Summarization of Multiple Chinese Technical Articles

Minghui WANG and Hediheko TANAKA
Department of Engineering, The University of Tokyo
Tokyo, Japan

Abstract

The automatic summarizer is considered as a kind of software system that can produce a condensed representation of the content of any document submitted to it. The research on document summarization recently has been becoming attractive and important because of the explosive increase of the scientific information. One of the major problems of the existing summarization systems is that it only summarizes a single article at a time. In this paper, we try to develop a system framework that focuses on performing a summarization of multiple Chinese papers on a specific domain that is considered more useful for a reader to grasp the outline of a research domain.

Firstly, We will investigate and describe some characteristics of a human expert on writing a survey for a specific domain. By considering the behavior on writing a summarization for a set of technical papers by a human expert, we can sketch out his cognitive activity or procedure on this task as in the following. (1) Careful consideration of the topic. (2) Retrieval of papers related to the specific domain and corpus construction. (3)Reading carefully and understanding each of the selected papers. (4)Extraction of information from papers of the domain. (5) Generating new ideas and viewpoint of his own after fusing various information together from each paper. (6) Accomplishment of the final multi-paper summarization based on the above study. 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. 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 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 the 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. 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 on the corpus and so on.

The first task a summarization system needs to perform is that of extracting the most important units in a text. These units can be words, phrasal expressions, clauses, sentence, fragment or paragraphs. . 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 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. 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 relationship. At present, there are three Reference Type classifications. (1) The Similarity-type reference means that the citation is to base on other researchers' theories. (2) The Difference-type reference represents that the reference to compare with related work or to point out their problems. (3) The Other-type reference refers to the reference other than the Similarity-type and the Difference-type. Considering the linguistic characteristics of the Chinese text, there are also some cue words to indicate such a connection between sentences beside the obvious citation information. We have selected some cue words to help the extraction for the reference area.

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 and it is also difficult to develop a tool to evaluate the method reasonable or workable at present. However, This seed work will allow us to conduct further study toward automatic multi-paper summarization.