Abstract

When researchers and students start a new research at the first step, they find some surveys and concentrate on the novelty of the research they want to work on. It is necessary to understand the elements and contents of research through a large amount of related-work surveys to get the information of the state-of-the-art technology. With the spread of the Internet, automatic summarization is a technology for grasping important information from massive data. Automatic summarization is a research project that automatically generates a short document that briefly describes the contents of a given document.

In recent years, a vast amount of academic papers have been to online and open resources. Collecting the essential information from them becomes an essential step in the initial stage of research activities. The contents in an academic paper reflect the viewpoint such as the background, purpose, method, experiment, evaluation, and conclusion. Catching the contents that reflects the viewpoints and recognize the critical sentences in the contents of each viewpoint can improve the effectiveness of research activities.

The purpose of this research is to develop a Viewpoint Refinement in Automatic Summarization (VPRAS) system for research articles that reflects the viewpoints such as the research background and purpose to support surveys by researchers and students. Since there is a limited dataset on the summary reflecting the viewpoints, we adopt machine learning techniques to classify sentences in the japanese article into the viewpoints. In addition to supervised machine learning, we introduced reinforcement learning and Dynamic Programming(DP) to extract the important sentence in each viewpoint. The agent automatically extracts summary sentences based on the reward function, to test the potentials of improving accuracy.

Extraction Summarization is regarded as a kind of document classification task. Chapter 3 introduces the method of text classification with viewpoints in our VPRAS model based on Deep-learning technology. We use the result of classification and apply reinforcement learning and DP(Dynamic Programming) to build a sentence extraction model to generate a summary. At the first step of building the dataset, we download academic articles of the Japanese language in PDF. Next, we use 'apache-tika' to recognize the texts in each PDF and make regular expressions in these texts to extract the body of text. The expert adds mainclass-label, subclass-label, and importance-label to each sentence in the main documents. The mainclass-labels are used in text classification by deep learning. The subclass-labels and importance-labels are used in text extraction. At the step of the Deep-learning model, we adopt the two methods of pre-training called Word2vec and PV-DM(Distributed Memory Model of Paragraph Vector) to

execute word-embedding which is one of the simplest deep learning techniques to build features that represent words, sentences, and documents. After acquiring the word-embedding vector in pre-training methods, Embedded words and sentences are inputted in the neutral network.

In the neutral network, we use Word2vec which reflects the feature of words as the input a classifier called LSTM(Long short-term memory) to execute text classification and use another classifier called SVM(Support vector machine) to classify the sentence-vector which embedded by PV-DM. In order to improve the recognition accuracy of text classification, we propose a combined-method that combines the advantage of Word2vec+LSTM and PV-DM+SVM. In combined-method, we acquire the result of each classifier to get the probability of each class and optimize these probabilities to do reclassification. In the classification by deep learning, there is a possibility that the error function does not decrease during the training process because of the different fields in the training article and test article. To solve this kind of problem, we adopt a function that configures with the probability of each class and cosine similarity to reclassification once again. We use the result of the final round of classification as the target sentences in the important sentence extraction model. At the step of text extraction, we calculate the value of each sentence by two methods. One is dependent on the importance-label, another one is dependent on both importance-label and subclass-label. Then, we calculate the cosine similarity between each sentence as a penalty to reduce information redundancy when extracting summary. Finally, we input the value of the sentence, the length of the sentence, the limited length of the summary, and the penalty of similarity into the knapsack-reinforcement learning model to extract the summary.

In the experiment of chapter 4, we conducted the simulation about the deep learning model with the pre-training method Word2vec and PV-DM. We also tested the effectiveness of the combined-method and reward function based on cosine similarity to verify our model's accuracy. In the experiment of reinforcement learning and DP, we also added the comparison model which only used the ranking of the value of sentences. In the part of the evaluation, we tested the recognition accuracy in importance-label, which were added by an expert, and calculated the Rouge-score of each summary. Finally, according to the result of the experiment, we discussed the feature of each method in our model and made an error analysis of them.

In chapter.5 of conclusion, we conclude what we did in this research and give the suggestion about how to revise our model to make a better recognition in the future work.

**Keywords:** Academic Paper, Automatic Summarization, ViewPoint, Deep Learning, Reinforcement Learning,Dynamic Programming