Obtaining Uncertainty to Generate Summarization

Jinguang Chen[1,2] and Tingting He[1,2]

[1]*Department of Computer Science and Technology, Huazhong Normal University,430079 Wuhan, China*

[2]*Engineering & Research Center for Information Technology on Education, Huazhong Normal University, 430079 Wuhan, China*

cjg2003@hutc.zj.cn, tthe@ccnu.edu.cn

Abstract

This paper describes Huazhong Normal University's participation in TAC 2010. For the guided summarization task, we use a better basic summarization system which makes many improvements to the method we used in TAC 2009. Our system is based on uncertainty methods, including cloud. Our teams IDs are 6 and 23, and they are among the best of all the 43 automatic summarization systems in TAC 2010.

1. Introduction

The Guided Summarization task aims to encourage summarization systems to make a deeper linguistic (semantic) analysis of the source documents. The guided summarization task is to write a 100-word summary of a set of 10 newswire articles for a given topic, where the topic falls into a predefined category. There are five topic categories: accidents and natural disasters, attacks, health and safety, endangered resources, investigations and trials. Participants are given a list of important aspects for each category, and a summary must cover all these aspects (if the information can be found in the documents).

Many problems of natural language processing (NLP) contain uncertainties. To deal with these problems, we use cloud model[1] and data field[1] method to quantitatively represent uncertainties to improve the effectiveness of summarization. Our work focuses on improving the effectiveness of query-focused multi-document summarization by handling uncertainties of words and sentences.

The rest of the paper is organized as follows. Section 2 introduces the basic system, Section 3 and 4 introduces our proposed method, section 5 presents experiments and evaluation results, and section 6 concludes with discussions and future research directions.

2. Basic System

The basic system selects sentences by using a feature fusion method to identify the sentences with high query relevance and high information density, i.e., the more relevant a sentence is to the query, and the more important a sentence is in the document set, the more likely the sentence is to be included in the final summary.

First, we score every sentence's representative feature (query-independent feature, QI henceforth) by obtaining its importance in document set. We use a vector space model(2) to obtain similarity between two sentences in document set. For a fixed collection of sentences, an *m*-dimensional vector is generated for each sentence, where *m* is the number of unique terms in the document set. In our method, weights associated with words are calculated based on TF-ISF(3).

Secondly, we re-score every sentence's query-focused feature (query-focused feature, QF henceforth) by obtaining its similarity with query. It has been shown that semantic representation of words in higher dimensions using Hyperspace Analogue to Language (HAL henceforth) spaces(4) can be extended to calculate the relevance score of a sentence towards the information need (5). HAL is used to obtain relevance score between words in document set and the query words. HAL is also known as "slip window co-occurrence", which uses co-occurrence of two words in a window of given length to obtain relevant information between them. The more times two words co-occur, and the nearer they are in the distance, the more relevant they are.

The final score of sentence S is obtained by linear combination:

$$\text{Score(S)}=\sigma\bullet\frac{QF(S)}{\sum_{i=1}^{N}QF(S_i)}+(1\text{-}\sigma)\bullet\frac{QI(S)}{\sum_{i=1}^{N}QI(S_i)}$$

where σ is the parameter to adjust the proportion of the two parts.

3. Cloud Model method

Fig. 1 illustrates the Cloud Model method. We can see that after sentences in document set and the query are preprocessed, their query-independent feature and query-focused feature are evaluated by the query-independent cloud (QIC) and the query-focused cloud (QFC), respectively. QFC consists of two subsequent models, i.e., word-level query-focused cloud (WQFC) and sentence-level query-focused cloud (SQFC). QIC and QFC are integrated together by feature fusion to decide the final score of sentences. We call our method cloud

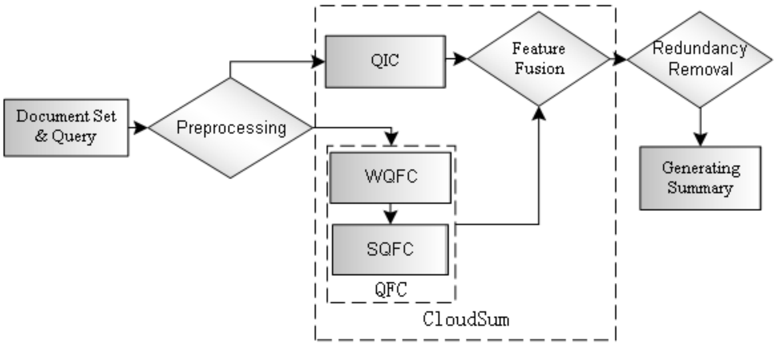

Fig. 1. Structure of the Cloud Model Method.

summarization method (CloudSum) and it consists of three cloud models: QIC, WQFC and SQFC. Details about CloudSum could be found in a published paper in reference 6.

