# KSU at the NTCIR-17 UFO Task

〇Tomokazu Hayashi and Hisashi Miyamori
Kyoto Sangyo University

NTCIR-17 Day 3: December14th 10:30~11:30（JST）

# Outline

# Introduction

Tables in securities reports have a complex structure created by merged cells in the tables, etc.

(Unit：One million yen)（単位：百万円）

| ヘッジ会計の方法 Hedge accounting method | 取引の種類 Type of transaction | 主なヘッジ対象 Main hedged subject | 契約額等 Contract amount and so on | うち１年超 In over one years | 時価 Current market price |
|---|---|---|---|---|---|
| 原則的処理方法 Processing method in general | 商品スワップ取引 Swap transactions of Merchandise<br>受取変動・支払固定 Receive-floating ・ Pay-fixed<br>原油 Crude oil | 営業未払金 Non-operating trade payables | 69,132 | 27,452 | △24,304 |
| | 商品オプション取引 Option transactions of Merchandise<br>売建 Going short<br>プット Put Option 原油 Crude oil | 営業未払金 Non-operating trade payables | 33,120 | 15,468 | △7,229 |
| | 買建 Going long<br>コール Call Option 原油 Crude oil | 営業未払金 Non-operating trade payables | 42,798 | 20,103 | △1,717 |
| 合計 Sum | | | 145,051 | 63,025 | △33,250 |

Figure 1: Examples of tables included in the Annual Securities Report

# Methods Pretrained model

## TUTA [1]

- Hierarchical information contained in the table is represented by a tree-based structure called a bi-dimensional coordinate tree
- State-of-the-art results were achieved on five datasets

**TUTA's assumption**

When there is a hierarchical structure in a table,
the size of the merged cells decreases gradually
from the top (or leftmost) to the bottom (or rightmost)
of the table.

[1] Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. TUTA: Tree-based Transformers for Generally Structured Table Pre-training.

