Automatic Bilingual Text Document Summarization

Sau-Han Lo and **Helen Mei-Ling Meng** and **Wai Lam**
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong,
Shatin,
Hong Kong
{shlo, hmmeng, wlam}@se.cuhk.edu.hk

Abstract

We develop a language independent text document summarization technique using Thematic Term approach. Corpus-based thematic terms capturing the main coverage of the whole corpus are automatically determined. Article-based thematic terms capturing the main theme of a document are extracted. Representative sentences are determined by considering both kinds of thematic terms and summaries can then be generated. We can successfully apply our summarization method to both English and Chinese documents. Many existing evaluation methods require either manual summaries or human judgments. We propose a novel evaluation technique based on information retrieval which does not need these requirements.

1 Introduction

We investigate language independent automatic text document summarization techniques. We focus on the problem of generating a kind of summary called extracts. Extracts are representative units taken from the original texts. We target to generate extracts that readers or users can make use of them to act as representatives of the original document. After reading the summaries, they can get the main idea of the original document and topic related to the whole corpus as well. Most existing text document summarizers are extract-based. Extracts can be different units of material of the original text such as sentences. Different criteria are used to determine the representativeness of each sentence. Sentences with high score will be collected to form the extractive summary. Those criteria are related to features such as stochastic measurements for the significance of key terms in the sentence (8; 13; 5) , sentence location in the source text (13; 1; 3), the presence of cue or indicator phrases (11) or title words (3) and professional names (13). Apart from using features, heuristic rules are developed to contribute to the score. Teufel considered the sentence length for generating extractive summaries (13). Kupiec utilized uppercase word feature, believing that proper names are important in the extracts (5; 7; 4). Besides using positive indicators to locate the important sentences, Goldstein tried to find negative indicators for summary (4). Those negative indicators are anaphoric reference, honorific, negation words, auxiliary verbs, integers, evaluative and vague words as well as conjunctions and prepositions.

We develop a language independent text summarization technique using corpus-based thematic terms and article-based thematic terms. Techniques are developed for automatically identifying and integrating both kinds of thematic terms. Therefore, the summary contains the main theme of a document as well as of the main idea of the whole corpus. Besides, we can successfully apply our summarization method to both English and Chinese documents.

In addition, we examine evaluation methods to access the quality of the summaries. Summarization performance can be evaluated by two approaches: one is intrinsic and the other is extrinsic evaluation. Intrinsic evaluation can be applied to sentence extractive summaries. The performance can be assessed using traditional recall and precision metrics. However, some experts need to determine the relevance judgment on each sentence prior to evaluation. Extra costs and time are required. On the other hand, extrinsic evaluation assumes that the summary will be used for certain tasks. In other words, many existing evaluation methods require either manual summaries or human judgments. We propose a novel evaluation technique based on information retrieval which does not need these requirements.

2 Our Thematic Term Approach

We propose a language independent summarization approach which extracts representative sentences from the original text article based on thematic terms. We consider two kinds of thematic terms, namely, corpus-based thematic terms and article-based thematic terms. These thematic terms are automatically extracted using information retrieval techniques. Unlike many existing summarization methods, our approach considers not only information contained in a single article, but also the information conveyed from the whole corpus. We attempt to apply our summarization method on both English and Chinese documents.

The architecture of the system consists of Document Preprocessor, Corpus Thematic Term Extractor, Article Thematic Term Extractor, and Sentence Score Generator. The Document Preprocessor converts the raw texts to suitable representation for subsequent processing. The purpose of the Corpus Thematic Term Extractor and Article Thematic Term Extractor is to extract corpus-based and article-based thematic terms respectively. The Sentence Score Generator is responsible for generating the score for sentences considering both kinds of thematic terms.

