NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001290

fuys Team at the NTCIR-17 UFO Task

Daigo Nishihara
Fukuoka University
Japan
td222007@cis.fukuoka-u.ac.jp

Hokuto Ototake
Fukuoka University
Japan
ototake@fukuoka-u.ac.jp

Kenji Yoshimura
Fukuoka University
Japan
yosimura@fukuoka-u.ac.jp

ABSTRACT

This paper reports the results of the fuys team's NTCIR-17 UFO Text-to-Table Relationship Extraction (TTRE). Since we thought that Value cells depend on Name cells, we came up with a method that uses the result of extracting Name cells to connect them together. The text of a HTML <mark> tag and texts of cells were used to find Name. These two were encoded and combined to perform a binary classification. We tried several combinations of mark tag text and cell text. The best results were obtained using mark tags and tables in the same section of the same company. We tried two different rules for binding Value cells. The rule of finding a cell by the row and column combination of the cell that became the Name yielded good results.

KEYWORDS

BERT, binary classification, Annual security reports

TEAM NAME

fuys

SUBTASKS

Text-to-Table Relationship Extraction (TTRE)

1 INTRODUCTION

Our fuys team participated in Text-to-Table Relationship Extraction (TTRE) task at NTCIR-17 UFO [1]. In this paper, we describe our proposed method and the results of this task.

2 TTRE

The TTRE task is to find cells in a table that are related to specific text in an annual securities report. The specific text to be linked to the table cell is marked with a HTML <mark> tag. The cells of the table related to this mark tag are bound as Name cells if the cells are item names, and Value cells if they are numeric values. The goal of this task is to link the text of the annual securities report to the table and make it easier to understand.

2.1 About Annual Securities Reports

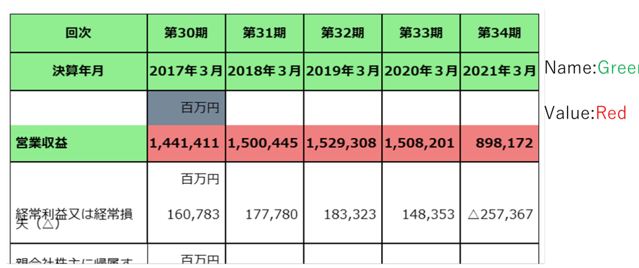
An annual securities report is a report that describes a company's sales and business activities. Since it is legally required to be submitted, there is a specific content to be written. In addition, the format is often fixed. This makes it possible to separate what is written in different sections. The content written for each section is as follows.

1. Company Overview
2. Status of Business
3. Status of Facilities
4. Status of the Submitting Company
5. Status of Accounting
6. Outline of Stock Administration of the Submitting Company

In this data, section 6, "Outline of Stock Administration of the Submitting Company," is written in a smaller volume and handled by fewer companies, so we will treat it as the same section as section 5. In addition, securities reports are converted to HTML files and can be handled as HTML.

2.2 About Name and Value

If the table's cell associated with the mark tag is an item name, it is a Name, and if it is a number, it is a Value. For example, for " 営業利益(Operating Profit)," cells colored green in the table in Figure 1 are Names. Also, the cells colored red in the table are Values.



回次	第30期	第31期	第32期	第33期	第34期
決算年月	2017年３月	2018年３月	2019年３月	2020年３月	2021年３月
	百万円				
営業収益	1,441,411	1,500,445	1,529,308	1,508,201	898,172
経常利益又は経常損失（△）	百万円 160,783	177,780	183,323	148,353	△257,367

Examples of Name and Value

Figure 1 Example of Name and Value.

3 METHOD

The author's actual reading of the distributed materials revealed the following.

- Cells that are Value are positionally related to cells that are Name.
- The cell that is the Name has characteristics in position and text.
- Most of the cells that become the Name are in the table near the MARK tag.

- Even if the companies are different, the content covered in the same section remains the same.

We believed that using these elements would increase accuracy, so we linked Name and Value.

3.1 About Name

We considered that the position of the cell that is Value depends on the position of the cell that is Name. Therefore, we used the method of connecting Name first and then using the result. We thought that by improving the accuracy of Name, the accuracy of Value would naturally improve.

Name can be thought of as a multi-label with candidate cells. However, this is not realistic because the number of labels would exceed 700. Therefore, we decided to use the text of the mark tag and the text of the table cell, encoded and combined, as the input text. We thought that by doing so, we could treat it as a simple binary classification of whether the mark tag and its table cell are related or not (Figure 2).