# Methods Conventional method

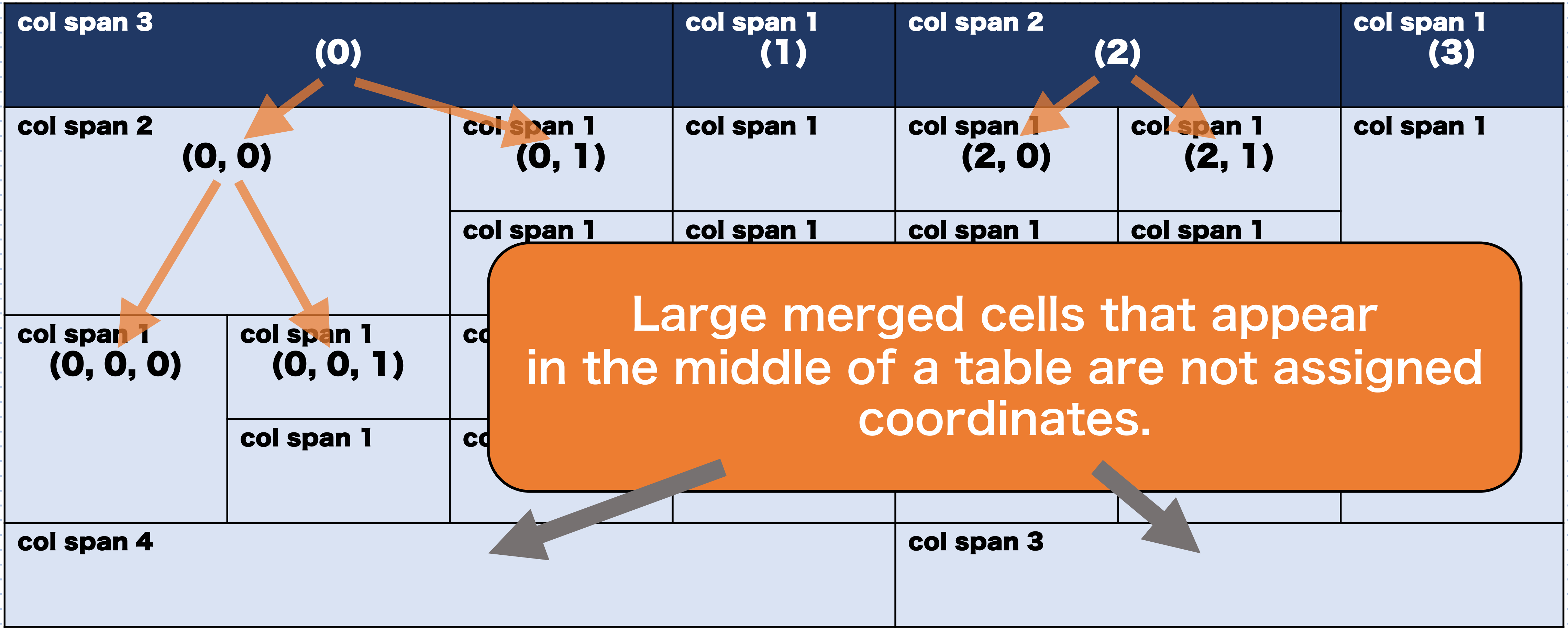


Figure 2: default vertical tree

# Methods Overview of the processing

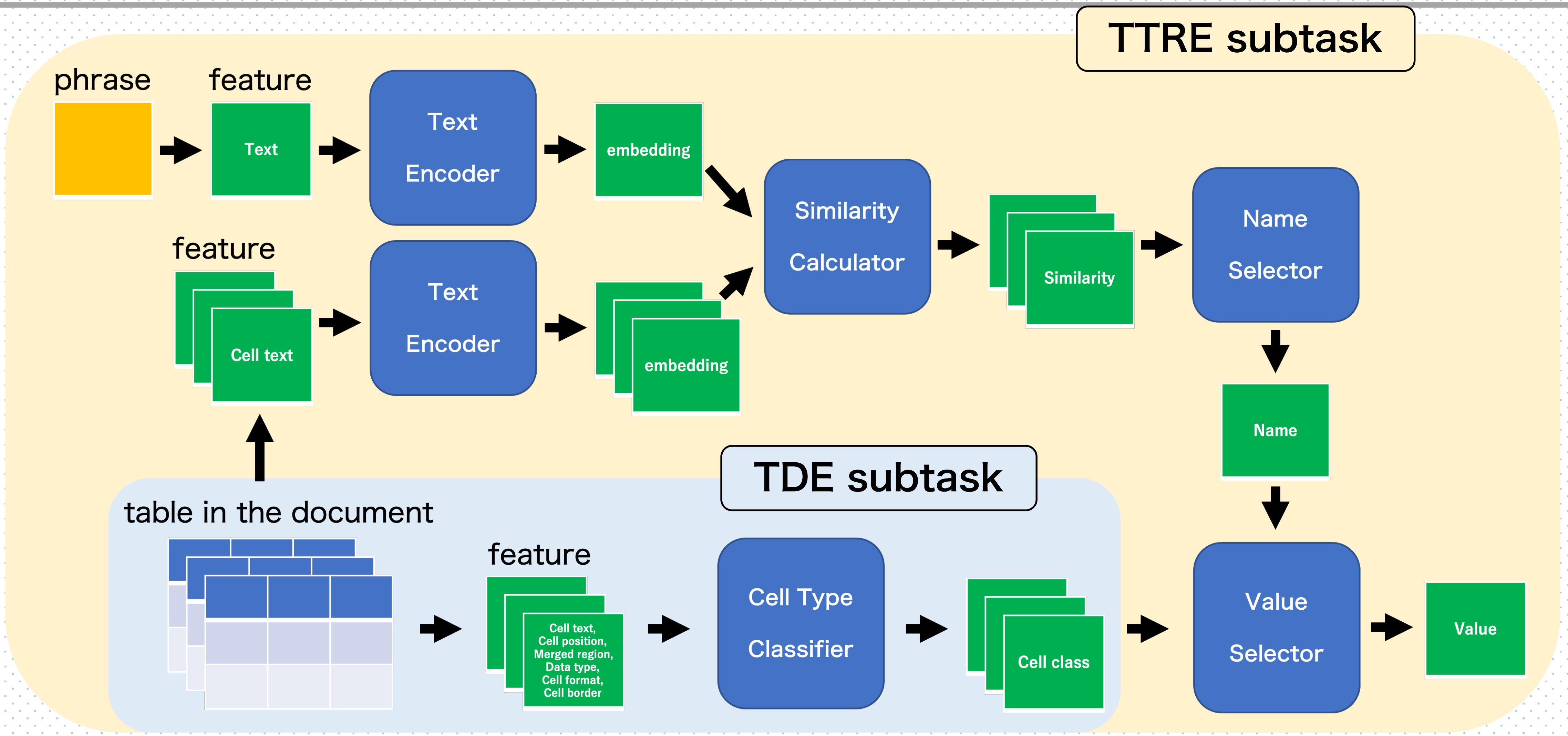


Figure 3: Overview of the processing of the proposed method in the UFO task

# Methods Table feature

Table 1: Feature set of table cells

| Cell Feature | e.g. |
|---|---|
| Cell text | • Cell text |
| Cell position | • Row / column indexes<br>• Tree-based coordinates |
| Merged region | • The number of merged rows<br>• The number of merged columns |
| Data type | • If cell string matches a date template<br>• If formula exists in the cell |
| Cell format | • If the bold font is applied<br>• If the background color is white<br>• If the font color is black |
| Cell border | • If cell has a top border<br>• If cell has a bottom border<br>• If cell has a left border<br>• If cell has a right border |

