fuys Team at the NTCIR-17
QA Lab-PoliInfo-4 Task

fuys team
Daigo Nishihara, Hokuto Ototake, Kenji Yoshimura
Fukuoka University
About MBLink

- Tasks that link the mayor's statement to a table that deals with the budget.
- The mayor's statement and the table dealing with the budget have been changed to an HTML file.
  → The mayor's statement can be treated as `<p>` tag.
  → Tables dealing with budget can be treated as `<table>` tag.

それでは、平成29年度の予算編成についてですが、収入状況は、市税の伸びが期待できないことに加え、地方譲与税や交付金、さらには実質的な地方交付税の減少が見込まれ、引き続き大変厳しい状況にあります。
Proposed Method

- Finding related tables by focusing on cells.
- Cells with only monetary amounts, unit notations, or symbols are not used.
- We'll use the <p>tag text with the ID and the table cell text.
- **Binary classification task** with the two encoded and combined as input sentences.

![Diagram showing the process of encoding and combining text and cells for classification]

Table cells are extracted in **three different methods**.
Proposed Method (Method 1)

- `<p> tag and cell-by-cell learning and inference.
- Extract words from all table cells.
- Learning is performed with
  - 1 if the word exists in the related table.
  - 0 if the word does not exist in the related table.

### Table

| 科目       | 本年度 | 前年度
|------------|--------|--------
|            | 予算額 | 構成比 | 初予算額 | 構成比 |
| 1港湾整備事業費 | 244,463 | 40.7   | 240,558 | 41.2   |
| 2公債費      | 355,634 | 59.3   | 344,017 | 58.8   |
| 3予備費      | 100    | 0.0    | 100     | 0.0    |
| 資金合計     | 600,197 | 100.0  | 584,675 | 100.0  |

cell word list

[“科目”
“本年度”
“前年度”
“比較”
“伸び率”
“予算額”
“構成比”
・・・]
Proposed Method (Method 1)

Inference

- Look at each cell in the table one by one.
- If any one of the inference results outputs a related (1), then it is a related table.

```
Table 1:
[ "科目", "本年度", "前年度", "比較" ]

Table 2
```

Related Table

まず、歳入についてでありますが、市税につきましては・・・

Cell Word list

results

Table 1:
0
0
1

Related Table
Proposed Method (Method 2)

- Output **frequently appearing words** (e.g., “合計(total)”, “科目(subject)”) as related in Method 1.
  → Many unrelated tables are linked together.

- **Delete words that appear frequently.**
- Delete words that appear 6 or more times and redo learning and inference.

```plaintext
Cell word list(text : number)

[  "産業廃棄物等処分事業": 2  
  "簡易水道事業": 2  
  "合計": 36  
  "科目": 18  
  "1市税": 3  
  "2地方譲与税": 3  
  ...]

Delete words that appear more than 6

New cell word list

[  "産業廃棄物等処分事業": 2  
  "簡易水道事業": 2  
  "1市税": 3  
  "2地方譲与税": 3  
  ...]
```
Proposed Method (Method 3)

- **Tag and column-by-column** learning and inference.
- Extract words by table column.
- Join the extracted words with delimiters to form a single sentence.
- Use `[SEP]` and ``(comma)`` as delimiters.

<table>
<thead>
<tr>
<th>科目</th>
<th>本年度</th>
<th>前年度</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>予算額</td>
<td>構成比</td>
</tr>
<tr>
<td></td>
<td>千円</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>初予算額</td>
<td>構成比</td>
</tr>
<tr>
<td></td>
<td>千円</td>
<td>%</td>
</tr>
<tr>
<td>1港湾整備事業費</td>
<td>244,463</td>
<td>40.7</td>
</tr>
<tr>
<td>2公債費</td>
<td>355,634</td>
<td>59.3</td>
</tr>
<tr>
<td>3予備費</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>歳出合計</td>
<td>600,197</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>240,558</td>
<td>41.2</td>
</tr>
<tr>
<td></td>
<td>344,017</td>
<td>58.8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>584,675</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Connect cell words by column with ``(comma)``
Proposed Method (Method 3)

- Label each column and combine with `<p>` tags for learning.
- Label related tables with 1 and unrelated tables with 0.
- If any one of the inference results outputs a related (1), then it is a related table.

