The summarization of a lecture sound by deletion or paraphrasing of an unnecessary part

Takashi Habata

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

February 15, 2001

Keywords: captioning, sentence reduction type summarization, spoken language.

The work which characterizes a sound like the closed captions is important from the viewpoint of hearing-impaired person support. In that case, it is necessary to characterize a sound as it is, but to summarize to suitable length. But, Since the result of a summary is what tells information as a substitute of a sound, it is necessary to avoid lack of the information by the summary as much as possible. On the other hand, in the case of a method of selecting important sentences, information may be mostly missing.

In the case of spoken language like a lecture sound, many redundant expression if included in one sentence. This redundant expression is an unnecessary part from the viewpoint of communication of information. Therefore, it is considered that it can summarize without lack of information by deleting such expression as an unnecessary part.

There is a captioning as an example of the sentence reduction type summarization for spoken language. This is activity which summarizes on real time and tells the contents of lecture sound by the handwriting character or the keyboard input. This activity is very important from the viewpoint of hearing-impaired person support. It is very useful to develop an automatic captioning system. Moreover, the captioning is good model of the sentence reduction type summarization for spoken language.

The purpose of this research is a development of sentence reduction type summary system for writen text of a lecture sound by deletion of the unnecessary part in a sentence. This system is the first step of the automatic captioning system.

Spoken Language has many characteristic expression. From these expression, Filler, Restatement, Repetition, Parenthesis and Polite expression is considered as redundant expression.

It investigated how these expression would be processed in the captioning. Filler can delete from the result of morphological analysis. Restatement and Repetition can delete

the degree of similar between a clauses and the morpheme which constitutes a clause as conditions. Although Parenthesis has not discovered the effective method of deletion, it was able to delete, when it had expression in the end of a phrase like "youyouni," which means "as" in English. In the research of closed caption generation, the techniquw of paraphrasing Polite expression is proposed. However, the technique paraphrase into an expression peculiar to a closed caption. Therefore, in the this reserch, the technique of paraphrasing was created.

Moreover, as a result of investigating a captioning, "toiu" which means "called", "of", etc. in English was also deleted as an unnecessary part. This expression was able to determine the deletion method from the prat-of-speech relation of the morpheme which appears befor and after "toiu". However, exceptional deletion method was applied by the kind of word after "toiu".

Based on these results of an investigation, the module which deletes or paraphrases each expression was created individually. The sentence reduction type summarization system is created by combining each of these module.

As evaluation of summary system, the rate of deletion, precision and recall were used. Then, the input of a system was the writen text of a lecture sound and correct answer data was the captioning data. The rate of deletion was about 20% from 15% and was influenced by a lecture presenter's individual variation. Three stages of perfect agreement, partial agreement and the correct answer in perfect disagreement estimated the precision of this system. The evaluation by the correct answer in perfect disagreement considers the plausible evaluation of this system, and the result was about 80%. Two stages of perfect and partial agreement estimated the recall of this system. The evaluation by partial agreement was about 50%.

As the future works, there is improvement of the parenthesis deletion module. At the present, since a deletion is a part of parenthetical expression which appears to lecture sound, the rate of deletion is low. Next, Restatement and Repetition deletion module also need to be improved as an improvement of precision. This module is performing the degree calculation of similar between clauses from character sequence matching of a clause composition morpheme. Howerve, it will be necessary to consider the degree of semantic similar of a content word. About the recall, it is necessary to investigate many captioning data. However, in this research, informative summarization is the purpose. Therefore, it is not necessary to improve recall until it reduces precision.