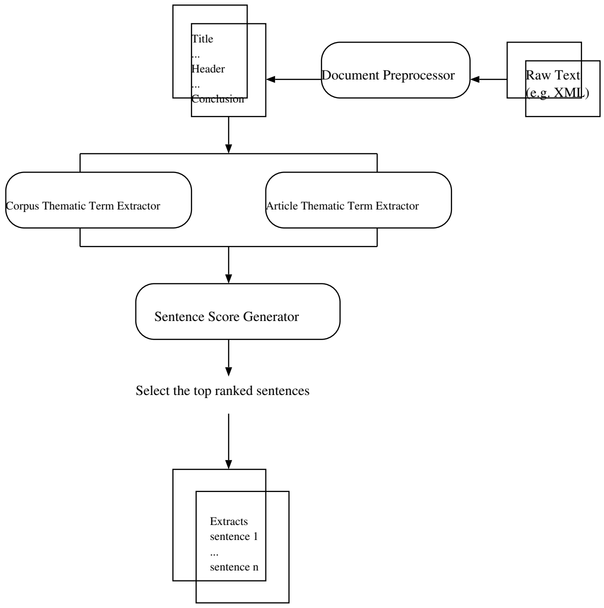

Figure 1: An Overall System Overview

2.1 Document Preprocessor

2.1.1 English Corpus

The English corpus used in our experiment is the Computation and Language (cmp-lg) corpus. It was prepared by MITRE Corporation based on extensions to some initial work carried out at the University of Edinburgh. There are 183 scientific papers in XML format and they are made available as a general resource to the information retrieval, extraction, and summarization communities. Markup tags cover core information such as title, author, date, as well as basic structure such as abstract, body, sections, lists, etc. Figures, tables, equations, cross-references and references were all replaced with placeholder tags.

2.1.2 English Corpus Preprocessor

XML formatted texts are converted to the "body" text and "abstract" text. Body text refers to the text excluding the abstract section. Abstract text refers to the author written abstract. The body and the abstract are processed by the basic document preprocessor step like stemming, case folding, removal of all XML tags, figures, equations, tables, stop words, and punctuation.

2.1.3 Chinese Corpus

The Chinese corpus used in our experiment comes from the Topic Detection and Tracking Evaluation Project (TDT) organized by DARPA and NIST. The corpus contains news data collected daily from 9 news sources in two languages (American English and Mandarin Chinese), Chinese sources are Xinhua News Agency, Zaobao News Agency, and Voice of America. The texts are encoded in GB code.

2.1.4 Chinese Corpus Preprocessor

The Chinese corpus is passed through a word segmentation module so that Chinese characters are grouped as segmented words which are more meaningful than discrete single character. The word segmentor is obtained from Linguistic Data Consortium (LDC). The algorithm makes use of dynamic programming to find the path which has the highest multiple of word probability.

2.2 Corpus Thematic Term Extractor

The objective of Corpus Thematic Term Extractor is to determine a set of corpus-based thematic terms. We use the Inverse Document Frequency (IDF) weighting scheme to find the salient terms from the corpus. The idea is that a term occurring in a few documents is likely to be a better discriminator than a term appearing in most documents. The inverse document frequency of a term j , IDF_j, is given by:

$$IDF_j = \log_2(N_c/DF_j) + 1 \tag{1}$$

where DF_j refers to the total number of documents containing term j and N_c refers to the total number of documents in a corpus. After calculating the inverse document frequency weight value for all the distinct terms in the corpus, the terms will be ranked according to the weight in descending order. We use the threshold, θ_c, to select the top ranked terms and form the set of corpus-based thematic terms.

2.3 Article Thematic Term Extractor

If we want to get some important and meaningful sentences, we should also consider the main theme in the original document. Intuitively, key terms found from a document are also important to contribute the sentence weight. Therefore, inspired by the idea in (7), we consider Inverse Sentence Frequency, ISF_j, defined as follows:

$$ISF_j = \log_2(N_s/SF_j) + 1 \tag{2}$$

where SF_j refers to the total number of sentences containing term j and N_s refers to the total number of sentences within a document. After calculating the Inverse Sentence Frequency value of all distinct words found in a document, we rank them by using a threshold, θ_a, we select a portion of terms with high ISF to form the article-based thematic terms.

2.4 Sentence Score Generator

After all corpus-based and article-based thematic terms are found, we calculate the score for each sentence as shown below:

$$S_k = \beta \sum_{j=0}^{T_k} (TF_j IDF_j) + (1-\beta) \sum_{j=0}^{T_k} (TF_j ISF_j) \tag{3}$$

where S_k is the weight for sentence k; T_k is the total number of tokens and TF_j is the total number of occurrence of term j found within a sentence k . We set β as the combination factor between the value of term j found from the whole corpus and found in an article. If terms found in sentence are salient for both corpus and article, then such sentence gains higher score. It implies that such sentence is representative within an article and in the corpus as well.