Table1:
```
[  
  “科目,1港湾整備事業,,歳出合計”  
  “本年度,予算額”  
  “本年度,構成比”  
  “前年度,当初予算額”  
  “前年度,構成比”  
  ...  
]
```

Table2

まず、歳入についてでありますですが、市税につきましては・・・

Cell word list
Proposed Method (Keyword extraction)

- There are multiple statements with the same meaning, but with different expressions and letterforms.

- **Keyword extraction** using TFIDF to eliminate fluctuations in statements.
- Use extracted words in place of `<p>` tag text.
Proposed Method (Keyword extraction)

- Extract the top 10 words
- Words are joined by delimiters to form a single sentence.
- Use `[SEP]` and “,(comma)” as delimiters.
- Use extracted words in place of `<p>` tag text in Method 3.
The most accurate method is method 3, using ",(comma)" as the delimiter.

<table>
<thead>
<tr>
<th>Method</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>([SEP])</td>
<td>(,)</td>
</tr>
<tr>
<td>Result</td>
<td>1.43%</td>
<td>21.36%</td>
<td>27.23%</td>
<td>36.66%</td>
</tr>
<tr>
<td>Recall</td>
<td>88.79%</td>
<td>70.08%</td>
<td>31.28%</td>
<td>40.53%</td>
</tr>
<tr>
<td>Precision</td>
<td>00.74%</td>
<td>16.34%</td>
<td>46.75%</td>
<td>52.06%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyword extraction</th>
<th>keyword extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>([SEP])</td>
<td>(,)</td>
</tr>
<tr>
<td>Result</td>
<td>11.63%</td>
</tr>
<tr>
<td>Recall</td>
<td>12.55%</td>
</tr>
<tr>
<td>Precision</td>
<td>49.58%</td>
</tr>
</tbody>
</table>
CONSIDERATION (Problem1)

- There is a `<p>` tag with no tables linked.
  - There were 22 out of 81 `<p>` tags in the output result that had nothing linked to them.

- The more rows in the table, the more words to join.
  - Words that provide material for human judgment (e.g., “一般会計 (general accounting)”) are easily buried.

Tables not linked by our proposed method. In the Gold data, it is linked with many `<p>` tags.
CONSIDERATION (Problem 1)

Solution

- Need to be flexible according to the shape of the table.
- Separate tables by half of a row if there are many rows
  - Text created by joining columns can be shortened.
  - It can prevent necessary words from being buried.

```
“会計別,一般会計,港湾整備事業,青果物卸売市場事業,水産物卸売市場事業,国民健康保険事業,住宅事業,
簡易水道事業,介護保険事業,産業廃棄物処分事業,後期高齢者医療事業,病院事業,水道事業,下水道事業,
産業廃棄物等処分事業,合計”
```

```
“会計別,一般会計,港湾整備事業,青果物卸売市場事業,水産物卸売市場事業,国民健康保険事業,住宅事業,
簡易水道事業”
```

```
“介護保険事業,産業廃棄物処分事業,後期高齢者医療事業,病院事業,水道事業,下水道事業,産業廃棄物等処分事業,合計”
```
CONSIDERATION (Problem2)

- Distributed data may not be well maintained.
- The tables with lists are linked in the Gold data.
- The tables with individual details are not linked in the Gold data.

- In our proposed method, the following table was linked to the statement about "地方交付税(local allocation tax)".

<table>
<thead>
<tr>
<th>交付税</th>
<th>予算科目</th>
<th>本年度予算額</th>
<th>前年度予算額</th>
<th>比較</th>
<th>節目</th>
<th>区分</th>
<th>金額</th>
<th>説明</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>千円</td>
<td>千円</td>
<td></td>
<td>千円</td>
<td></td>
<td>千円</td>
<td></td>
</tr>
<tr>
<td>11 地方交付税</td>
<td>16,581,000</td>
<td>15,340,000</td>
<td>1,241,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1地方交付税</td>
<td>16,581,000</td>
<td>15,340,000</td>
<td>1,241,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1地方交付税</td>
<td>16,581,000</td>
<td>15,340,000</td>
<td>1,241,000</td>
<td>1地方交付税</td>
<td>16,581,000</td>
<td>普通交付税15,615,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table with details of local allocation tax
Thank you for listening.