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/10.20736/0002001309

JPXIteam at the NTCIR-17 UFO Task

Nobushige Doi
JPX Market Innovation & Research, Inc.
Japan
n-doi@jpx.co.jp

Mayuri Tanaka
JPX Market Innovation & Research, Inc.
Japan
may-tanaka@jpx.co.jp

ABSTRACT

The JPXIteam participated in the table data extraction subtask of the NTCIR-17 UFO Task. This study outlines our methodology to address this challenge and analyzes the official results. Our approach to solving this subtask involved few-shot text classification using ChatGPT. This paper discusses the implications of these results, highlighting the contributions of this study in advancing table structure recognition.

KEYWORDS

few-shot text classification, ChatGPT, annual securities reports, table structure recognition, tabular data understanding

TEAM NAME

JPXIteam

SUBTASKS

Table Data Extraction

1 INTRODUCTION

Table data extraction (TDE), a component of the NTCIR-17 UFO Task[9], is a significant undertaking. It plays a crucial role in effectively comprehending and handling tabular data commonly found in annual security reports. TDE involves categorizing each cell within a table into distinct classes, including Metadata, Headers, Attributes, and Data. These classifications are instrumental for extracting information relevant to finance and its various applications.

One of the primary challenges in this research is the requirement for additional datasets. Recognizing tables of diverse forms and structures requires substantial annotated data. However, the datasets available for this task could be limited in size.

This study presents a novel approach that employs few-shot text classification method as a solution. In particular, we devised a workflow for the automatic classification of each cell within a table into Header, Attribute, or Data. Subsequently, we evaluated the accuracy of this classification using ChatGPT[11]. Notably, as part of the experimental approach, only a single cell was included in this study.

2 RELATED WORK

Research on extracting structured data from tables within electronic document files, including HTML and PDF formats, has been extensively conducted[2, 12–14]. While deep learning-based methods have been proposed for tabular data extraction, these conventional methods often require large amounts of annotated data. ChatGPT displays promise in zero- or few-shot text classification, a technique for categorizing text without relying on annotated data[5–8, 10].

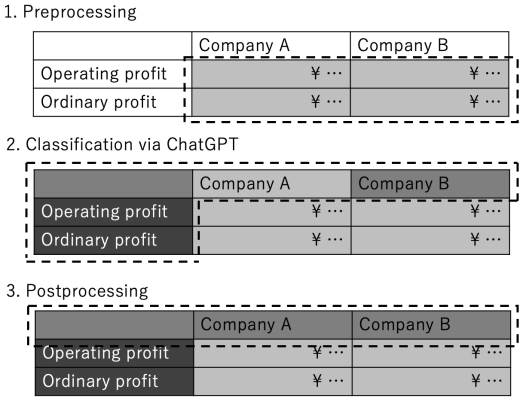

Figure 1: Flows of the proposal method. The cell fill represents the classification result. The dotted line represents examples of the range of cells to be classified in each step.

Furthermore, ChatGPT has been reported to help extract information from text in various fields[1, 3, 4, 15]. Methodologically, several studies have employed prompt engineering to steer ChatGPT responses. These associated investigations have demonstrated that these models can achieve performance comparable to models trained via supervised learning, particularly in the context of the English language. However, it has some limitations, particularly in under-resourced languages and within the explainability domain. This study examines whether the proposed method is effective in Japanese, a language with limited linguistic resources.

3 METHODS

The method employed in this study comprises three primary steps: preprocessing, text classification using ChatGPT, and postprocessing. This integrated approach was devised to efficiently and accurately classify table data. In particular, the preprocessing stage entails rule-based filtering, central processing step employs advanced text classification through ChatGPT, and postprocessing stage guarantees consistency across table elements. These processes are illustrated in Figure 1. Subsequent sections provide detailed explanation of each of these steps.

3.1 Preprocessing

Rule-based preprocessing of the tabular data is performed during the initial phase of the method. For instance, cells exclusively comprising numbers are categorized as Data. This preprocessing helps reduce the overall computational load by bypassing certain ChatGPT processing steps. Furthermore, cells corresponding to the Metadata are recognized in this preprocessing stage.

3.2 Classification via ChatGPT

After the preprocessing step, text classification is performed on the remaining cells using ChatGPT. The text from each cell is input as a prompt to ChatGPT, and the cells are categorized as Header, Attribute, or Data. If ChatGPT produced text that did not fall into the categories of Header, Attribute, or Data, the prompt was repeated until one of those categories was generated. An example of this prompt is provided in the Appendix. The next section details how this prompt was developed. In this section, we describe the characteristics of this prompt.

First, we meticulously examined the dataset provided for this task. This prompt contains explicit definitions and detailed information for each category: Header, Attribute, and Data. We initiated writing these definitions because no information about them was provided as part of this task.