After the score of each sentence is computed, we rank all sentences in descending order. With different compression rates, top ranked sentences are selected out. Then the sentences are arranged according to their chronological order in the original article to form a summary.

3 Summarization Evaluation

Currently, there is still no golden or standard evaluation method to assess how good of a summary is. Some researchers adopt the Cosine Similarity Measure which calculates the score of vocabulary overlapping between the system extracts and the human written abstracts / summaries. We also try this as one of our evaluation methods. Unfortunately, we find that there are some drawbacks of this method. If there is no human written abstracts / summaries provided, we can not apply this method. Furthermore, if a writer uses different vocabularies to construct the abstract / summaries, there is no clear overlapping between the system extracts and the author written one. As a result, the similarity score will be very low. In view of this, we propose a new evaluation method, called Average Inverse Rank (AIR) method.

3.1 Cosine Similarity Measure

Cosine similarity measure is one of traditional methods to assess the goodness between system extracted summaries and author written abstract (9; 10). System extracts and author written abstract are firstly converted to corresponding vector forms. Next we calculate the dot product between two vectors to get the similarity score. The weight of the vector is calculated as follows:

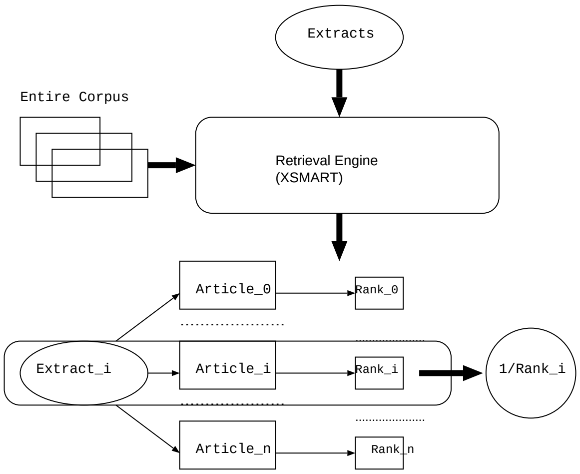

Figure 2: The idea of Average Inverse Rank (AIR)

$$W_{e,j} = (TF_j)(\beta IDF_j + (1-\beta)ISF_j) \tag{4}$$

where $W_{e,j}$ is the weight of the term j in Extract e, TF_j is the term frequency of the term j in Extract e.

$$W_{A,j} = (0.5 + 0.5TF_j) \times (\beta IDF_j + (1-\beta)ISF_j) \tag{5}$$

where $W_{A,j}$ is the weight of the term j in Abstract A, TF_j is the term frequency of the term j in Abstract A

$$C(A,e) = \frac{\sum_{j=0}^{T_A}(W_{e,j} \times W_{A,j})}{\sqrt{\sum_{j=0}^{T_A} W_{e,j}^2 \times \sum_{j=0}^{T_A} W_{A,j}^2}} \tag{6}$$

where $C(A,e)$ is the similarity function related to the Abstract A and Extract e and T_A is the total number of distinct terms found in the Abstract A.

3.2 Average Inverse Rank (AIR)

Generally, good summaries should be relatively shorter in length compared with the original document and capture the representative information of the lengthy document. The idea of AIR is based on how good the summary can act as a query to retrieve the original full-length document. The mechanism is that given a collection of the original documents, if a particular summary can retrieve its original document from the pool of the document collection, then this summary is capable of characterizing the original document. To implement this evaluation technique, we make use of an information retrieval engine known as XSmart(6). XSmart extends from Smart(2) so that it can handle both Chinese and English. The original documents are first indexed by the indexing engine in XSmart. Then each summary is treated as a query and a retrieval run is conducted. The result of the retrieval is a ranked list of the original documents. The specific rank of the particular document corresponding to the summary can be determined. Then taking the inverse of the rank will get the *Inverse Rank*. Using the same procedure, we can find out the Inverse Rank of all the extracts. Finally, taking the average of it can give the *Average Inverse Rank* of the summarization system.