# Methods Proposed method

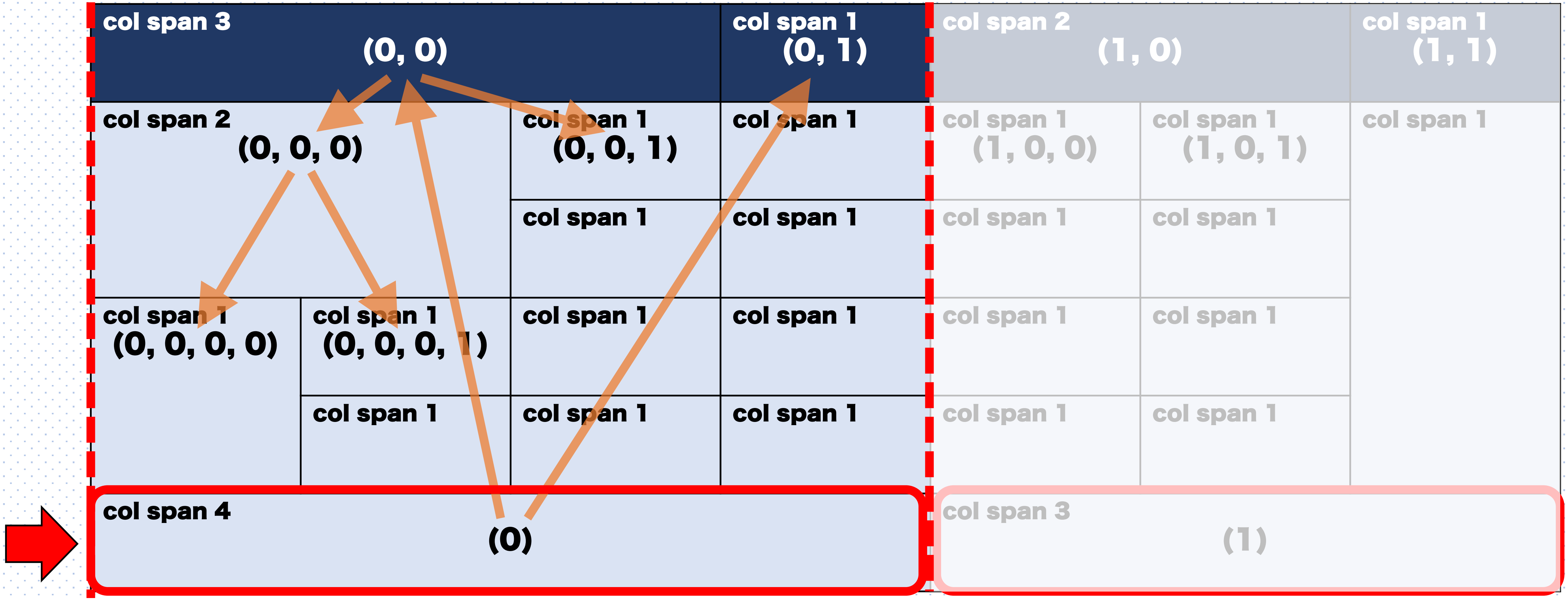


Figure 4: exhaustive vertical tree

# Methods Proposed method

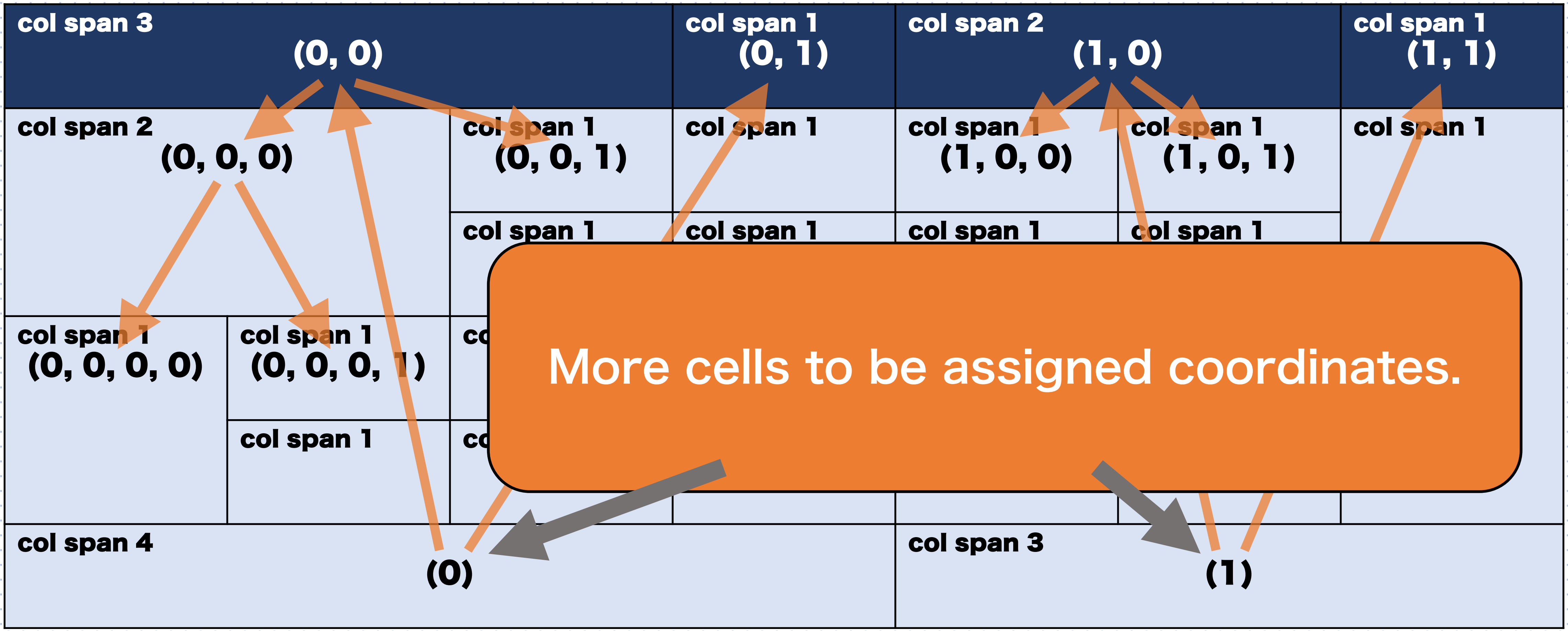


Figure 4: exhaustive vertical tree

# Methods Definition of terms

Table 2: Description of tree type

| Tree Type | Description |
| --- | --- |
| default tree | Bi-dimensional coordinate tree determined by TUTA |
| default vertical tree | vertical (column) tree |
| default horizontal tree | horizontal (row) tree |
| exhaustive tree | Bi-dimensional coordinate tree determined by the proposed method |
| exhaustive vertical tree | vertical (column) tree |
| exhaustive horizontal tree | horizontal (row) tree |