Second, this prompt includes sample classifications for few-shot text classification. Although it may increase the usage cost of ChatGPT due to the additional tokens in the prompts, this is anticipated to enhance classification accuracy.

3.3 Postprocessing

In the final stage, postprocessing fine-tunes the table as a cohesive unit. When a certain percentage of Headers or Attributes is detected in the row or column direction of the table, all cells in that row or column are categorized as Headers or Attributes. This postprocessing enhances consistency across the entire table.

4 HOW TO CREATE PROMPTS FOR CHATGPT

This section outlines the creation of prompts for ChatGPT as part of the ChatGPT-based classification detailed in Subsection 3.2. The prompts enhance the understanding of the model and improve the classification accuracy of each cell in the table. The following procedures were implemented to achieve this goal.

4.1 Understanding the Context

Before creating the prompt, it is essential to understand the contexts. As discussed in Subsection 3.2, this task lacked predefined definitions for Headers, Attributes, and Data. Consequently, we developed definitions for these categories based on an examination of the dataset, which was integrated into the prompts as follows:

Header: Strings representing column names

Attribute: Strings signifying the primary key of each record

Data: Strings other than the primary key found within each record

4.2 Incorporate Examples

Incorporating examples into prompts can significantly enhance the accuracy of few-shot classification. Within this prompt, we described examples of Headers, Attributes, and Data through a feedback loop, ensuring a balanced approach to avoid content bias. Ultimately, each example was defined as follows:

Header: Revenue, Total, Amount (Million), (Name), Percentage (%), Name of segment, Position, Breakdown

Attribute: Company name, Person's name, Title, Security name, Range, Period

Data: (Note), Minato-ku, Tokyo, 100 shares

Table 1: Classification Results by Class

Class	# of Ref.	Precision	Recall	F-measure
Header	8949	0.7756	0.7244	0.7491
Attribute	7611	0.7218	0.4436	0.5495
Data	28895	0.8579	0.9629	0.9074
Metadata	44	0.8065	0.5682	0.6667

Table 2: Examples of Primary Misclassifications in Headers or Attributes

#	Translation of Text	Reference	Result
1	Name of subsidiary	Header	Attribute
2	Resolution of the Board of Directors on 25 May 2020	Attribute	Header
3	Return on equity	Header	Data
4	Book value at the beginning of the period	Attribute	Data
5	≥115% and <120%	Attribute	Data
6	2026 onwards	Attribute	Data

4.3 Minimizing the Number of Tokens

Although providing examples and definitions is valuable, brevity is equally crucial to minimize token usage in each ChatGPT query. Thus, prompts were carefully crafted to eliminate redundancy and unnecessary words.

4.4 Feedback Loop

The prompts underwent iterative refinement based on feedback from the classification experiments. The train dataset, consisting of actual table contents and assignments, was used to assess the effectiveness of the prompts. Subsequently, misclassifications were analyzed to fine-tune and enhance the prompts, aiming for improved results.

5 EXPERIMENTS

In this study, we employed few-shot text classification method to categorize 45,499 cells in the test dataset as Header, Attribute, Data, or Metadata. We used the the gpt-3.5-turbo-0301 for the classification via ChatGPT. Note that cells are classified into Metadata in only Preprocessing. The overall f-measure achieved was 0.8287, indicating a high level of classification accuracy, although not perfect.

Table 1 presents the classification results by class. When examining Table 1, a notable trend emerged, exhibiting higher Precision and Recall in classifying Data and a lower Recall in classifying Attributes. Additionally, challenges were observed where text that should have been classified as a Header or Attribute was misclassified. This observation suggests that the proposal method is effective in classifying Data, but not the best at classifying Headers or Attributes, or it is difficult to distinguish Headers or Attributes.

Table 2 presents examples of primary misclassifications regarding Headers or Attributes in the train dataset. We have concluded that these misclassifications also may occur in 45,499 cells in the test dataset. The next section discusses each of these misclassifications based on the provided examples.

前連結会計年度の買収により取得した資産、引き継いだ負債の修正による前連結会計年度の連結財政状態計算書への影響額は次のとおりであります。

（単位：百万円）

	遡及修正額
非流動資産	
有形固定資産	3,030
無形資産	17,048
その他の非流動資産	157
取得資産の公正価値	20,235
非流動負債	
繰延税金負債	5,025
引受債務の公正価値	5,025
取得資産及び引受債務の公正価値（純額）	15,210
のれん	△11,371

のれんの増減については、注記「14. のれん及び無形資産」に記載しております。なお、上記無形資産は下記で構成されております。

（単位：百万円）

	加重平均償却年数	遡及修正額
商標権	非償却対象	1,881
顧客関係	17年	13,693
専有技術	9年	1,474

Figure 2: Same texts are differently classified as Header and Attribute in the train dataset. Red cells represent the Header and blue cells represent the Attribute. The yellow "retrospective adjustment amount" texts are classified as Header and Attribute.

