPA Broadcast News Transcriptions and Understanding Workshop, February, 1998: 194-218.

[2] James Allan, Victor Lavrenko. UMass at TDT 2000. Available at http://www.nist.gov/speech/tests/tdt/tdt2000/papers.htm, 2000

[3] Walls, F., H. Jin,S. Sista, and R. Schwartz. Topic Detection in Broadcast News. In: Proceedings of the DARPA Broadcast News Workshop, Herndon, 1999: 193-198.

[4] Juha Makkonen, Helena Ahonen-Myka, Marko Salmenkivi. Applying Semantic Classes in Event Detection and Tracking. In: Proceedings of

[5] International Conference on Natural Language Processing, Mumbai, India, 2002: 175-183.

[6] Stephanie Strassel, David Graff, Nii Martey. Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora. In: Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

[7] The 2003 Topic Detection and Tracking (TDT2003) Task Definition and Evaluation Plan. Available at http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm, April, 2003

[8] J.M. Conroy, J.D. Schlesinger. CLASSY 2007 at DUC 2007. In Proceedings of the 2007 Document Understanding Conference(DUC 2007), New York, 2007

[9] S. B. Goldensohn, D. Evans, et al. Columbia University at DUC 2004. In Proceedings of the 2004 Document Understanding Conference(DUC 2004). Boston, MA, 2004

[10] D. Bollegala, N. Okazaki, M. Ishizuka. A Machine Learning Approach to Sentence Ordering for Multidocument Summarization. In Proceedings of the Annual Meeting of the Association for Natural Language Processing, Japan, 2005: 482~488

[11] N. Daniel, D. Radev. Timothy Allison. Sub-event Based Multi-Document Summarization. HLT NAACL Workshop on Text Summarization, Edmonton, Canada, 2003: 9~16

[12] G. C. Stein, T. Strzakowski, G. B. Wise. Summarizing Multiple Documents Using Text Extraction and Interactive Clustering. In

Proceedings of the Pacific Rim Conference on Computational Linguistics, Canada, 1999: 200~208

[13] Y. H.Gong, X. Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In Proceedings of ACM SIGIR'01, 2001: 19~25

[14] I. Mani. Recent Developments in Temporal Information Extraction. Proceedings of the Conference on Recent Advances In Natural Language Processing, 2004

[15] The 2003 Topic Detection and Tracking Task Definition and Evaluation Plan. http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm, April, 2003 03/evalplan.htm, April, 2003