We performed learning and inference using BERT[2] and searched for table cells to be linked using binary classification. For the BERT pre-training model, we used the Japanese language model created by the Inui/Suzuki Laboratory at Tohoku University, cl-tohoku/bert-base-japanese-whole-word-masking. The BERT-related part was implemented by "BertForSequenceClassification," a class for document classification in Transformers, an open-source library developed by HuggingFace[1]. The model was created by performing fine tuning on a pre-trained model with the number of epochs as 10, batch size as 32, learning rate as 1e-5, and maximum input length as 512.

We used the <ix:nonnumeric> tag to find tables near the mark tag. During inference, the table in the same <ix:nonnumeric> tag as the mark tag was used to find Name (Figure 3). We looked at the cells in the table one by one and if the text was the same as the text in the MARK tag, we tied the cell's ID to that table. If not, we made inferences using cells that satisfied the following rules. The input sentence was the text of the mark tag and the text of the table cell, encoded and concatenated. This input text was used to perform a binary classification, and cells with a result of 1 were linked as Name.

- Cell is not located in row 4 or more and column 4 or more. (Figure 4)
- Cells that are not just symbols or monetary expressions.

As with inference, the training was done with a combination of the mark tag and a cell that is a candidate for the Name in the same ix:nonnumeric tag. In addition, we tried two other methods of combining mark tags and cells.

Method 1. How to combine with a table that exists within the same ix:nonnumeric tag as the mark tag (Figure 5).

Method 2. How to combine a mark tag with a table that exists in the same company and in the same section (Figure 6).

Method 3. How to combine mark tags and tables that exist in the same section of all companies (Figure 7).

Each method extracted a candidate cell for Name from the table and created an input sentence with all combinations of the text of that cell and the text of the mark tag.

The model was trained as a binary classification of 1 if the text in the mark tag is related to the cell in the table and 0 otherwise. Since sections 1, 2, 3, 4, and 5-6 each deal with the same content, we thought that creating a model for each section would increase accuracy. Therefore, we also tried to create a total of five models by dividing the input sentences in sections, with each combination.

3.2 About Value

Cells that are Value were attached on a rule basis. Since we thought that Value was dependent on the position of Name, we decided to attach it using the result of Name. Our reading of the material shows that the cells in the first row and first column of the table are rarely Values. Therefore, we decided to omit the cells in the first row and first column.

We used two rules to tie the Value to the cell. In Rule1, the cell and Value were tied together by a rule that attaches all the same rows and columns of the cell that is the Name (Figure 8). In Rule2, the cell and Value are linked by the following rule based on the location information of the cell that is Name (Figure 9).

- If Name's cell has only one row, tie all cells in the same column as Name's cell.
- If Name's cell has only one column, then tie all cells in the same row as Name's cell.
- If the number of rows and columns of cells in Name are both greater than or equal to 2, then tie the cells that can be found by combining rows and columns.

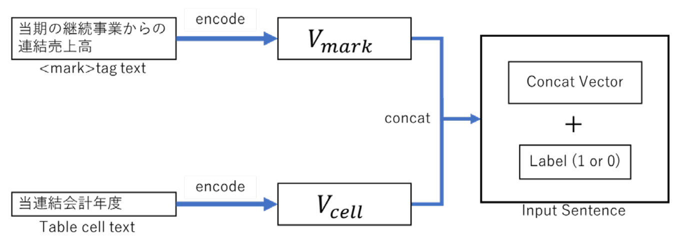


Figure 2 How to create an Input Sentence.

[1] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

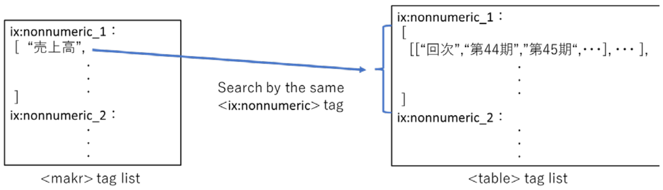


Figure 3 How to find a table.

