Generating Summaries of Multiple Technical Articles

Minghui WANG*†, Hediheko TANAKA* and Yixin ZHONG†

*Department of Engineering, The University of Tokyo, Japan

†The Research Center of Intelligence, Beijing University of Posts and Telecommunications, China

mhwang@mtl.t.u-tokyo.ac.jp

Abstract

The research on document summarization has been becoming attractive and important because of the explosive increase of the scientific information. In this paper, we try to develop a system framework that focuses on performing a summarization of multiple papers on a specific domain which is considered more useful for reader to grasp the outline of a research domain.

Keywords: Automated Text Summarization, Summarization of Multi-paper, Sentence Extraction, Citation Information

1 Introduction

The abundance of information and the difficulty to acquire wanted information on a specific topic may be one of the major problems for a reader in this age of information overload. Usually, the reader has to go through Megabytes of data everyday to select articles of interest and read the relevant parts of them in order to get needed materials [Kathleen McKeown and Dragomir R. Radev (1995), Dragomir R. Radev. (1996)]. Hence, he needs *search* and *selection* services (one of the very popular topics in present Internet society), as well as for *Summarization* facilities.

The automatic summarizer is considered as a software system which can produce a synopsis of any document submitted to it. With few exceptions, automatic approaches to summarization have primarily addressed possible ways to determine the most important parts of a text [Luhn,H.P (1958), Daniel Marcu (1997), Eduard Hovy and ChinYew Lin (1999)]. The so-called summary produced by most of the currently existing text summarization systems, including some Web browsers is actually portions of the text, which is just a *Extrac*t. A truly comprehensive and informative text summary should combine together various concepts of the original text and generate a new text, or called *Abstract*. Another major obvious problem of the existing summarization systems is that it only summarizes a single article at a time [Kathleen McKeown and Dragomir R. Radev (1995), Hietsugu NANBA and Manabu OKUMURA (1999)]. As we know, especially on technical areas, most of papers have a very well written abstract with them. So, such a extract-based summarizer for single article doesn't lead to any economy of time for the user. On the other hand, a reader wants to seize the recent development and new achievement on his specific research field but have no time to read all the published papers. In fact, it is impossible for a reader to collect and read all papers published in his interesting areas in this information overload time. In this situation, summarizing multiple articles or survey of a specific domain can make it easier for user to grasp the outlines of the domain.

In this paper, we try to investigate the problem of summarization on a set of articles. We use a corpus on the Neural Network Algorithms collected from Journals and conferences papers in Chinese in which 89 papers are included. Firstly, We will analysis and describe some characteristics of a human expert on writing a survey for a specific domain in section 2. Section 3 will present our methodology on summarizing multi-paper in details and explain the corpus used in the study. Section 4 is concerned with the system realization and algorithms. And finally is the conclusion and discussion.

2 Essentials of Summarizing Multi-paper by Human Expert

It is well known that a survey paper whether on Magazine and Journals or on the academic conference is often written by one who has completed a lot of research work and has enough knowledge in the filed, or who at least has read many papers in the domain. We call this kind of person human expert in contradistinction to automatic machine summarizer. After considering the behavior on writing a survey paper of a human expert on his own research field, It is not difficult for us to outline his cognitive activity or procedure in this task as the following.

(1) Careful consideration of the topic

Who will be the reader of this paper?
Did someone else write the same paper before?
What is the structure of the paper?
Is there any new development achieved recently?
…….

(2)Retrieval of papers related to the specific domain and construction of corpus

Maybe there happened to have many related papers on his hand. However he also needs to index the papers especially recently published papers on the area.

(3)Reading carefully and understanding each of the selected papers

This may be a time-consuming work. He needs to carefully perceive every viewpoint of the authors as well as the correctness of the experiment result. He needs also to comprehend or grasp the main idea of each paper.