4 Experiments

For both Chinese and English corpora, we divide the whole corpus into two portions. The first portion consists of 30 percent of the whole corpus and it is used for parameter fitting. The remaining portion is used for evaluation.

For English corpus, since a manual summary is available, we introduce a baseline summarization for comparative study. It can be observed that writers always present and emphasize the key concept of the whole passage in the introduction and conclusion sections. With this intuition, we select sentences randomly from these two sections given a compression rate and form the baseline summary.

4.1 Cosine Similarity Measure

Since the English corpus has manual summaries, we can conduct cosine similarity measure for the English corpus only. In Fig. 3 "best" refers to the best cosine similarity value taken from all different β values ranged from 0 to 1. Conversely, "worst" implies the lowerest cosine similarity value comes from different β values. Our approach outperforms the baseline, showing that the extracts contain more important information than the introduction and conclusion sections do. This confirms that extracts generated by considering both corpus thematic terms and article thematic terms are able to give more representative information. Moreover, there is a great difference in the best β case

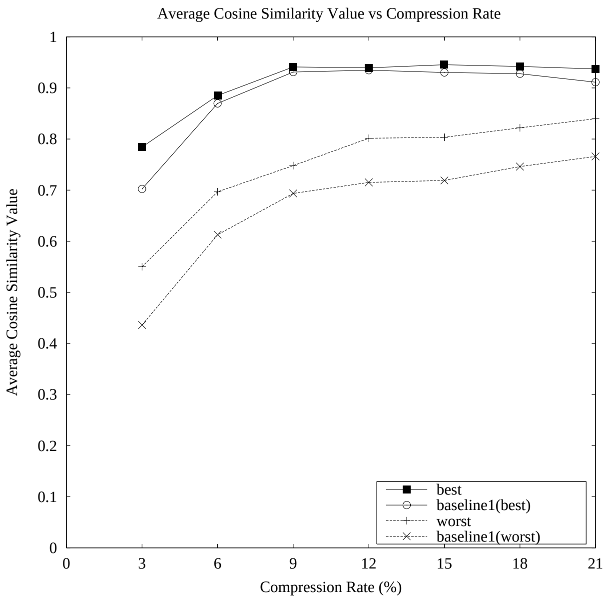

Figure 3: The cosine similarity measure of our thematic term approach and the baseline approach

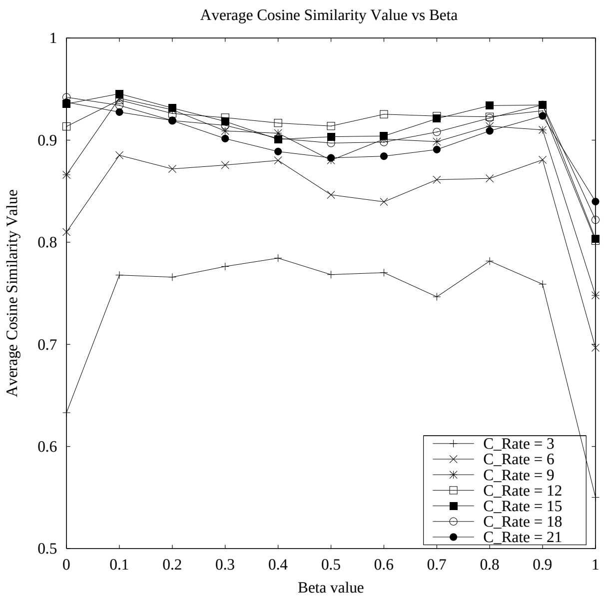

Figure 4: The cosine similarity value with different beta set under different compression rate

and the worst one. This implies that if poor β is chosen, then the quality of extracts is affected.

