# Criminal Deduction Using Similarity Analysis Between Mystery Stories

Shotaro Hattori
Faculty of Science and Engineering
Hosei University
Tokyo, Japan
Shotaro.hattori.6z@stu.hosei.ac.jp

Akihiro Fujii
*professor of* Science and Engineering
Hosei University
Tokyo, Japan
fujii@hosei.ac.jp

***Abstract— The " Knowledge Graph Reasoning Challenge" an annual event in which participants try to infer a mystery novel using Knowledge Graph and other tools, has been attracting a lot of attention. The purpose of this research is to propose and implement a new method using BERT for this challenge. To prove this, we focused on the degree of similarity between the two novels and estimated the culprit. As a result of the analysis focusing on the words and actions of the culprit, we succeeded in extracting the culprit's name. However, the successful extraction was based on a specific condition, and further discussion is needed to pursue more precise accuracy in the future.***

***Keywords—BERT, BERTscore, Natural Language Processing, Similarity***

## I. INTRODUCTION

In the "Knowledge Graph Reasoning Challenge"[1] , various methods have been devised to estimate the culprit. However, very few studies have focused on the similarity between novels to estimate the culprit. In this study, we would like to estimate the murderer using similarity. We extract a scene that describes the words and actions of the murderer and estimate the murderer from scenes in other novels that have a high similarity to this scene.

## II. METHODS

Half a year ago, we used to calculate similarity between novels by focusing on word-to-word similarity. However, when we learned that there is a method for calculating similarity using BERT[2], a language model that can take context into account, we decided to use BERTscore[3]. For person name extraction, we decided to use spaCy[4], for which various training models are publicly available.

### *A. BERTscore*

BERTscore is an evaluation metric that uses the above language model BERT to calculate the similarity between texts using the context-aware variance representation obtained from BERT.

Specifically, first, the generated and correct texts are input to BERT to obtain a vector representation of the tokens. Next, the vector representations are used to create a cos-similarity matrix between the tokens. Finally, the maximum similarity for each token is used to compute the Precision, Recall, and F values, which are then used as the similarity score.

### *B. SpaCy*

spaCy is a natural language processing library that runs in python. The library has pre-trained statistical models and word vectors, which are useful for processing large amounts of text and analyzing text content.

In this study, we used the pre-trained model ja_core_news_md[5]. This is a standard Japanese model from spaCy, containing 480,000 word vectors and 20,000 unique vectors.

### *2-1. Preprocessing*

20 mystery novels (Sherlock Holmes series) were downloaded from Aozora Bunko. Next, the small *hiragana* or *katakana* characters printed alongside *kanji* characters, as well as illustration numbers and other noise were removed from the text of all 20 works through the program. From these 20 pre-processed works, two works were selected for actual comparison of similarity.

### *2-2. Scene Segmentation*

The number of letters in the main text of each novel was counted, and the novel was divided into scenes so that the number of letters was divided into 10 parts.

In cases where the text was not clear or the number of letters could not be neatly divided into 10 sections, adjustments were made as necessary.

### *2-3. Similarity Calculation*

In order to extract the scenes in which the criminal appears using BERTscore, we calculated the similarity by focusing on the following two conditions.

(1). The scene in which the culprit is seen for the first time.
(2). The scene where the crime was committed.

The above conditions (1) and (2) are the information to be studied in advance from one of the two novels selected from the 20 novels in 2-1.

### *2-4. Presumption of the culprit*

We used spaCy to extract the names of the people in the scenes of the other novels with the highest similarity to the scenes we focused on in 2-3.

## III. RESULTS

In this study, the novel "The Adventure of the Speckled Band" was chosen as the informational novel, and "The Adventure of the Red-Haired League" was chosen as the novel in which the culprit was presumed.

In "The Adventure of the Speckled Band ," the number of scenes corresponding to condition (1) was 2, and the number of scenes corresponding to condition (2) was 4.

According to 2-3, Figure 1 shows the results of the similarity calculation for condition (1), and Figure 2 shows the results of the similarity calculation for condition (2).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.364 | 0.334 | 0.316 | 0.309 | 0.274 | 0.397 | 0.190 | 0.253 | 0.321 | 0.358 |

Figure 1 Results of similarity calculation under condition (1)

The scene with the highest similarity to condition (1) was scene 6.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.214 | 0.194 | 0.213 | 0.342 | 0.279 | 0.061 | 0.258 | 0.149 | 0.205 | 0.223 |

Figure 2 Result of similarity calculation for condition (2)

The scene with the highest similarity to condition (2) was scene 4.

Next, Figure 3 and Figure 4 below show the results of person name extraction according to 2-3.

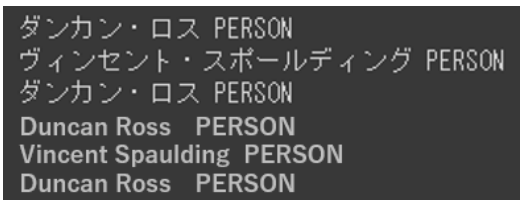


Figure 3 Results of estimating the culprit under condition (1)

Under this condition, the culprit of the "The Adventure of the Red-Haired League" could not be identified, and unrelated nouns and place names were extracted.


Figure4 Results of the estimation of the culprit under condition (2).

In this condition, both Duncan Ross and Vincent Spalding were identified as the culprits of the "The Adventure of the Red-Haired League. "

## IV. DISCUSSION

*4-1. Consideration of condition (1)*

Figure 3 shows that we did not find the culprit.
This is because the appearance of the culprit was not consistent from one work to another, so we were not able to narrow down the list of culprits, and we expected that the information would be complicated.

*4-2. Consideration of condition (2)*

Figure 4 shows that the culprit was right on target.
The scene where the crime took place describes the behavior of the perpetrator in detail. In other words, the relationships among the people in the vicinity of the murderer are described in detail.

Therefore, we can expect that scenes from other works that have a high degree of similarity to this scene are important scenes in those works, and that the culprit is likely to appear in them.This is the reason why the guess of the murderer was successful in (2).

## V. CONCLUSION

In this study, we applied similarity calculation using BERT to estimate the culprit of a mystery novel. The experimental results showed that in condition (2), the name of the murderer could be extracted. However, the scene selected based on the similarity was not the scene where the crime took place. We considered that the scene selected here was a scene in which the actions of the murderer were described in detail, and therefore, it was an important scene in the novel.

The current problem is to solve the problem of extracting unrelated nouns and place names when the names of the criminals are extracted. In "The Adventure of the Red-Haired League," no noise was detected, but in "The Adventure of the Crooked Man" noise was detected along with the name of the culprit. It is also necessary to confirm that stable results can be obtained by estimating the culprit in other works that have not been tested.

## REFERENCES

[1] Kawamura, Takahiro, et al. "Report on the first knowledge graph reasoning challenge 2018." Joint International Semantic Technology Conference. Springer, Cham, 2020.
[2] KENTON, Jacob Devlin Ming-Wei Chang; TOUTANOVA, Lee Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT. 2019. p. 4171-4186.
[3] TianyiZhang*,VarshaKishore*,FelixWu*,KilianQ.Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. International Conference on Learning Representations, 2020.
[4] spaCy · Industrial-strength Natural Language Processing in Python https://spacy.io/[2022/11/18]
[5]shttps://github.com/explosion/spacy-models/releases/download/ja_core_news_sm-3.4.0/ja_core_news_sm-3.4.0-py3-none-any.whl