回次		第30期	第31期	第32期	第33期	第34期
決算年月		平成29年３月	平成30年３月	平成31年３月	令和２年３月	令和３年３月
営業収益	（百万円）	1,756,980	1,822,039	1,878,137	1,844,647	823,517
経常利益又は経常損失（△）	（百万円）	563,973	583,569	632,653	574,282	△262,064
親会社株主に帰属する当期純利益又は親会社株主に帰属する当期純損失（△）	（百万円）	392,913	395,502	438,715	397,881	△201,554
包括利益	（百万円）	399,856	404,198	446,213	388,418	△165,901
純資産額	（百万円）	2,726,729	3,084,739	3,508,065	3,872,103	3,686,609
総資産額	（百万円）	7,052,675	8,908,682	9,295,745	9,603,126	9,600,370
１株当たり純資産額	（円）	13,681.22	15,602.66	17,703.74	19,514.81	18,510.87
１株当たり当期純利益又は１株当たり当期純損失（△）	（円）	1,996.52	2,015.48	2,238.95	2,027.86	△1,025.46
潜在株式調整後１株当たり当期純利益	（円）	－	－	－	－	－
自己資本比率	（%）	38.2	34.3	37.3	39.9	37.9
自己資本利益率	（%）	15.7	13.8	13.4	10.9	△5.4
株価収益率	（倍）	9.09	9.99	11.48	8.54	－
営業活動によるキャッシュ・フロー	（百万円）	580,565	609,595	600,319	595,227	△169,354
投資活動によるキャッシュ・フロー	（百万円）	△1,909,547	△1,676,489	△597,502	△552,494	△134,718
財務活動によるキャッシュ・フロー	（百万円）	1,425,188	1,434,788	△33,635	△32,993	262,638
現金及び現金同等物の期末残高	（百万円）	414,559	782,454	751,636	761,376	719,941
従業員数［外、平均臨時雇用者数］	（人）	28,593 [8,275]	28,867 [8,494]	29,128 [8,799]	29,603 [9,112]	30,153 [8,078]

Figure 4 Scope of search for Name.

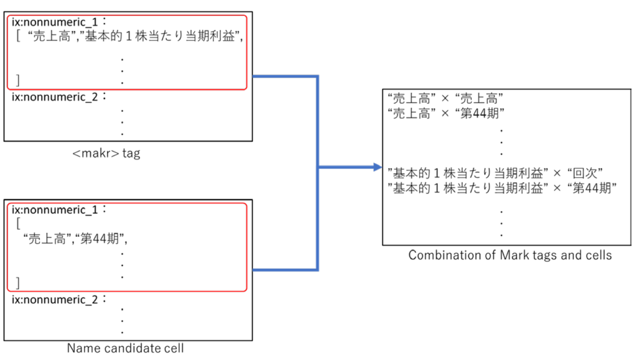


Figure 5 Example of Method 1.

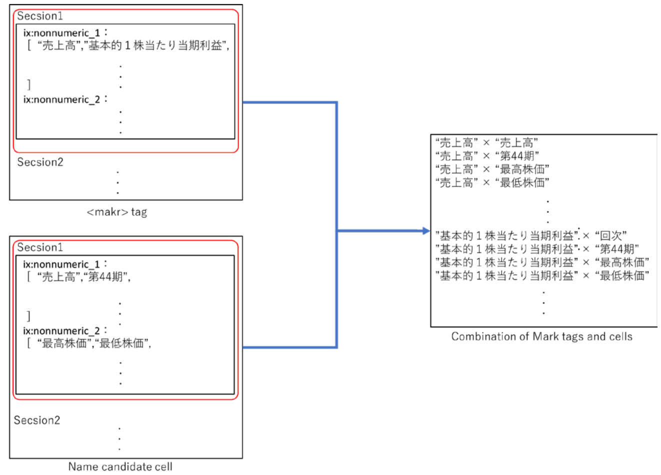


Figure 6 Example of Method 2.

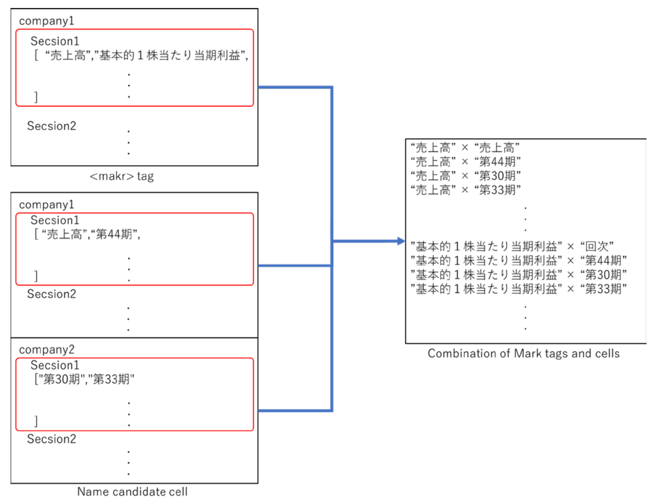


Figure 7 Example of Method 3.