# Experiments Method to be evaluated

**TUTA** no tree

- Method that does not consider the tree structure of the table

**TUTA** default tree

- Conventional method

**TUTA** exhaustive

- Proposed method

# Methods Method to be evaluated

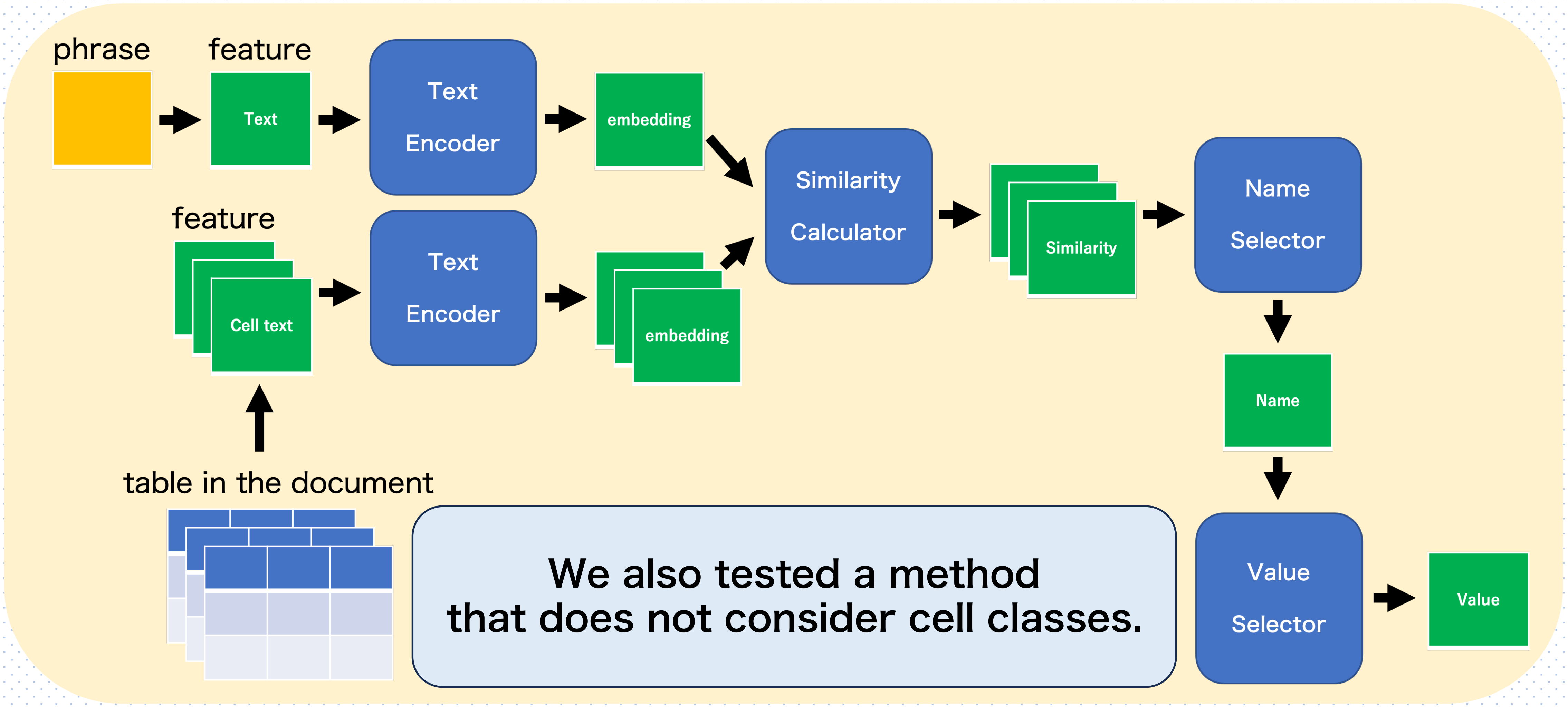


Figure 3: Overview of the processing of the proposed method in the UFO task

# Methods Method to be evaluated

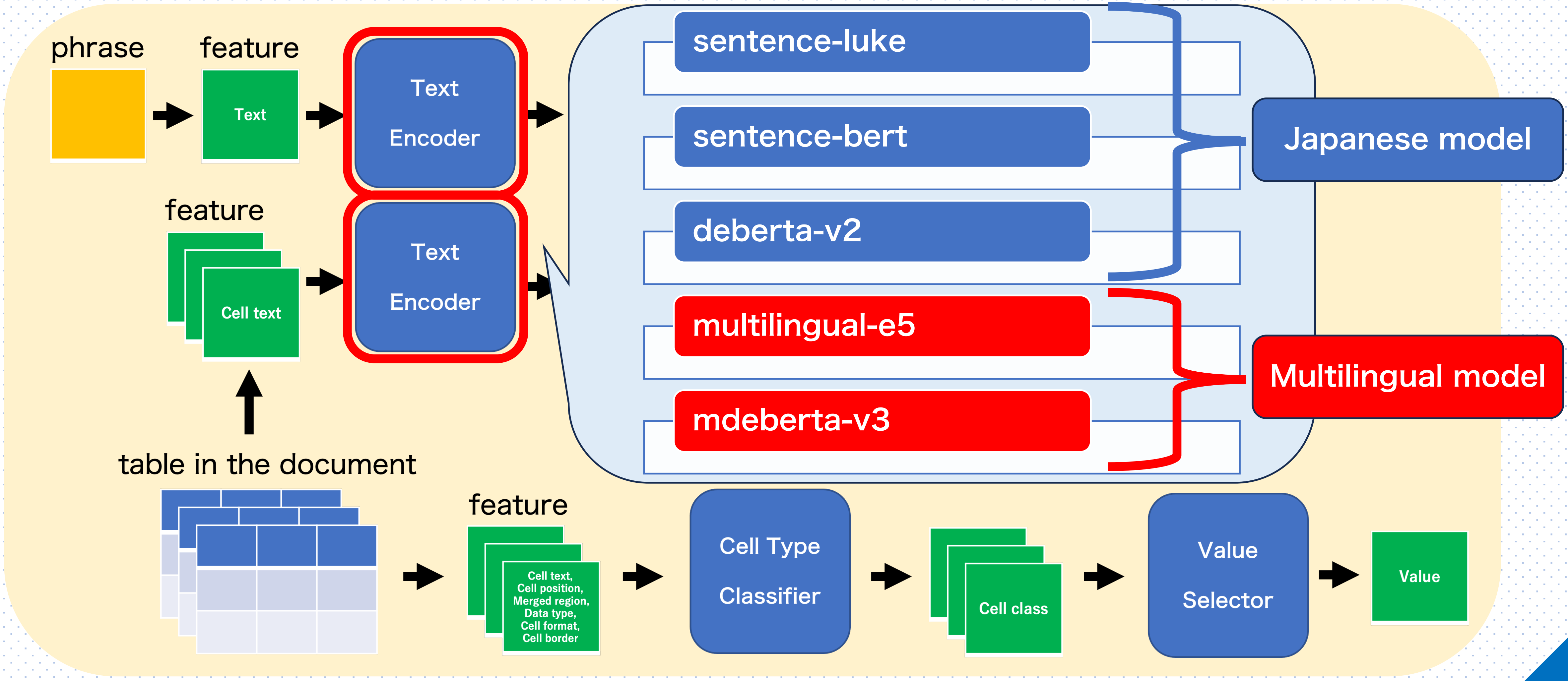


Figure 3: Overview of the processing of the proposed method in the UFO task

# Experiments Results of TDE subtask

Table 3: Scores of TDE subtask in formal run and late submission

| ID | Method | Precision | Recall | Macro-F1 |
|---|---|---|---|---|
| 81 | **TUTA** no tree | 0.7940 | **0.8216** | 0.8058 |
| 140 | **TUTA** default tree | **0.8815** | **0.8267** | **0.8496** |
| 150 | **TUTA** exhaustive tree | **0.8533** | 0.8196 | **0.8352** |

# Experiments Results of TTRE subtask

Table 4: Scores of TTRE subtask in formal run and late submission

| ID | Method | Name | | | Value | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | F1 |
| 122 | **multilingual-e5** | **0.3556** | **0.4574** | **0.3221** | 0.0857 | **0.5069** | 0.1186 | 0.2204 |
