ICT CAS at DUC 2007

Jin Zhang, Hongbo Xu, Xiaolei Wang, Huawei Shen, Yiling Zeng

Institute of Computing Technology

Chinese Academy of Sciences

{zhangjin, hbxu, wangxiaolei, shenhuawei, zengyiling}@software.ict.ac.cn

Abstract

This paper presents our multi-document summarization system ICTGSP-S at DUC 2007. We propose a new method for representing and summarizing documents by integrating subtopics partition with graph representation. The method starts from the assumption that capturing sub-topic structure of document collection is essential for summarization. The evaluation results show the benefit of this approach.

1 Introduction

DUC 2007 consists of two independent tasks. The main task is the same as the DUC 2006 task and will model real-world complex question answering, in which a question cannot be answered by simply stating a name, date, quantity, etc. Given a topic and a set of 25 relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic statement. Successful performance on the task will benefit from a combination of IR and NLP capabilities, including passage retrieval, compression, and generation of fluent text. The update task is to produce short (~100 words) multi-document update summaries of newswire articles under the assumption that the user has already read a set of earlier articles. The purpose of each update summary will be to inform the reader of new information about a particular topic. The documents for summarization come from the AQUAINT corpus, comprising newswire articles from the Associated Press and New York Times (1998-2000) and Xinhua News Agency (1996-2000) [1].

This is the first time that our group attended to the DUC evaluation. ICTGSP-S is a summarizer we developed during our participation to the DUC 2007 main task and update task. Our approach is based on a new method proposed by ourselves – GSP-S – based on the topic's implicit organization. The topic of document cluster consists of different subtopics. Therefore, the summarization task can be converted into the process of n-best subtopics finding, and the key for our task here is to find the best subtopics and to select a salient sentences to stand for certain subtopic.

The rest of this paper is organized as follows. Section 2 relates an overview of our system in detail. The evaluation results from NIST and experiments are reported in section 3, followed by the conclusion in section 4.

2 Our System for DUC 2007

Summarization is a product of electronic document explosion, and can be seen as the condensation of the document collection. As summary is concise, accurate and explicit, it became more and more important.

As the current multi-document summarization (MDS) systems are not convenient to do extensive experiments. Therefore, we decided to design and implement a new extractive MDS system, which is based on the sub-topic representation for document collection. Our new system named GSP-S (Graph-based Sub-topic Partition - Summary) takes sub-topics as the basic process units for summarizing.

As same to the conventional MDS systems, our system is also divided into three modules: text preprocessing, summarization algorithm, and post-processing in Figure 1.

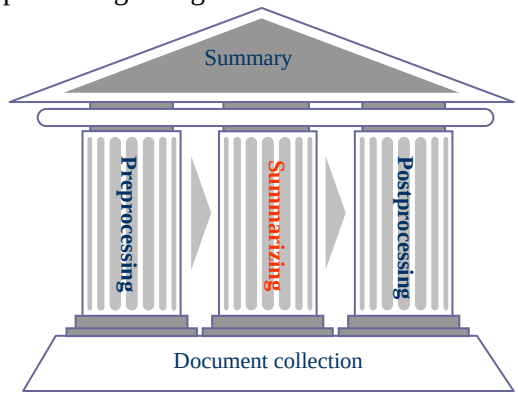

Figure 1: MDS System Architecture

2.1 Preprocessing

The pre-processing step includes document cleaning, format normalization, sentence segmentation, word stemming, etc. Besides, we have also done document collection model construction to prepare for the second step – summarizing.

Document Cleaning In order to utilize the dataset more efficiently, some contents should be filtered from the document, such as the tags, the news agencies' name, and the ESC characters, etc. For this type of contents is the noise information in the document.

Format Normalization In order to utilize the structure of the document, the paragraph structure should be identified firstly. In our system, the paragraphs are transfer into the normal format – each paragraph a line follows with a paragraph mark.

