TAC 2009 Update Summarization Task of WUST

Maofu Liu, Bo Yu, Fang Fang and Hao Sun
College of Computer Science and Technology,Wuhan Universitity of Science and Technology,
Wuhan 430065, P.R.China
{liumaofu,yubo,fangfang}@wust.edu.cn

Abstract

The TAC 2009 update summarization task is to generate not more than 100-words fluent summaries for the given document sets of newswire articles. The newswire articles are relevance to the same event topic, stated by title and narrative If we take the background information of the topic into consideration according to the topic statement, it will help to understand the content of the news articles document set comprehensively and accurately. In our system, we introduce the Wikipedia article, related to the topic, to provide such background information. The experiment results show that it can make encouraging improvement comparing to the run not taking Wikipedia article into consideration.

1 Introduction

The update summarization task of TAC 2009 (http://www.nist.gov/tac/) aims to generate a short (not more than 100 words) fluent summary for a set of newswire articles. For the document sets of TAC 2009, relevant documents are as close together in time as possible comparing to those of TAC 2008.

The test dataset is composed of 44 topics. Each topic has a topic statement (title and narrative) and 20 relevant documents which have been divided into two sets, Document Set A and Document Set B. Each document set has ten documents, and all the documents in Set A chronologically precede the documents in Set B.

For a given topic, the task is to write two summaries, one for Document Set A and one for Document Set B, according to the topic statement. The summary for Document Set A should be a straightforward summary, but The update summary for Document Set B should be written under the assumption that the user of the summary has already read the documents in Document Set A.

Wikipedia is a huge online encyclopedia with a large number of users to create and add entries to it. Wikipedia has more than three million of articles in English now, almost covering all the topics. We pick-up appropriate name entities or key words from title and narrative of topic, search them on the Wikipedia to obtain the relevant Wikipedia article.

In our update summarization system, we generate the summary for document set with the assistance of Wikipedia article. We take the first paragraph of entry explanation provided by Wikipedia article, relevance to the topic, as a special single-document summary. We take advantage of this special summary to filter sentences of document set B. The similarity between sentence of the candidate multi-document summary and the one of the special summary will be calculated, and sentences whose significane score is lower than the threshold value should be removed.

The remainder of this paper is organized as follows. Section 2 introduces the summarization system. Section 3 focus on the update summarizationn. Section 4 presents evaluations results. Finally, Section 5 concludes the paper and talks about the future works.

2 System Overview

The statistics on the biology indicate that when the text are manually indexed, 42.7% of the words are selected from the original text, 47% of the words can be got by synonymousness of the words (Christiane, 1998). According to this distribution, the summary sentence can be gained directly or indirectly.

It considers text as a linear combination of the sentence, takes the sentence as a linear combination of the word with automatic extraction(Lin, 2004). By calculating score of each feature items in the sentence and sorting sentence by the score, summary is generated.

Our system consists of four phases, pre-processing, features selection, summary generation and post-processing.

2.1 Pre-processing

In accordance with sub-word dictionary and stopword list, pre-processing phase segments the text and tags part of speech, statistics of the keywords frequencies, records the location of the words and other basic information of words. Our system divides every document of the document corpus provided by TAC2009 into sentence units by using GATE（http://gate.ac.uk/） as text segmenting tool.

We use GATE to firstly segment word and clause, then extract the stem of each word in the TAC2009 document, mark the part of speech of each word and check whether it is a stop word, lastly calculat the TF * IDF values for each word , and generate pre-processing file (C. R. Chowdary, M. Sravanthi, et al., 2008).

2.2 Features Selection

According to the results of pre-processing for each sentence, score of four feature items, sentence TF*IDF, position, subject-related degrees and the length of the sentence, will be calculated.

(1) Sentence TF*IDF

TF*IDF is product of the frequency of a phrase in the document and reciprocal of the frequency of the phrase appearing in the entire document corpus. In other words, more frequently phrase appears in this document and not in other documents, higher amount of information the terminology have. This feature indicates that more key words the sentence contains in the document, higher the this score will be. This score is the sum of TF*IDF value of the sum of all the words after removing stop words(Mihalcea, 2004).

$$ST_{i,k} = \sum_{w\in Sen_{i,k}} (TDoc_{i,k}(W) + TTopic_{i,k}(W)) \qquad (1)$$

Where $Sen_{i,k}$ means the first k-documents for the first i-sentences, w means $Sen_{i,k}$ in the non-stop words, $ST_{i,k}$ means the score of TF * IDF feature items in $Sen_{i,k}$, $TDoc_{i,k}$, $TTopic_{i,k}$ respectively means, the word w in the document in the TF * IDF score, as well as information in the subject of the TF * IDF score.

(2) Sentence Position

Our system considers text as a linear combination of the sentence. First sentence of each document text is considered as the most important, and importance of other sentences reduced by location backwards (E. Hovy and C. J. Fukumoto, 2006).

$$P_{i,k} = \frac{n - i + 1}{n} \qquad (2)$$

Where $P_{i,k}$ means scores of location as a feature item, *n* means the total values of the sentences in number *k* document.

(3) Similarity between Topic and Sentence

Topic is the center of the document. The greater similarity between each sentence is, the more important information the sentence contains, then the sentence becomes more important (U.von Luxburg, 2007).

$$S_{i,k} = SSenTopic_{i,k} + SDocTopic_{i,k} \qquad (3)$$

Where $SSenTopic_{i,k}$ means the directly similarity score between the sentences and the subject heading sentences, $SDocTopic_{i,k}$ means the indirect similarity score between the sentences and the subject heading sentences.

