NTCIR-18 U4 ~~UFO 2.0~~

Unifying, Understanding, and Utilizing Unstructured Data in Financial Reports

Yasutomo Kimura

Otaru University of Commerce

Eisaku Sato

Otaru University of Commerce

Kazuma Kadowaki

The Japan Research Institute, Limited

Hokuto Ototake

Fukuoka University

NTCIR-18 kick-off meeting
March 29th 2024

Motivation

NTCIR-18 U4 task is designed to develop techniques for extracting structured information from tabular data and documents, with a special focus on **annual securities reports**.

Resolve issues in NTCIR-17 UF0 task, and promote research on table structure understanding.

- Not utilizing (correct data) XBRL for annual securities reports.
- Specific to securities reports and not applicable to tables in other fields.

Annual securities reports

有価証券報告書

第78期　自　平成28年 4 月 1 日
　　　　至　平成29年 3 月31日

大和ハウス工業株式会社

E00048

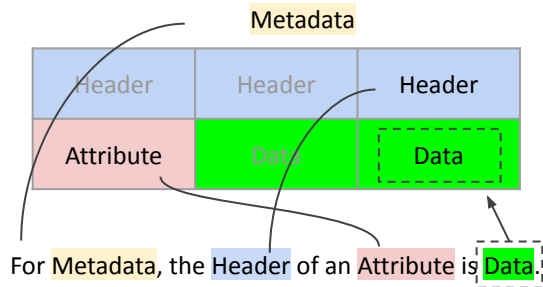# What is the problems of previous task (NTCIR-17 UFO) ?

https://sites.google.com/view/ntcir17-ufo

TDE — Table Data Extraction

TTRE — Text-to-Table Relationship Extraction

What is the problems?

- Not utilizing (correct data) XBRL for annual securities reports.
- Specific to securities reports and not applicable to tables in other fields.

NTCIR 18 U4 task

- Utilize CSV of annual securities reports (correct data)
- Perform QA tasks considering field expansion
  - Prepare a dataset of various question texts using ChatGPT.

Step 1 Create a question template from the CSV

Step 2 Create a commonly worded question using ChatGPT

要素ID	項目名	コンテキストID	相対年度	連結・個別	期間・時点	値
jpcrp_cor:NetSales...	売上高、経営指標等	Prior4YearDuration	四期前	その他	期間	319...0
jpcrp_cor:NetSales...	売上高、経営指標等	Prior3YearDuration	三期前	その他	期間	351...0
jpcrp_cor:NetSales...	売上高、経営指標等	Prior2YearDuration	前々期	その他	期間	379...0
jpcrp_cor:NetSales...	売上高、経営指標等	Prior1YearDuration	前期	その他	期間	414...0
jpcrp_cor:NetSales...	売上高、経営指標等	CurrentYearDuration	当期	その他	期間	438...0

ChatGPT

大和ハウス工業の2020年度の「売上高、経営指標等」は？

大和ハウス工業の2020年度の「売上高、経営指標等」は？

2020年度の大和ハウス工業における「売上高、経営指標等」は？

「売上高、経営指標等」について、2020年度大和ハウス工業では？

Step 3 Perform the following two subtasks using questions

Table Retrieval タスク

文書 から 質問文 を用いて 表 を検索・抽出

Table QA タスク

表 から 質問文 を用いて セル を回答・抽出

Table Retrieval (TR) subtask on NTCIR-18 U4

Sub task 1 : Table Retrieval (TR):

The subtask we aim to undertake involves extracting data from tables. Given a question and multiple tables across all sections in the ASRs, the system developed by the participating team will identify and search relevant tables from this entire collection.

Input :   Question and all tables in ASRs (HTML format)

Output :   Related tables

Evaluation: The basis shall be an accuracy-based score.

Example:

task1-Q1		
	"context"	"S100ITAZ"
	"question"	"What is the **table** including "Disposal of Treasury Stock" in Daiwa House Industry's individual ShareholderEquity for 2019?" "大和ハウス工業の2019年の個別のShareholdersEquityにおける「自己株式の処分」を含む表は？"
	"answer"	S100ITAZ-0000000-tab1

Table Question Answering (Table QA)

Sub task 2 : Table Question Answering (Table QA):

Our goal is to extract **the appropriate values** contained in the tables by means of natural language questions. Given a question and a target table in the ASRs, the system developed by the participating team will answer the question.

Input :   Question and a target table in ASRs (HTML format)

Output :   Answer (Cell values included in the table)

Evaluation: The basis shall be an accuracy-based score.

Example:

task2-Q1			
	"context"	"S100ITAZ"	
	"question"	"What cells indicate "disposal of treasury stock" in Daiwa House Industry's individual ShareholderEquity in 2019?" 大和ハウス工業の2019年の個別のShareholdersEquityにおける「自己株式の処分」を示すセルは？	
	"answer"	S100ITAZ-0000000-tab1-r1c1	

Schedule

Mar 2024   NTCIR-18 Kickoff Event
May 2024   Dataset release*
May 28-31 2024 Presentation of U4 task at a conference on Artificial Intelligence

JSAI2024 Organized session (OS-12 政治経済におけるAIの利活用)

Title：Proposed Tasks for Inter-company Comparisons Using TOPIX100 Annual Securities Reports
TOPIX100の有価証券報告書を用いた企業間比較のためのタスクの提案
Authors：Eisaku Sato, Yasutomo Kimura

Jun-Dec 2024   Dry run*
Sep 2024-Feb 2025   Formal run*
Feb 1, 2025   Evaluation results return
Feb 1, 2025   Task overview release (draft)
Mar 1, 2025   Submission due of participant papers (draft)
May 1, 2025   Camera-ready participant paper due
Jun 10-13, 2025 NTCIR-18 Conference