Sentence Segmentation Sentence is the basic element in extractive summarization in general. In order to integrate sentence segmentation into our system, a rule-based segmentation method is designed and implemented except from using DUC tool of sentence breaking.

Word Stemming In English, many words indifferent forms but with the same root, e.g. clued and clue, often share the same meaning. Therefore, the similarity score can not be measured with the words themselves but be measured with words' root.

2.2 Summarizing

Although the document collection used to generate a MDS may be relevant to the same general topic, they do not necessarily include the same information. Extracting all similar sentences would produce a verbose and repetitive summary, while extracting some similar sentences could produce a summary biased towards some sources, as it was noted in [2]. However, the graph-based extractive summarization algorithm succeeds in identifying the most important sentences in a document collection based on information exclusively drawn from the collection itself. We propose an iterative graph-based algorithm to obtain the most important sub-topics in global space. The algorithm starts from the assumption that capturing sub-topic structure of document collection is essential for summarization. It firstly creates graph representation of document collection, then selects the salient sentence with a new rank criterion and obtains the most important sub-topic in global graph space iteratively, finally forms the summary supported by the real salient sentences of different sub-topics.

2.2.1 Rank Criterion

To assess the salience of nodes in graph, we propose a new sentence ranking criterion served as basis for our method. This criterion has inspired by the ideas in information retrieval and feature selection. Since the summarization is controlled by choosing the central sentences, which we call "salient sentences", it is in principle possible for the salient sentences to be scored according to the word based features - the statistical features or semantic features according to words or phrases - and the global features.

$$g(u) = f_1(u) \cdot f_2(u) \qquad (1)$$

where $g(u)$ is the salience score of sentence, $f_1(u)$ is the score of word based features, and $f_2(u)$ is the score of global features. We can use the product of the two classes of features to assess the salience of sentence u, for they belong to two different feature spaces.

2.2.2 Feature Selection

Our approach to text summarization allows both generic and topic-oriented summaries by scoring sentences with respect to both statistical and linguistic features. For topic-oriented

summarization, a topic vector is calculated using the title and narrative content of the specified topic of a document collection. Each sentence is scored according to the word based features and the global features.

Here, global features mainly consider the length, the position, the temporal order and some text patterns of the sentence. A simple fact is that short sentences cannot carry enough information corresponding for the topic. Thus, too short sentences are not appropriate candidates of summary sentences and will not be considered. And due to the constraint of summary length, too long sentences are not appropriate too. There are some patterns which are unsuitable for being in the summary. The sentences which have these patterns will be discounted for summary sentence. For example, the sentences with somebody saying, e.g., somebody say/ said/ says, "...", should be used a penalty, instead of being banned.

2.3 Post-processing

In our system the information redundancy between sentences can be decreased with our algorithm, so we do not need use the MMR [3] to avoid redundancy. However, we should conduct some remedy to avoid the redundancy inside the sentences. Therefore, we employ a method combined with some naïve rules and similarity discrimination to simplify sentences. In which just the elements of the sentence – clauses in general – not similar to the topic can be simplified. With this method, the information comprehensiveness in summary can be developed with the increase of sentences' count for the reason of sentence simplification. Except from sentence simplification, the sentences' reordering is another important task in the step of post-processing. In order to assure the readability of a summary, the sentences should be organized with a reasonable order. Here, a reordering rank method combined with the documents' time order in document collection and the sentences' location order in the document is proposed to calculate the order score.

3 Evaluation

3.1 Test Data and Metrics

Different from DUC 2006, DUC 2007 divided into two tasks – main task and update task. The main task is similar to previous task in DUC 2006, which provides 45 document sets for test valuation (unlike the 50 document sets in DCU 2005 and DUC 2006). And each document set includes a fixed number – 25 documents and its query. Each query contains a query title and a query narrative. A query title is usually a phrase which describes briefly the topic. A query narrative is usually composed of several factoid or definition questions, which need answers given in the summary. NIST assessors created 4 reference summary for each topic. There are 32 participants in DUC2007 main task, each participant submit one summary. All submitted systems are either manually or automatically evaluated, including linguistic quality, responsiveness, ROUGE-2, ROUGE-SU4 [4], and Pyramid [5].