In Fig. 4, "C_{Rate}" stands for compression rate. As the compression rate increases, the similarity value approaches more or less the same. The reason is that if more sentences

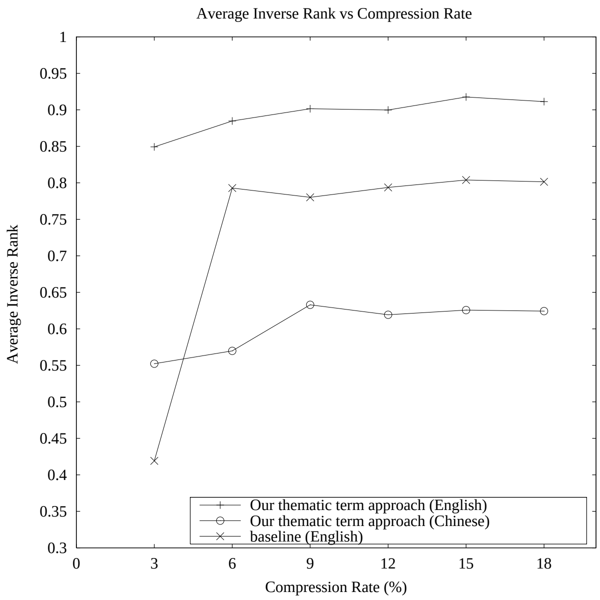

Figure 5: AIR Results on our thematic term approach and the baseline approach

are included in extracts and the number of vocabularies in author written abstracts are unchanged, there is a higher chance of more vocabularies to be included in author written abstracts as the compression rate increases.

4.2 Average Inverse Rank Measure

Fig. 5 shows that the AIR measure increases as the compression rate increases. The reason is that the summaries possess greater ability to retrieve the original documents from a pool of document collection. Moreover, there is one observation about the compression rate. It reveals that about 10 percent of the full-length documents are already to be good surrogates representing themselves.

For English baseline summary, there is a sharp increase at the beginning and then it lies flat. It is because there are not enough sentences covering the required compression rate in some cases. Hence, the AIR value becomes steady.

Fig. 5 also shows that our thematic term approach outperforms the baseline for English corpus. Those extracts contain more useful terms.

5 Conclusion and Further Enhancement

We have developed an automatic text document summarization by considering both the corpus thematic terms and article thematic terms. Experiments reveal that combination of them is able to give a representable extractive summary especially under the low compression rate. We also successfully apply our summarization technique to both Chinese and English documents. In addition to this, a new summarization evaluation method, called Average Inverse Rank is introduced. It offers an alternative way to cope with the drawback of traditional cosine similarity measure. One future extension of our work is to investigate a technique for generating event-based summary from multilingual news document.

References

[1]Baxendale, P. B.: Man-Made Index for Technical Literature — An Experiment, IBM Journal of Research and Development 2:4, 1958, 354-361.

[2]Buckley, C: Implementation of the SMART information retrieval system. Cornell University Technical Report 85–686, 1985.

[3]Edmundson, H. P.: New methods in automatic extracting. Journal of the Association for Computing Machinery, 16(2):264–285, April 1968

[4]Goldstein, J., Kantrowitz, M., Mittal, V. O. and Carbonell, J.: Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Proceedings of SIGIR. Berkeley, CA, August 1999.

[5]Kupiec, J., Pedersen, J. and Chen, F.: A Trainable Document Summarizer. Proceedings of SIGIR 1995, Association for Computing Machinery, New York, 1995.

[6]Lam, W, Wong, C. Y. and Wong, K.F.: Performance Evaluation of Character-, Word-, and N-Gram-Based Indexing for Chinese Text Retrieval. Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages IRAL, pp. 68–80, 1997.

[7]Larocca Neto, J., Santos, A.D., Kaestner C.A.A., Freitas A.A. and Nievol J.C.: A trainable algorithm for summarizing news stories. Proceedings of the PKDD-2000 Workshop on Machine Learning and Textual Information Access. Lyon, France. Sep. 2000.

[8]Luhn, H. P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2 (2), April, 1958, pages 159–165. 1958

[9]Marcu, D.: Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, editors, Advances in Automatic Text Summarization. MIT Press, 1999. To appear.

[10]Nakao, Y.: An Algorithm for One-page Summarization for a long Based on Thematic Hierarchy Detection. Proceedings of ACL, pp. 302–309, 2000.

[11]Paice, C. D.: The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Oddy, R. N.; Robertson, S. E.; van Rijsbergen, C. J.; and Williams, P. W., eds.: Information Retrieval Research. London: Butterworth. 1981.

[12]Teufel, S. and Moens, M.: Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (eds.), Advances in automatic Text Summarization. MIT Press, 1999.

[13]Teufel, S. and Moens, M.: Sentence extraction as a classification task. In: ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, pp. 58–65. 1997.