(4) Sentence Length

The normal distribution model is used to calculate this score. More the length of the sentence is closer to the average length, higher scores this feature obtains. The average length of sentence is the value that the sum of the words of all documents under the same topic is divided by the sum of sentence (J. Xiao and J. Yang ect, 2007).

$$L_{i,k} = \frac{1}{\sqrt{2ps}} e^{-\frac{(x_i - u)^2}{2s^2}} \qquad (4)$$

Where m means the average length of all the sentences in the same document, x means the number of words contained in $Sen_{i,k}$.

2.3 Summary Generation

Each participation team will need to submit two set of results obtained by different methods. In our first run (run 1), we get similarity between the topic and sentences in document sets by the four feature items to pick-up the sentences that contain important information to generated summary. To generate summary, the document set A is inputed into the system, and our system calculated each score of four feature items, then according to this scores sentences are sorted in accordance with descending order.

When using Wikipedia to generate the first summary in our second run (run 2), we identify suitable key phrases or key words from 44 different topic documents to input into Wikipedia to search at first, and then take the first paragraph of content in each corresponding Wikipedia articles as a new document set, called document set W. We firstly generate the first summary A1 based on the document set A, and then sentence similarity between the first summary A1 and document set W will be calculated, and the sentences are sorted in the descending order. The sentences that contain first 100 words will be assembled to the second summary A2.

2.4 Post-processing

The summary should be not more than 100 English words in length. We adopt English grammar as guidance to make a few rules with regular expressions to remove clauses with obvious characteristics, appositive and "somebody says," or something like this (A. Gretton, B. Scholkopf, et al., 2003). For example, clause starting with a comma and followed by "wh-" or other guidance words will be considered as a non-restrictive attributive clause. These sentences shall be deleted. In our system, we have developed 11 rules to eliminate non-restrictive attributive clause, time phrases, and so on.

3 Update Summary

Mission for update is firstlly to remove the sentences in the document set B that contain the same or similar information with ones in the document set A, and keep the rest of sentences correlative with topic as much as possible. We can calculate similarity between all sentences of document set B and each sentence in document set A, and remove the sentences with higher correlation in document set B. But it costs too much in time with poor efficiency. We also could do similarity calculation just between all sentences of document set B and each sentence in summary A. It is proved to be feasible and effective. So we input document set B into system, system will do similarity calculation as mentioned above, and sort sentences in descending order. The sentences whose score is larger than the threshold value are considered to contain repetitive information, and should be removed from Docset B. After that, we have Docset C. Docset C is inputed system to generate summary B1 like summary A1. Summary B1 is the summary for update in run1. So we have two summary A1 and B1 in run1 last.

When using Wikipedia to generate update summary in run 2, similarity calculation will be done between summary A1 and document set B. The sentences with high similarities shall be removed from the document set B, and the result is intermediate document set C. Similarity calculation will be done between document set C and document set W, and sentences with lower correlation will be deleted to get document set D. This document set is the material collection need to be deal with for update to generate summary B2, so we have two summaries A2 and B2 for run 2 finally.

4 Evaluations

Table 1 shows the scores in ROUGE-2 and ROUGE-SU4 with using Wikipedia (Run 2) and not (Run 1). Table 2 shows the scores of two groups of summaries in manual evaluation. We can find that the scores using Wikipedia are better than those of not using Wikipedia.

Table 1. Comparative ROUGE Scores using Wikipedia and Not

	ROUGE-2	ROUGE-SU4
Run 1	0.02625 (0.02297 - 0.02964)	0.06551 (0.06172 - 0.06936)
Run 2	0.04262 (0.03888 - 0.04644)	0.08446 (0.08082 - 0.08802)

Table 2. Comparative Manual Scores using Wikipedia and Not (Py.: Pyramid; Ling.: Linguistic Quality; Resp.: Overall Responsiveness)

	Document Set A			Document Set B		
	Py.	Ling.	Resp.	Py.	Ling.	Resp.
Run 1	0.062	3.636	2.455	0.050	3.659	2.227
Run 2	0.133	3.614	2.864	0.081	3.909	2.636

5 Conclusions

In this paper, we describe the four selected features of our update summarization system, and use Wikipedia to filter sentences for generating a summary to improve the relevance between sentence and topic. The experiment shows that the results of using Wikipedia are better than those of not using Wikipedia.

But the generated asummary is not ideal enough in readability, linguistic structure and redundancy degree. This is our next important point to improve. We will compress sentences with the grammar and language usage information.

References

Christiane, F, 1998. WordNet: an Electronic Lexical Database. MIT Press.

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

C. R. Chowdary, M. Sravanthi, and P. Sreenivasa Kumar. "QueSTS: A Query Specific Text Summarization System" In Proceedings of the 21st International FLAIRS Conference, pages 219 - 224, Florida, USA, May 2008. AAAI Press.

Mihalcea, R., and Tarau, P. “TextRank: Bringing order into texts”. Proceedings of EMNLP 2004, 404–411. Barcelona, Spain: Association for Computational Linguistics.

E. Hovy, C. J. Fukumoto, and Lin, L. Zhou. Automated Summarization Evaluation with Basic Elements. In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC), 2006.

U. von Luxburg. A tutorial on spectral clustering. Statisticsand Computing, 17(4):395–416, 2007.

J. Xiao, J. Yang, and X.Wan. Manifold-ranking based topicfocusedmulti-document summarization. In IJCAI, pages2903–2908, 2007.

A. Gretton, B. Scholkopf, D. Zhou, J. Weston, and, O. Bousquet. Ranking on Data Manifolds. In Proceedingsof NIPS 2003, 2003.