Automatic Summarization of Academic Paper based on Segment Structure

WONHA SHIN (1510026)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 10, 2017

Keywords: Automatic Summarization, Extractive Summarization, Single Document Summarization, Support Vector Machine.

In general, a survey of previous research is necessary to explore research trends or open up a new research field. However, it requires much labor, since we have to read a lot of academic papers. An abstract, which is written in the beginning of the most of the papers, is useful to reduce the labor of the survey. We read not all text in the paper but the abstract only to check if the paper is essential for the survey. After choosing essential papers, we read all text to know the details of them. However, the abstract in the paper is often too concise to know all necessary information for the survey. For example, the abstract may not contain discussion about difference between the paper and its related work, or the detail results of an experiment. If we want to know the above information, we should read all text.

A goal of this thesis is to propose a method of automatic summarization of the academic papers to reduce the labor of the research survey. We aims at automatically generating a long or extended abstract that contains all necessary information for the survey, such as a goal, novelty or characteristics compared to related work, overview of the proposed method, the evaluation of the proposed method and so on. Hereafter, we call such an abstract "comprehensive summary". The proposed method is a kind of a single document extractive summarization. Although several attempts have been devoted to summarization of the single academic paper, they did not consider a structure or segment of the paper. In the proposed method, the paper is divided into the segments first. For each segment, important sentences are extracted by a method that is optimized to the segment. Finally, the extracted sentences are merged to produce the comprehensive summary.

Detail procedures of the proposed method is as follows. In this study, we suppose that a paper in LaTeXformat is given as an input of the system. First, a given paper is divided into the segments. Considering the typical structure of the academic papers, five types of the segments are defined: "introduction", "related work", "proposed method", "evaluation" and "conclusion". Two methods for the segmentation are proposed. One is a method based on the section title, the other is a method based on cue phrases of related work. In the first method, a set of keywords is prepared for each segment, then a section whose title contains the keyword is extracted as the segment. In this method, keywords

for the "proposed method" segment are not made, since various expressions are used in the title of the section that explains the proposed method. Therefore, the four segments except for "proposed method" are extracted by keyword matching, then the rest of the sections are extracted as the "proposed method" segment. In our preliminary experiment, many "related work" segments are failed to be extracted. The second method based on cue phrases of related work is introduced to extract "related work" segments more. First, the paper is divided into the paragraphs. Next, a list of cue phrases, which represent a typical expression found in the related work section, is manually prepared. Finally, a paragraph that contains the cue phrases and its preceding two paragraphs are extracted as the "related work" segment.

Next, methods to extract important sentences from each segment are proposed. It is expected that different types of the sentences are important in the different types of the segment. For example, a sentence that describes the goal and contribution of the paper is the important sentence in the "introduction" segment. A sentence that explains the difference between the paper and the previous work is the important sentence in the "related work" segment. A sentence that explains the outline of the proposed method and a figure that shows the overview of the proposed method are important in the "proposed method" segment. A sentence that describes an experimental setup and a figure/table that reports the results of the experiment is important in the "evaluation" segment. Therefore, different methods are designed for individual types of the segments. That is, each method is optimized to extract important sentences in one type of the segment.

To extract the important sentences from the "introduction" segment, a binary classifier that judges if a sentence is important or not is trained by Support Vector Machine (SVM). Since the important sentences in the "introduction section" may also appear in the abstract of the paper, a training data is automatically constructed as follows. For each sentence in the "introduction" segment, similarity between it and the sentence in the abstract is calculated by measuring the overlap of 3-gram of words in them. If the similarity is high, the sentences is regarded as the important sentence. After constructing the training data, SVM is trained from it. Features used for training are n-gram of words where n is 1, 2 or 3. A simple feature selection is applied, that is, the features whose frequency in the training data is one are removed.

To extract the important sentences from the "related work" segment, a score of each sentence is calculated, then the highly ranked sentences are extracted and added to the summary. The score is defined as follows. More cue phrases of related work the sentence contains, the higher the score of the sentence is. The greater the sum of TF-IDF scores of the words in the sentence is, the higher the score of the sentence is.

An experiment to evaluate the proposed method is conducted. LaTeX corpus of The Association for Natural Language Processing is used for the experiment. Thirty papers are used for the test data, and 388 papers are used for the training or development data. First, the segmentation methods are evaluated. The precision of the method based on the section title is 100% except for the "proposed method" segment (that is 83%). While the recall is 62 to 100%. The recall for the "related work" segment is worse. As for the method based on cue phrases of related work, the accuracy of it is 65%. Anyway, the method can contribute to extract more "related work" segments.

Next, the method to extract the important sentences are evaluated. The method for

the "introduction" segment can choose the important sentences with 30% precision, recall and F-measure. The precision, recall, F-measure of the method for the "related work" segment is 21%, 24% and 22%, respectively. Furthermore, the method can extract more important sentences from the "related work" segments identified by the method based on section title than by the method of cue phrase of related work.

Future work of this study is summarized as follows. First, methods to extract the important sentences from the "proposed method" and "evaluation" segment are not implemented yet. A method to combine the important sentences extracted from the segments and make a comprehensive survey is not, either. Exploration of these methods are urgent. According to the results of the experiments, there is much room to improve the methods to extract the important sentences. Furthermore, the subjective evaluation of the proposed system is necessary. It will evaluate how the comprehensive summary generated by the proposed method is effective for the survey.