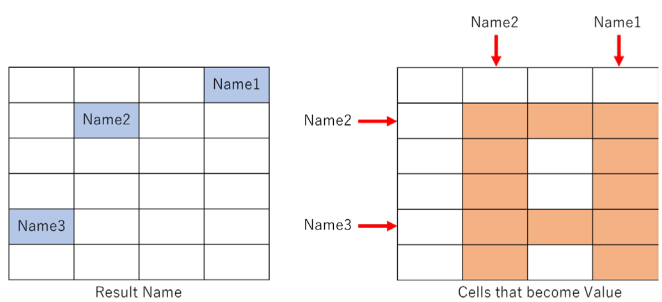


Figure 8 How to attach Rule 1.

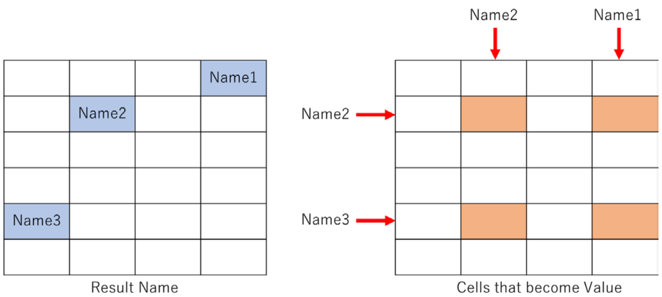


Figure 9 How to attach Rule 2.

4 RESULTS

When one model was prepared with Method 2 and Rule 2 was applied, the result of Total was 23.47%, which was the highest.

Looking at the accuracy of Name for each method, 22.50% was obtained with one model for Method 1, and 21.15% with five models (Table 1). Method 2 with one model was 27.07%, and with five models was 24.99% (Table 2). Method 3 with one model was 20.47%, and with five models was 22.88% (Table 3). As can be seen from the results, the accuracy of Name was highest when one model was prepared for Method 2.

For Value, the accuracy was higher when Rule 2 was applied. In addition, for each rule, the highest accuracy was obtained with Method 2, which has the highest Name accuracy.

Table 1 Results with Method 1.

	Name	Value (Rule1)	Total
Model 1	22.50%	17.12%	19.81%
Model 5	21.15%	16.36%	18.76%

Table 2 Results with Method 2.

	Name	Value (Rule1)	Total
Model 1	27.07%	19.43%	23.25%
Model 5	24.99%	17.81%	21.40%

Table 3 Results with Method 3.

	Name	Value (Rule1)	Total
Model 1	20.47%	15.89%	18.18%
Model 5	22.88%	16.86%	19.87%

Table 4 Results of Rule 1 and Rule 2 in Method 2.

	Name	Value	Total
Rule1	27.07%	19.43%	23.25%
Rule2	27.07%	19.88%	23.48%

5 CONSIDERATIONS

Name and Value will be considered for each.

5.1 About Name

Looking at the results for each method, the Name accuracy is highest for Method 2 with one model (Tabel 5). The highest F value is obtained with one model in Method 2, making this model the best performing one. Since Method 1 uses the tables in the same ix:nonnumeric tag for training, the number of training sessions in which the label is 0 is reduced. This is thought to have resulted in a high recall and low precision. Since Method 3 uses tables that exist in the same section for training, there are many training sessions in which the label is 0. This is thought to have resulted in a low recall and high precision.

5.1.1 Problem

Two problems were found in the way Name was tied together in this proposed method.

The first is that since we are dealing with securities reports, there are many proper nouns that appear in a single company. Looking at the actual contents of the cells in the table, there are numerous proper nouns, such as the names of products and executives. When the percentage of correct answers is given for each section, the results for Section 1 show the highest accuracy (Table 6). Section 1 contains many words that are used by all companies, such as 利益(profit)" and "資産(assets)," and fewer occurrences of proper nouns than in the other sections. This is the reason why the percentage of correct answers in Section 1 was relatively higher than in the other sections. Conversely, the appearance of proper nouns in Section 2 and later sections is more frequent, making it more difficult to find cells in the related tables.