(4)Extraction of information from papers of the domain

Detection of important fragments from each paper
Recognition of similarity among papers
Discovery of difference among papers

(5)*Generating new ideas and viewpoint of his own after fusing together various information from each paper*

Outline of the new approaches and methodology in the domain

Review of the main achievements of the domain

Supposition of the future research direction of the domain

High point of the existed problem of the current researches

(6)*Accomplishment of the final survey paper based on the above study*

Through the overview of human expert based survey writing activities, we can conclude that: summarization of multiple papers should be much more difficult than summarizing only single article. Firstly, we should take into account how to collect the target papers for summarization, which corresponds to (2) of the above human based activity. Secondly, a multi-paper summary system should clearly describe the similarity and difference among papers, or it should be able to extract useful information exactly from papers like human expert does in the step (4) of the above procedure. Furthermore, it is desirable that the output of a multi-paper summarization system should describe or generate new ideas-based sentences or text which cannot be extracted directly from the original papers corresponding to the step (5). However, the task to generate such a survey automatically seems very difficult and impossible at present.

3 Corpus and Methodology

3.1 Corpus Description

Summarization research is notorious for its lack of adequate corpora [Daniel Marcu (1999)]. Today, there exist only a few small collections of texts whose units have been manually annotated for textual importance. Not only the English text but also the other languages face the problem. This phenomenon is also the trouble of the whole Natural Language Processing research and has prevented rapid progress in the field. In order to conduct study toward multi-paper summarization, We use the corpus created by the Research Center of Intelligence, Beijing University of Posts and Telecommunications, as our main materials. This corpus consists of a collection of Neural Network Learning Algorithms articles in Chinese, which were selected from the Chinese Research Journals and Conference Proceedings in recent years [BUPT Technical Report (1999)]. There are totally 89 papers on the corpus. The average length of the texts has about 3,000 Chinese characters. The longest one had about 6,300 Chinese characters. The shortest one has only about 500 Chinese characters. The corpus was manually marked with some symbols at the head of the line such as T, T1, N, U, A1, A2, K1, /P, J etc.. T represents Title, N represents the Name of the author, A represents the Abstract of the paper given by the author and so on.

3.2 Methodology

All researches on the automatic generation of generic abstract assume that the first task a summarization system needs to perform is that of extracting the most important units in a text [Daniel Marcu (1997), Dragomir R. Radev. (1996)]. These units can be words, phrasal expressions, clauses, sentence, fragment or paragraphs. Determining the salient parts is considered to be achievable because one or more of the following assumption hold: (1)Important sentences in a text contain words that are used frequently [Luhn, 1958] (2)Important sentences contain words that are used in the title and section headings. (3)Important sentences are located at the beginning or end of paragraphs (4)Important sentences use cue word such as "greatest" and "significant" or indicator phrase such as "the main aim of this paper" and "in this paper we propose…". (5)Important sentences are located at the location of reference areas[Hietsugu NANBA and Manabu OKUMURA (1999)].

Most current summarization systems use one or several above techniques to extract the so-call important part based on the powerful computer and then simply aggregating them together to output as the summarization of the system. In our multiple text summarization research, we try to utilize *Reference Information Sentences* of each paper as the important part to be extracted, which was firstly applied for multiple English texts summarization by [Hietsugu NANBA and Manabu OKUMURA (1999)] and seems very successful on the task. This is based on the assumption that Reference

Information Sentences contain some information of the similarity and difference between the paper and referred papers. Understanding the essence and difference among the collected papers of the specific domain would be very important to generate a survey of multiple papers according to the observation of human expert-based survey activity in section 2. Although other information such as the title, keywords etc. may be also very important to the summary. Therefore firstly, we browse the text from the beginning to the end to look for the referred position, and extract fragments of the paper where the author describes the essence of referred paper and the difference with his paper. Then with the information of reference areas, we can generate multiple papers summarization through categorizing the types of reference relationships which will be explained in the following.