| 127 | **multilingual-e5 + TUTA** no tree | **0.3556** | **0.4574** | **0.3221** | **0.2682** | 0.5062 | **0.2719** | **0.2970** |
| 146 | **multilingual-e5 + TUTA** default tree | **0.3556** | **0.4574** | **0.3221** | **0.2658** | 0.5062 | **0.2704** | **0.2962** |
| 156 | **multilingual-e5 + TUTA** exhaustive tree | **0.3556** | **0.4574** | **0.3221** | 0.2646 | **0.5065** | 0.2659 | 0.2940 |

# Discussion Assumption

- **Why did the TUTA** no tree **show the best performance in the TTRE subtask?**
  - Table features were input into the TUTA model, and each cell was classified into one of four classes: Metadata, Header, Attribute, and Data.
  - Next, cells that belonged to the same row or column as the cell whose name was presumed to be Name were extracted, and cells that were classified into the Data class were designated as Value.

# Discussion Hypothesis

- **Why did the TUTA** no tree **show the best performance in the TTRE subtask?**

**We expected that the performance of the model for classifying Data classes in the TDE subtask would have a significant impact on the TTRE subtask.**

# Discussion Evidence

Table 5: Scores of experiments with TDE dataset

| ID | Method | header | attribute | data | metadata | macro-F1 |
|---|---|---|---|---|---|---|