Second, the accuracy of Name was limited. In the proposed method, candidate cells for Name were searched from tables in the same ix:nonnumeric tag as the mark tag. However, in reality, there are cells that are candidates for the name in tables outside the ix:nonnumeric tag. When the gold data was extracted from the tables inside the ix:nonnumeric tag, it was found that only 65.78% of the cells were connected at best.

5.1.2 Solution

The first solution is to replace proper nouns with different ones. By converting a person's name to "人名(person's name)," a company's name to "会社名(company name)," and so on, we would treat common proper nouns as the same thing. By doing so, we thought we could reduce the myriad of proper nouns and make them easier to deal with.

A second solution would be to search for cells that are candidates for Name across all tables in the same company and in the same section. In fact, when Name was applied to all tables in the same section, the result was 29.95% for Name. However, the result for precision is 31.73%, which is significantly lower than the 44.71% for precision when the range of tables searched is narrowed down with the ix:nonnumeric tag. Also, applying Rule 2 and tying Value to the mark tag resulted in a Value of 18.57% and a Total of 24.26%. From this, it can be seen that the results for Name and Total are higher when all tables in the same company and in the same section are targeted.

However, the precision and Value results for Name are lower. Therefore, if the number of tables in the same section increases in the future, it is possible that cells from unrelated tables will be linked, and the results of Name and Value will drop. This result indicates that there are related tables outside the ix:nonnumeric tag. Therefore, it is thought that by narrowing down the tables to look for using a different method than the ix:nonnumeric tag, it is possible to handle an increase in the number of tables in the same section and improve accuracy.

5.2 About Value

The differences between Rule 1 and Rule 2 were compared using Method 2, which was the most accurate. The results were 19.43% for Rule 1 and 19.88% for Rule 2. The result was slightly higher for Rule 2. When the percentage of correct answers for Name was set high, the results were 50.15% for Rule 1 and 50.42% for Rule 2. When the correct answer rate for Name was set at 65.78%, which is the limit value, the results were 35.37% for Rule 1 and 38.80% for Rule 2.

As can be seen from the results, Rule 2 results in a slightly higher Value result. However, there is no dramatic difference between Rule 1 and Rule 2. The reason for this result is thought to be that Rule 2 was based on Rule 1, with only minor rules added, so it is difficult to see clear differences. Both Rule 1 and Rule 2 use the position of Name. Since the cells to be looked at are based on the same cell, it is thought that there is not much difference in the results.

Table 5 Results of Name's analysis for each method.

	Name	recall	precision
Method 1 (Model 1)	22.50%	48.68%	26.49%
Method 1 (Model 5)	21.15%	44.10%	25.92%
Method 2 (Model 1)	27.07%	39.17%	44.71%
Method 2 (Model 5)	24.99%	40.28%	37.24%
Method 3 (Model 1)	20.47%	28.11%	48.17%
Method 3 (Model 5)	22.88%	33.93%	41.41%

Table 6 Results of each section in Method 2.

	Name	recall	precision
Section1	47.52%	62.99%	65.79%
Section2	40.93%	57.71%	55.34%
Section3	38.45%	53.50%	51.11%
Section4	34.90%	48.19%	51.93%
Section5-6	27.07%	39.17%	44.71%

Table 7 Result when <ix:nonnumeric> tag is not used.

	Name	Name recall	Name precision	Value	Total
Result	29.95%	58.88%	31.73%	18.57%	24.56%

6 CONCLUSIONS

To solve TTRE task, we proposed a method for each Name and Value. The method to link Name was considered as a binary classification task using <mark> tags and table cells. Three different training methods were used to create the classification model, and the best performance was achieved with an F-score of 27.07%. Value implemented a rule-based association method using the results of Name. The result was an F-score of 19.88%. The overall score was 23.48%. One problem with Name was the large number of proper nouns and the limiting value. This problem can be solved by replacing proper nouns with other words and by expanding the range of the search table.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K12740, JP21H03769.

REFERENCES

[1] Yasutomo Kimura, Hokuto Ototake, Kazuma Kadowaki, Takahito Kondo and Makoto P. Kato, “Overview of the NTCIR-17 UFO Task,” In Proceedings of the NTCIR-17 Conference, 2023. https://doi.org/10.20736/0002001321

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, NAACL-HILT, 2019.