4. Data Field method

Fig. 2 illustrates how the data field constructed. We can see that after sentences in document set were scored by the basic system as described in section 2, they are further processed by the Data Field method. The Data Field method firstly selects the best n percent sentences from all sentences according to scores obtained by the basic system. Then a data field matrix is constructed considering two aspects: a. importance of sentences; b. "votes" or "recommendations" between each other. In the Data Field method, a sentence with more importance as well as nearer distance to another sentence S will add greater weight to the potential of sentence S. A parameter optimizing process is then utilized to find the best data

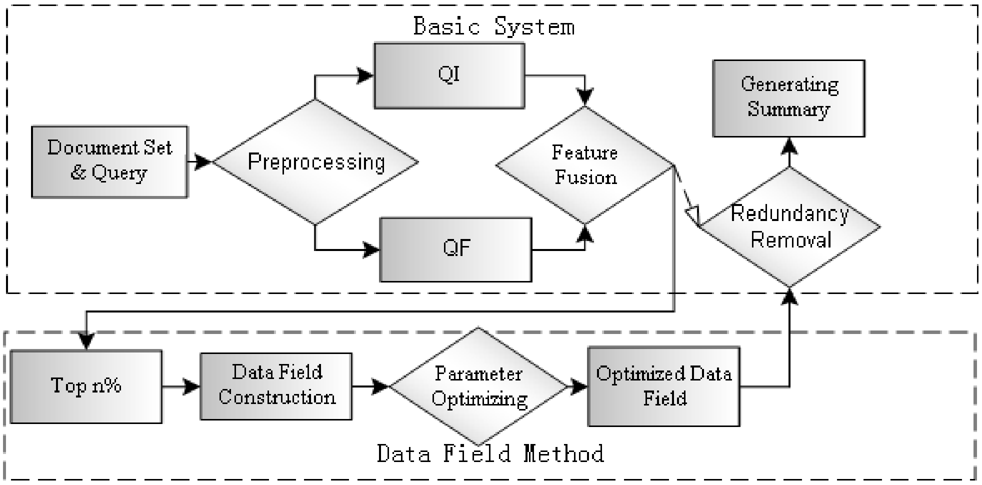

Fig. 2. Structure of the Data Field method.

field which is most appropriate for describing the character of relationship between the selected sentences. The optimized data field is used to choose the best sentences to generate the summary. We call this system DFSum. Details about DFSum could be found in a forthcoming paper titled *Query-focused Multi-document Summarization Using Data Field.*

5. Experiment

5.1 Experimental Set-up

Preprocessing of data sets has three main steps: sentence extracting, removal of stop-words and word-stemming. We use a perl module *breaksent-multi.pl* to extract sentences from documents. Stop-words in both documents and queries are removed, and the remaining words are stemmed by *Morphology* developed by Stanford University, which computes the base form of English words based on a finite-state transducer, and are considered as terms. All terms are converted to lowercase, terms containing punctuation are removed with the exception of terms containing “-”, which are split in two, and terms with length no more than 2 are also removed. For the benefit of repeatability, no further treatment is taken.

After the sentences are scored, we use an evolved Maximal Marginal Relevance (MMR)(7) method to avoid including redundant information in the summary. The basic idea of our redundancy removal process is: once a sentence is selected into the summary, its influences on the remaining sentences will be removed. The iteration of selecting abstract sentences ensures that redundant information between the current candidate sentence and all of the former selected sentences is not included.

5.2 Experiment Results

We use CloudSum and DFSum to participate TAC2010’s guided summarization task[a]. In order to make CloudSum and DFSum, which are designed for the query-focused summarization task, adapt to the guided summarization task, we make a change in replacing the query with “categories and their aspects” information provided with the task. Table 1 and 2 show ROUGE-2わROUGE-SU4わBEわPyramidわManual evaluation result for task A and B, respectively. CloudSum and DFSum’s team ID is 23 and 6, respectively, which ranked among the best of all the 43 machine systems.

[a] http://www.nist.gov/tac/2010/Summarization/index.html

System	Peer ID	ROUGE-2	ROUGE-SU4	BE	average modified (pyramid) score	average overall responsiveness
CloudSum	23	3	2	2	8	3

Table. 1. Rank of the CloudSum in TAC 2010 guided summarization track, task A.

System	Peer ID	ROUGE-2	ROUGE-SU4	BE	average modified (pyramid) score	average overall responsiveness
DFSum	6	5	6	6	3	6

Table. 2. Rank of the DFSum in TAC 2010 guided summarization track, task B.

6. Conclusion

This paper briefly described two method derived from uncertain theory to generate query-focused multi-document summarization. Experimental results on the TAC2010 data sets demonstrated the effectiveness of our method. In light of the inherent limitations of the mainstream uncertainty-handling theory, and considering the common existence of random distribution and often-ignored fuzziness, we believe our approach can be utilized widely in many other areas of NLP in addition to the summarization area as presented in this work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 60773167), the Major Research Plan of National Natural Science Foundation of China↖ No. 90920005), the 973 National Basic Research Program (No. 2007CB310804), the Program of Introducing Talents of Discipline to Universities (No. B07042), and the Natural Science Foundation of Hubei Province (No. 2009CDB145), Chenguang Program of Wuhan Municipality (No. 201050231067).

Reference

[1] D.Y. Li and Y. Du. *Artificial Intelligence with Uncertainty*, Chapman & Hall/CRC2007.

[2] G. Salton, A. Wong, and C.S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18:11 (1975), 613–620.

[3] J.L. Neto, A.D. Santos, C.A.A. Kaestner, and A.A. Freitas. Document clustering and text summarization. Proceedings of PADD 2000, 41–55.

[4] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical

co-occurrence. *Behavior Research Methods Instrumentation, and Computers*, 28 (1996), 203-208.

[5] J. Jagarlamudi, P. Pingali, and V. Varma. A Relevance-Based Language Modeling Approach to DUC 2005. Proceedings of Document Understanding Conference (2005).

[6] Jinguang Chen, Tingting He. Query-focused Multi-document Summarization Using Cloud Model, Information-An International Interdisciplinary Journal. 2011.1.

[7] J. G. Carbonell and J. Goldstein. The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. Proceedings of SIGIR 1998, 335–336.