| 81 | **TUTA** no tree | **0.8729** | **0.9935** | **0.8970** | 0.4600 | 0.8058 |
| 140 | **TUTA** default tree | 0.8604 | 0.9900 | 0.8911 | **0.6579** | **0.8496** |
| 151 | **TUTA** exhaustive tree | **0.8638** | **0.9917** | **0.8926** | **0.5055** | **0.8134** |

**It is important to improve the performance of the model for classifying Data classes**

# Discussion Results of TTRE subtask

Table 4: Scores of TTRE subtask in formal run and late submission

| ID | Method | Name | | | Value | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | F1 |
| 122 | **multilingual-e5** | **0.3556** | **0.4574** | **0.3221** | 0.0857 | **0.5069** | 0.1186 | 0.2204 |
| 127 | **multilingual-e5 + TUTA** no tree | **0.3556** | **0.4574** | **0.3221** | **0.2682** | 0.5062 | **0.2719** | **0.2970** |
| 146 | **multilingual-e5 + TUTA** default tree | **0.3556** | **0.4574** | **0.3221** | **0.2658** | 0.5062 | **0.2704** | **0.2962** |
| 156 | **multilingual-e5 + TUTA** exhaustive tree | **0.3556** | **0.4574** | **0.3221** | 0.2646 | **0.5065** | 0.2659 | 0.2940 |

# Discussion

- Comparing Value between the method using TUTA and the method without TUTA, there was no significant difference in recall and a significant difference in precision.

