Automatic Summarization Focusing on Document Genre and Text Structure

Yohei Seki†,‡
Department of Informatics,
The Graduate University for Advanced Studies (Sokendai)†
National Institute of Informatics (NII)‡
Tokyo 101-8430, Japan
seki@nii.ac.jp
Advisor: Noriko Kando

This dissertation proposes a new automatic summarization method focusing on *document genre* and *text structure*, and verifies its effectiveness. "*Document genre*" refers to the type of document, such as a diary or a report. "*Text structure*" refers to the functional aspects of the text and divides the text into sentence units or components, according to their functional roles. This type of structure includes both the components and their organization within the text of a specific document genre. We used text structure and document genre to extract important sentences from source documents and to generate output summaries.

To date, automatic summarization research has focused principally on the topic of the document, using term frequency as a clue. The problem with this approach is that the resulting output summary lacks coherence and balance and does not differentiate between types of information. For example, sometimes the user is looking for facts concerning a certain topic, while at other times the user is interested in opinions or general discussion.

Our proposed method allows the user to search for specific types of information (for example, opinion, fact or encyclopedic knowledge). Therefore, this proposed method produces summaries according to the type of information specified by the user as well as the topics of the documents. Text structure is also effective in producing more balanced and coherent output summaries. The three main aspects of the problem in this dissertation are as follows:

A. Extracting balanced contents of the source documents.

B. Summarization to discriminate between types of information (fact, opinion, and knowledge) that the user's desire to know.

C. Generating output summaries to improve the readability and reduce redundancy.

We used text structure and document genre to extract the important sentences from the source documents in *A* and *B*, while we used text structure of output summaries to produce summaries in *C*.

A. **Single-Document Summarization Using Text Structure: Summarization for a Specific Document Genre, Newspaper Editorial**
As a preliminary stage of the research, single document summarization of single genre (newspaper editorial) was investigated. Five sentence types were automatically annotated in the source documents: main description, elaboration, background, author's opinion and prospect. This research has suggested that these sentence types be used as clues to extract important sentences, in addition to using term frequency, etc. as clues. To balance the contents for a summary, the text structure of news editorial was

manually analyzed and found to be constructed as a repetition of a topic description and corresponding opinions. We then proposed an automatic summarization method to reflect this structure and balance the main description, elaboration, and opinion/prospect.
The proposed method was evaluated by preparing sets of questions asking for the contents of source documents and by testing whether a businessman could answer the questions using only the information contained in the summaries generated by each of the the three systems, i.e., the proposed system and two baseline systems: the lead method and the term frequency-based extraction method without text structure. Although the results did not show the statistically significant improvement but, the correct answer rates for the proposed system showed an improvement of 5.8% and 8.1% over two baseline systems. This result demonstrates the promise of summarization using text structure. In the next investigation this approach was used on a more complex problem: multi-document summarization.

B. **Multi-Document Summarization Using Document Genre and Text Structure: Summarization of Multiple Genre Documents Based on the User's Information Needs**
The user tailors the search according to his or her information needs. In this research, document genre and text structure were used to summarize documents by discriminating between various types of information on the topics that the user wishes to identify. Document genre was determined by a combination of values according to four dimensions of genre: situation dependency, argumentation, impersonal styles, and factual reporting. We used document genre to discriminate between information types within a document set. Inter-coder consistency of document genre labels between three annotators was high ($\kappa = 0.5 \sim 0.65$) and the labels were annotated automatically using support vector machine (SVM). Six sentence types were used to analyze the text structure: main description, elaboration, background, author's opinion, authority's opinion and prospect. As in *A*, the inter-coder consistency of sentence type annotation between three annotators was high ($\kappa = 0.45 \sim 0.6$).
The effectiveness of this proposal was evaluated by using a newly constructed summarization test collection, *ViewSumm30*, which containing human-made reference summaries focusing each of the three different information types: facts, opinions, or knowledge for each of the 30 document sets. As a result, the proposed system showed improvements of 5.4%, 33.6%, and 24.6% for coverage in fact reporting-type, opinion-focused, and knowledge-focused summaries, respectively, over the baseline system, which did not use text structure and genre. The result was statistically significant (Wilcoxon's signed rank test, $P < 0.05$). For extrinsic evaluation, professional editors prepared sets of questions, asking for facts, opinions, or knowledge, which were described in the source documents for each of the 30 document sets. We then tested whether these questions could be answered using only the information contained in the summaries generated by the two systems: a baseline system that did not discriminate between information types and the proposed system. The correct answer rates were higher for the summaries produced by the proposed system and the result was statistically significant (Wilcoxon's signed rank test, $P < 0.05$). This showed that the proposed method produced relevant summaries consistent with the user's information needs.

C. **Summary Generation Using Text Structure: Generating Reports for Specific Document Genre with the Three-stage Model**
To solve the problem of the extracting method not producing coherent summaries, we focued on the text structure of the output summaries. This method allowed us to generate concise reports that reduce redundant information originating at different times or from different sources.
In this research, text structure was used to summarize numeric data and generate concise reports in natural language on a specific document genre, such as weather forecasts etc. This structure was represented using XML.
The weather reports generated by this method were evaluated by comparing them with human-generated natural language weather forecasts for the same area and time. For 22% of generated documents, more than three items of the triplets for weather, time, and location were found to agree. We

implemented a lexical paraphrase function by considering their context, and an aggregation function by collecting similar types of information in coherent sentences. By these functions, the original data was aggregated to $10 \sim 20\%$ size of it and this improved the readability of the output summaries.

In this dissertation, the author's principal aim is to verify the effectiveness of document genre and text structure for automatic summarization. Summarization methods according to the components of document genre and text structure were evaluated and the results showed their effectiveness.