The update is a new task added in this year, which provides 10 document sets selected from the document sets of main task. There are 25 documents in each document set, and the documents will be ordered chronologically and then partitioned into 3 subsets, A-C, where the time stamps on all the documents in each subset are ordered such that time(A) < time(B) < time(C). There will be approximately 10 documents in Subset A, 8 in Subset B, and 7 in Subset C [6]. There are 24 participants in DUC2007 update task, each participant submit three summaries corresponding to the subset A, B, C. Same with main task, all system in update task are either manually or automatically evaluated, including linguistic quality, responsiveness, ROUGE-2, ROUGE-SU4, and Pyramid.

3.2 Evaluation Results

3.2.1 Main Task Results

Among the manual evaluation results of 32 systems, our submitted system ranks 11th in the content evaluation, 7th in the linguistic quality evaluation, 3rd in the grammaticality evaluation, 10th in the non-redundancy evaluation, 15th in the clarity evaluation, 12th in the focus evaluation, and 13th in the structure and coherence evaluation. In the automatic evaluation, our submitted system obtains the 17th, 17th and 14th respectively in ROUGE-2, ROUGE-SU4 and BE. Table 1 shows

the detailed scores of our submitted system – ICTGSP - in DUC 2007 by manual evaluation, and table 2 is the scores of our system in automatic evaluation.

Table 1: The Main Task Manual Scores of ICTGSP-S

Main Task	Rank	Score	Best
Content	11	2.89	3.4
Linguistic Quality	7	3.48	4.11
Grammaticality	3	4.4	4.64
Non-Redundancy	10	3.76	4.18
Ref Clarity	15	3.31	4.09
Focus	12	3.49	4.24
Structure&Coherence	13	2.47	3.69

Table 2: The Main Task Results of ICTGSP-S

Main Task	Rank	Score	Best
ROUGE-2	17	0.0975	0.124
ROUGE-SU4	17	0.15109	0.177
BE	14	0.05445	0.066

Besides the detailed scores of our system in main task, we can also obtain the detailed scores of all the systems participated in DUC 2007. Based on evaluation scores analysis, we can plot a figure of all the systems in figure 2 and figure 3. In figure 2, all the systems' score are ordered by the avg. content scores, while the scores in figure 3 are ordered by ROUGE-2 scores, and our GSP-S system is marked by the red bar. From these two figures, our system is more stable than most other systems in manual and automatic scores for the reason that our system's vibration is rather small.