Consider the following text in the corpus:

```
F:5.txt
T1:BP 悚? 媢灋岦? 嶼輮揑尒?
N丗? 憧涼
U丗峲揤塇岦岺彁尋嬞強丆杒嫗丆                    100094
/p (1) 敧廫擭戫拞堅丆        Rumelhart[3]搼 採弍  梮 堦? 強
? 婇峇? 攄   (Back Propagation ? ? ?  BP)揑懚? 恖
岺悎? 悚?    (ANN)岦? 嶼輮丆搴? 揑       姓檏  梮 懥? 悚?
拞? ? 偵? 悘? 揑岦? ? ? 丆        嵟暈  梮 憗堅揑? ? 姶柄
婍 (Perceptron)[4]廖擻? 峴? 偞暘? 揑婓艦偞丆湵
?  ANN ? 捬孾忋揑堦椞栺彁旞            [1,2]丅(2)扐惀  丆 恖?
嵯尋嬞檻? 椌拞         ? ? 丆 BP 悚? 栫梌灋晄慞擵? 丆旇
擛懚嵯嵊暘? 彨丆岦? 懍搙枅搼? ? 丅尒憧                   BP 悚?
嶼輮栟灋惀岦? 懍搙涬? 廰棷揑? ? 丅                (3)
Rumelhart[5]搼恖椃? 搨搙壛忋? 偞? 棃? 彇戫楯斊
悢婄梡忋揑嵿? 暘検婋栖梮岦? ? 彁拞揑悅? 丆棡帶
壛懍潤? 丟   Le Cunn[6]憖椃娻? 曽輮惛? ? 惍? 悘?
埲倈婖朌嵄嵊暘? 彫棃? 擾崡栙灣揑岦? 丟              Lehman[7]
搼恖椃壛殑惡揑曽輮棃援弍嵊暘? 彫丄壛懍潤? 丟
Storetta[8]搼恖孾椃惷媈惓憗斂悢? 偵棃尒憧岦?
懍搙丟   Caillon[9]搼拞? 檻惂? 悘? ? 鍮? 棃婋栖岦
? 拞揑悅? 丆壛懍潤? 丆搼搼丅              (4)? 傲 ? 嵄尒? 嶼
輮嬒嵯晄攦彁搙忋尒憧梮              BP 悚? 揑潤? 偞擻丆        扐 懥
悢嶼輮揑岦? 懍搙檻潤? 偞? ? 悘? 弶? 揑? 庢丄?
擷? 杮揑揺偞滊晎姶丆? ? 偞晄慞? 丆椦嵄嶼輮柋榬
鷺壛梮惐揑? ? 嶲悢丆? 棃嶼輮忋惐揑嶲悢? 墛?
? 丅(5)杮夓  嵯?   BP 悚? ? 峴嶼輮尋嬞揑雄? 忋丆          採
弍  嵜惱孮? ? 嵎棊? ? ? ? 岦? 椪檻? 椣塶孾揑? 椞
夢憧   BP 悚?  (FBP)岦? 嶼輮丆湀? 灋? 寁偞
(Robustness)? 峴梮暘慞旇? 丅
```

Mark '/p' means the beginning of a new paragraph which was created by the maker of the corpus. In order to explain our ideas, the bold Mark (1) **to (5) a**re created to represent the fragment of the text by ourselves. Now from the text, We can obtain the following information: Fragment (1) introduces the theme of the referred papers [1~4]. Fragment (2) points out the problems of this research field. Fragment (3) Describes the methods to solve the problem by another group of referred papers [5~9]. Fragment (4) further points out the new problem to solve the problem. Fragment (5) describes the authors' own method copes with the problem pointed out in Fragment (2)**.**

By reading the fragment from (1) **to(5),**we can understand the relationship between this paper and the referred papers[1~9].We call this kind of fragment in the text *Reference Area*. With the information in it, we can also identify the reason for citation by authors. We classify the reason for citation into the following three categories, we call these categories *Reference Types*.

¦ The Similarity-type reference**:**
The references to base on other researchers' theories.

¦ The Difference-type reference**:**
The reference to compare with related work or to point out their problems.

¦ The Other-type reference
Refer to the reference other than the Similarity-type and the Difference-type

From the above *Reference Type* classification, it is clear that the Similarity-type reference usually represents the information of similarity (method or viewpoint) between the paper and the referred paper. While the Difference-type reference describes the information of difference between the paper and the referred paper. We think the reference of the Difference-type is more important than other, because from reference areas of the Difference-type, we can obtain the following information:

(A) Introduction of previous research
(B) Description about the problem of the previous research
(C) The purpose of the authors' research

In case of the above example, Fragment (1) **and (3)** corresponds to (A), Fragment (2) **and (4) corresponds to(B)**. And Fragment (5) **corresponds to (C)**. (A) can be considered as a kind of summary of the referred paper from the authors' viewpoint. (A) can also be regarded as a fragment that describes the similarity of research topics between his paper and others. On the other hand, the problem of previous work and the purpose of research are described in Fragment (2)**,(4)** and (5). These fragments can be regarded to describe difference topics between his paper and others. So, It would be the important information for generating the multi-paper summarization.