**This indicates that it is very effective to determine the cell type and exclude cells other than those of the Data class when determining Value.**

# Discussion Results of TTRE subtask

Table 4: Scores of TTRE subtask in formal run and late submission

| ID | Method | Name | | | Value | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | F1 |
| 122 | **multilingual-e5** | **0.3556** | **0.4574** | **0.3221** | 0.0857 | **0.5069** | 0.1186 | 0.2204 |
| 127 | **multilingual-e5** + **TUTA** no tree | **0.3556** | **0.4574** | **0.3221** | **0.2682** | 0.5062 | **0.2719** | **0.2970** |
| 146 | **multilingual-e5** + **TUTA** default tree | **0.3556** | **0.4574** | **0.3221** | **0.2658** | 0.5062 | **0.2704** | **0.2962** |
| 156 | **multilingual-e5** + **TUTA** exhaustive tree | **0.3556** | **0.4574** | **0.3221** | 0.2646 | **0.5065** | 0.2659 | 0.2940 |

# Discussion

- In all methods, recall was higher than precision.

This may be due to the fact that Name is determined only by the similarity between a given phrase and cell values, resulting in the acquisition of tables that are different from those that should actually be referenced.

# Additional Experiments

Table 6: Scores of experiments with DeEx dataset

| Method | metadata | notes | data | attributes | header | derived | macro-F1 |
|---|---|---|---|---|---|---|---|
| **TUTA** no tree | 0.8574 | **0.4825** | **0.9937** | **0.8114** | **0.8701** | **0.8092** | **0.8041** |
| **TUTA** default tree | **0.8585** | 0.4419 | 0.9930 | **0.8382** | 0.8547 | 0.7391 | 0.7876 |
| **TUTA** exhaustive tree | **0.8671** | **0.5589** | **0.9931** | 0.8045 | **0.8638** | 0.7491 | **0.8061** |
| **TUTA** raw tree | 0.8516 | 0.4271 | 0.9911 | 0.7888 | 0.8070 | **0.7767** | 0.7737 |

# Discussion Hypothesis

- Why did the proposed method perform poorly when validated on the TDE dataset?

We considered this to be due to the different characteristics of the tabular data contained in the TDE and DeEx datasets.

# Discussion Evidence

Table 7: Depth of the tree structure representation of the tabular data

| | TDE dataset | | DeEx dataset | |
|---|---|---|---|---|
| | mean | std | mean | std |
| default vertical tree | 0.3658 | 1.0948 | 0.4444 | 2.2411 |
| default horizontal tree | 0.2162 | 0.6906 | 0.1995 | 1.0503 |
| exhaustive vertical tree | 0.3919 | 0.8152 | 0.5079 | 1.0413 |
| exhaustive horizontal tree | 0.2255 | 0.6488 | 0.1814 | 0.6380 |

1.1994 1.3332

# Discussion

- Why did the proposed method perform poorly when validated on the TDE dataset?

We concluded that the method of representing tables in a tree structure works well when the tables have a complex structure.

# Conclusion

## TDE subtask

- The method of representing a table as a tree structure is most active when the table has a complex structure.

## TTRE subtask

- When searching for cells, it is effective to exclude cells that correspond to the Data class.
- However, it is not the overall performance of the cell type classification, but the performance for extracting the Data class that is important.

KYOTO SANGYO UNIVERSITY

Thank for your attention.