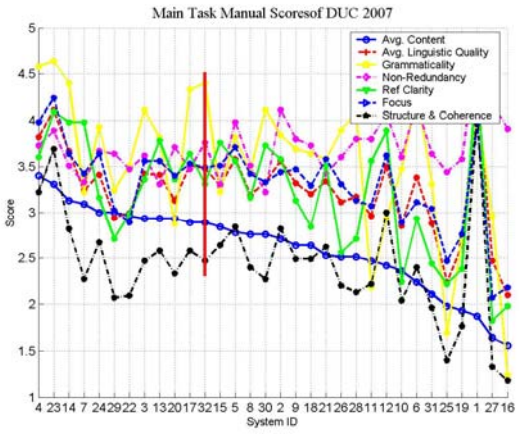

Figure 2: The Main Task Manual Scores of DUC 2007

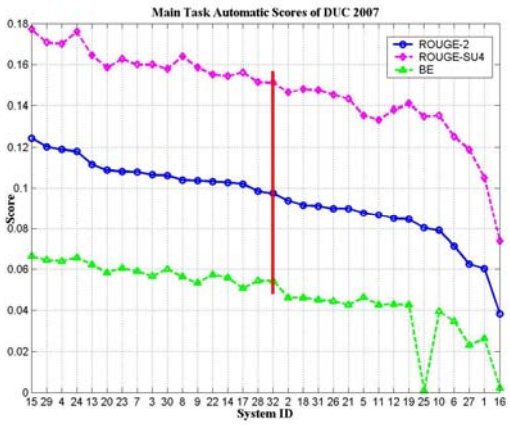

Figure 3: The Main Task Automatic Scores of DUC 2007

Our system gains a middle-above rank among all the systems, but the scores of the submitted system are not the true ones for our system. Due to our negligence, a result without word stemming was submitted. Consequently, the performance of our system is reduced a lot. Table 3 shows the automatic evaluation results of our system with word stemming, and our system can obtain the 11th and 12th respectively in ROUGE-2 and ROUGE-SU4. In the same way, the manual evaluation would be improved a lot.

Table 3 The New Main Task Results of ICTGSP-S

Main Task	Score	Rank
ROUGE-2	0.10585	11
ROUGE-SU4	0.15823	12

3.2.2 Update Task Results

In the second task of DUC 2007, update task, our submitted system ranks 9th in ROUGE-2, 7th in ROUGE-SU4, and 11th in BE with the automatic evaluation among 24 systems. In the manual evaluation, our GSP-S system gets the third place in responsiveness evaluation. The following table (Table 4) is our detailed scores in the above four evaluations.

Table 4: The update task results of ICTGSP-S

Update Task	Rank	Score	Best
ROUGE-2	9	0.0876	0.1119
ROUGE-SU4	7	0.1285	0.14306
BE	11	0.0463	0.07219
Responsiveness	3	2.767	2.967

Same as we did above, we can plot a figure of all the 24 systems in the update task with the results returned by NIST. In figure 4, all the systems are ordered by the ROUGE-2 scores, and the pink curve in the figure is the smoothing responsiveness (Responsiveness' in Figure 4) in order to contrast with the other three curves. Our submitted system, the one marked by the red bar, is competitive in the ROUGE-2, ROUGE-SU4 and responsiveness evaluation.

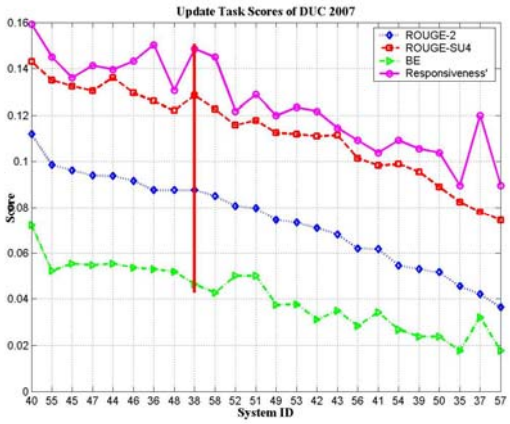

Figure 4: The Update Task Scores of DUC 2007

4 Conclusion and Future work

In this paper, we described our participation in DUC 2007. Through this participation, we get a lot of lessons and experiences in summarization research, and there is still room to improve our graph-based sub-topic partition method, GSP-S. How to use the useful features to obtain the salient sub-topics and how to formalize our algorithm more effective would become the important research aspects for our system's performance. It seems that there is a good foundation for our future research.

5 Acknowledgements

The work is supported by National Grand Fundamental Research 973 Program of China "Large-Scale Text Content Computing" under Grand NO. 2004CB318109.

References

[1] DUC, http://duc.nist.gov/

[2] R. Barzilay, K. R. McKeown, and M. Elhadad. Information Fusion in the Context of Multi-Document Summarization. In Proceedings of the 37th Annual Meeting of the Association for Computational

[3] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, Auguest 1998.

[4] Lin.C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization, Barcelona. ACL.

[5] Pyramid, http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html

[6] Update Task, http://duc.nist.gov/duc2007/tasks.html#pilot