4. Multi-paper Summarization by Extracting Reference Information

As described in the above section, If we assume the survey of multiple papers as Detection and extraction of Similarity and Differences among papers. Therefore, extracting and displaying reference areas can be a good support for writing a survey. In this section, we describe our method to realize the multi-paper summarization system.

4.1 Extraction of Reference Area

Reference areas can be considered as a set of sentences (or fragment) which have a connection with the sentence including citation in the paragraph, for example, fragment ① to ⑤ in the above text. Beside the obvious citation information, there are also some cue words to indicate such a connection between sentences. Those cue words are considered helpful for reference area extraction. Examples of cue words are shown below:

1st person pronoun	変? ，嶌戄 …
3rd person pronoun	憤? …
Indefinite pronoun	? 嵊 ,悤? …
Adverb	? 斐 ,? ,庡椱 ,? 堦? …
Verb	採弍 ,晒柂 ,揑撻 …
Special Noun	柿曖 ，嶉柿曖 …
Negative expression	扐惀 ,扐 ,? 慠 …扐惀 …

4.2 Identification of the Reference Type

In a reference area, if a negative expression appears at the beginning of the sentence including citation, the reference area can be considered as the Difference-type. Similarly, if the expression like "柿曖嶉 …媸? 忋 " appears in the sentence including citation, the reference area can be considered as the Similarity-type. Therefore we prepare a list of cue words and make a set of rules for the reference type identification.

The flow chart of reference area extraction is shown in the following.

Step 1: Input a text from corpus
Step2: Read one paragraph from the text
Step3: Check the sentence one by one and examine whether it contains the citation information
Step4: If citation mark is found, apply the reference area extraction rules to determine the fragment
Step5: Identify the type of the reference area.
Step6: If the end of the paragraph is met, go to step 2.
Step7: If the end of the file is met, go to step 1.
Step8: Combine the whole fragment and form a text file to output.
Step 9: End.

5. Conclusion

The ability to automatically provide summarization of multiple textual materials in a specific domain will critically aid in effective use of the Internet in order to avoid overload of information. In this paper, we proposed a prototype system framework to perform the summarization of multi-paper on the technical domain by extracting the reference information of each paper in the corpus. This work is still at early stage. The whole system hasn't been completely implemented yet. However, This seed work will allow us to conduct further study toward automatic multi-paper summarization.

References

Luhn,H.P (1958). The Automatic Creation of Literature Abstract. IBM Journal of Research and Development, 2(2):159-165,1958.

Kathleen McKeown and Dragomir R. Radev (1995), Generating Summaries of Multiple News Articles. In Proceeding of 18th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.74-82,Washington.

Dragomir R. Radev. (1996). An Architecture for Distributed Natural Language Summarization. In Proceedings of the 8th International Workshop on Natural Language Generation, pp.45-48, England.

Daniel Marcu (1997). From Discourse Structures to Text Summaries. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent scalable Text Summarization, pp.82-88.

Hietsugu NANBA and Manabu OKUMURA (1999), Towards Multi-paper Summarization Using Reference In formation. The International Joint Conference on Artificial Intelligence, pp.926-931.

Eduard Hovy and ChinYew Lin (1999), Automated Text Summarization in SUMMARIST. In Inderject Mani and Mark Maybury, eds, Advances in Automatic Text

Summarization, The MIT Press.

Daniel Marcu (1999). The automatic construction of large-scale corpora for summarization research. The 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pages 137-144, Berkeley, CA, August.

BUPT Technical Report (1999), Research on the Key Problems of Some Intelligent Services Provided by Computer Intelligent Network, the China Hi-Tech. Program 863-317-9601-06-03.