Supporting the construction of mystery novel knowledge graphs using BERT summarization

Kazuma Hasegawa
Department of Applied Informatics, Faculty of Science and Engineering
Hosei University
Japan
kazuma.hasegawa.5w@stu.hosei.ac.jp

Akihiro FUJII
Hosei University
Japan
fujii@hosei.ac.jp

Abstract*—In the Knowledge Graph Reasoning Challenge, the plot of a mystery novel is converted into a knowledge graph. Building the graph requires the work of choosing the parts to put into the knowledge graph. This research aims to automate this task. Specifically, BERT, a Natural Language Processing model, is used to summarize the entered text, with the results proposed as the parts to put into the knowledge graph. Casting summarization as a classification problem, a classification model was created. Its accuracy was found to have an F-value of 0.59. Given the likelihood of improving the accuracy, it is thought that organizing the dataset could have a significant impact. This is to be pursued as a next step.(abstract***)***

Keywords—BERT, summarization, Natural Language Processing (key words***)***

I. INTRODUCTION

The Knowledge Graph Reasoning Challenge is held by the Japanese Society for Artificial Intelligence Semantic Web and Ontology Working Group. For this challenge, a mystery novel (from the Sherlock Holmes series) was divided into scenes, with the description of each scene and the relationships among them converted to a publicly available knowledge graph. Participants then used these to identify the killer using a variety of methods.

In order to convert a mystery novel to a knowledge graph, the parts to be entered into the graph are extracted and summarized into short passages. Currently, all of the work is done manually. As this takes a lot of time, automating it would be helpful.

In this study, we focused on automating the task of extracting parts from the mystery novel to the knowledge graph. Using BERT, which has achieved state of the art performance across a range of tasks, we established a text summarization model. By applying this model, we aim to be able to automate the conversion of mystery novels to knowledge graphs in future.

II. METHODS

Initially, we hypothesized that the parts that would be converted to the knowledge graph would be those that were the most meaningful in the novel. We defined meaningful scenes as those with an impact on the story: those that dealt with the emotions of the characters in the story, their actions, character, and the circumstances of the crime and location.

Text summarization was selected as the method for extracting the important scenes. In this study, we applied exttractive summarization in practice and came up with a model for proposing extracts to be converted to a knowledge graph as a summary.

A. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a Natural Language Processing model that uses a Transformer able to generate vector representations of words taking the context into account, as well as vector representations of overall text. Using attention, it learns which inputs to focus on. BERT can be specialized to a specified task by pretraining it with linguistic characteristics and fine-tuning.

For the purposes of this research, the Japanese-language BERT model made available by Tohoku University is used, a model pretrained on the available Huggingface Transformers library. A model was created of summarization as a classification problem - of whether a given sentence is to be classed as summary or not (non-summary). From BERT output, the model obtained a 768-dimension sentence vector representation. This was then converted to a 2-D vectors by inputting this vector representation into the fully-connected layer, classified with either summary or non-summary values.

B. Dataset

The eight Sherlock Holmes stories publicly available as knowledge graphs are *The Adventure of the Speckled Band*, *The Adventure of the Dancing Men*, *The Adventure of the Devil's Foot*, *A Case of Identity*, *The Adventure of the Abbey Grange*, *The Resident Patient*, *The Adventure of the Crooked Man* and *The Adventure of Silver Blaze*. Of these, the only one that has publicly available data marking the parts of the original text transferred to the knowledge graph is *The Adventure of the Speckled Band*. Thus, this work was used as the training data in our model. As advance preparation, dashes, *furigana*, spaces and so forth were eliminated. The text was then cut into coherent 1-3 sentence paragraphs. This was to prevent the lines from becoming too brief. Label 1 (summary) was attached to sentences with marks, and Label 0 (non-summary) to those without. *The Adventure of the Dancing Men* was used as test data. This story only has data publicly available in the form of a knowledge graph. In creating the test data, labels were attached subjectively, referring to the shortened sentences used in the actual knowledge graph. Looking at the datasets, *The Adventure of the Speckled Band* had 478 pieces of data (summary: 278, non-summary: 200), while *The Adventure*

of the Dancing Men had 497 (summary: 315, non-summary: 182).

III. RESULTS

Table 1 shows the average value of the 10 experiments run from training to testing.

Table 1. Results of the experiment

Accuracy score	**Precission score**	**Recall score**	**F-score**
0.71	0.61	0.57	0.59

Accuracy is rather poor, in part due to the low volume of training data. Looking at the prediction results in detail, of those parts predicted to be in the summary, some similar expressions were found in the two works. Two examples are shown in Table 2. Both sentences in the table are summary sentences.

Table 2. Summary sentences with similar expression

The Adventure of the Speckled Band	***The Adventure of the Dancing Men***
今のわたくしではお礼も十分には致しかねますが、あとひと月ふた月のあいだに結婚して、お金を自由にできるようになりますので、そうすれば相応のお礼もできるかと思います。	それで私どもは早速、結婚の手続きをすませ、夫婦としてノーフォークに帰りました。
廊下のランプの光を浴びて、姉が戸口に現れたのですが、恐怖で真っ青になった顔、すがるような手つき、千鳥足でふらふらと進み出てくるのです	そのとき、妻の顔は気絶しそうなほどに真っ青で、手紙を読むと、それをそのまま火の中に投げ込んでしまいました。

In the first row, the sentences from the stories can be parsed as information about getting married. The second row gives us the information that a lady went pale with fright. Elsewhere, there are a number of instances of similar phrases in the two works, such as on the subjects of money and anger. Similar expressions were correctly classified as “summary.”

Meanwhile, passages about sending a telegram and shooting a gun are considered summaries in *The Adventure of the Dancing Men*, but given the absence of such expressions in *The Adventure of the Speckled Band,* they were not classified as summaries. There are likely to be differences in classification of scenes where the expressions of *The Adventure of the Speckled Band* and *The Adventure of the Dancing Men* overlap, and scenes where an expression appears in only one of the stories.

IV. DISCUSSION

Analyzing the results regarding attention, it is likely the focus was on phrases to do with marriage, money, anger and turning pale with fright. Further insight into attention should reveal the individual words focused on. This will be followed up in future research.

Due to the fact that phrases may vary depending on which story they appear in, we are likely to need to training data not just from *The Adventure of the Speckled Band* but rather, to come up with training data that covers a broad range of passages from a collection of other works as well. In future, we will see if there is any change in accuracy if we create a dataset from other works in the Sherlock Holmes series. This will also require recalibration of the model and training settings.

V. CONCLUSION

In this study, we created a model using BERT in order to automate the task of extracting the most important scenes from a mystery novel. Casting summarization as a classification problem, we conducted fine-tuning and tested the accuracy of the model. It appears that the content and volume of the dataset affected the performance. One of the main current issues in this study is the lack of understanding of the representation due to the lack of data sets.

Going forward, we will continue our research with the primary goal of increasing the dataset. In addition, we will visualize Attention and adjust the model and learning settings.

REFERENCES

[1] Sakamoto Toshiyuki. ***Tsukutte wakaru! Shizen gengo shori AI.*** C&R Kenkyujo, 2022

[2] Yang Liu, Fine-tune BERT for extractive summarization, Sep 2019, https://doi.org/10.48550/arXiv.1903.10318.