First, as illustrated in Examples 1 and 2, the Header and Attribute labels were mistakenly interchanged. Such misclassifications were notably common when the distinctions between Header and Attribute were subtle, thereby rendering manual classification challenging in some instances. This finding suggests a reduced need for strict differentiation between Header and Attribute in table structure recognition.

Second, scattered instances of misclassification were observed where words related to accounting terms were incorrectly categorized as Data, such as in Examples 3 and 4. These observations indicate that ChatGPT may occasionally struggle to recognize technical accounting terminology. Furthermore, although not included in Table 2, a similar issue was noted with proper names, including company and personal names, being mistakenly classified as Data.

Third, there were scattered misclassifications of numerical ranges, as observed in Examples 5 and 6. Despite the word "range" in the Attribute example in the prompt, numerical ranges tended to be classified as Data rather than Attributes, possibly due to the implied association of "Data" with values. These findings suggest that there are limitations in guiding classification outcomes solely based on descriptions in the prompt.

As an additional misclassification case, instances arose during the experimental process where multiple valid classifications exist for the same text within cells. For instance, consider representing two yellow "retrospective adjustment amount" cells in Figure 2. Figure 2 shows a screenshot of the tables included in the train dataset. In the upper table of Figure 2, "retrospective adjustments" is classified as Header (red), whereas it is categorized as Attribute (blue) in the lower table. Such instances may introduce challenges tied to subtle distinctions between the Header and Attribute, as exemplified in Examples 1 and 2. Therefore, even if the text of the cell is identical, accurate classification depends on considering the structure of the entire table.

In summary, the three-part approach involving preprocessing, ChatGPT-based classification, and postprocessing demonstrated effectiveness to a certain extent. However, there is room for improvement, particularly in differentiating between Header and Attribute and when single-cell classification is challenging. Future study could explore incorporating more contextual data into the prompts, adopting a holistic approach that considers the entire table structure to address these challenges.

6 CONCLUSIONS

In this study, we have explored the application of few-shot text classification using ChatGPT to recognize table structures in annual securities reports. Our experiments have demonstrated the efficacy of ChatGPT in automatically identifying these table structures to a certain extent. Our findings highlight the promise of leveraging ChatGPT in document structure recognition.

However, it is important to acknowledge the room for improvement in the classification accuracy of our proposed approach, particularly in the misclassification between Headers and Attributes. This limitation can be attributed to the inherent variability of cells, assuming different semantic roles depending on the table's structure and contextual information. This variability underscores the challenge of achieving precise classification based solely on the content of individual cells.

As we look to the future, one avenue for further research involves a more comprehensive integration of the context of multiple cells and the overall table structure into the prompts provided to ChatGPT. This approach could enhance the accuracy and robustness of table structure recognition.

In conclusion, our study contributes to the ongoing efforts to automate data extraction from financial documents, and it underscores the need for continued research to address the challenges posed by diverse table structures and the nuances of contextual information. By doing so, we can unlock the full potential of AI-driven document analysis in finance and beyond.

REFERENCES

[1] Alexander Brinkmann, Roee Shraga, Reng Chiz Der, and Christian Bizer. 2023. Product Information Extraction using ChatGPT. arXiv:2306.14921 [cs.CL]

[2] Xin Luna Dong, Hannaneh Hajishirzi, Colin Lockard, and Prashant Shiralkar. 2020. Multi-modal information extraction from text, semi-structured, and tabular data on the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3543–3544.

[3] Matthias A Fink, Arved Bischoff, Christoph A Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, and Tim F Weber. 2023. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 308, 3 (2023), e231362.

[4] Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the Feasibility of ChatGPT for Event Extraction. arXiv:2303.03836 [cs.CL]

[5] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (jul 2023). https://doi.org/10.1073/pnas.2305016120

[6] Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot Clinical Entity Recognition using ChatGPT. arXiv:2303.16416 [cs.CL]

[7] Yibo Hu, Erick Skorupa Parolin, Latifur Khan, Patrick T. Brandt, Javier Osorio, and Vito J. D'Orazio. 2023. Synthesizing Political Zero-Shot Relation Classification via Codebook Knowledge, NLI, and ChatGPT. arXiv:2308.07876 [cs.CL]

[8] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate

Speech. In *Companion Proceedings of the ACM Web Conference 2023*. ACM. https://doi.org/10.1145/3543873.3587368

[9] Yasutomo Kimura, Hokuto Ototake, Kazuma Kadowaki, Takahito Kondo, and Makoto P. Kato. 2023. Overview of the NTCIR-17 UFO Task. *Proceedings of The 17th NTCIR Conference* (12 2023). https://doi.org/10.20736/0002001321

[10] Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. arXiv:2303.03953 [cs.CL]

[11] OpenAI. [n. d.]. Introducing ChatGPT. https://openai.com/blog/chatgpt (accessed on 1 September 2023).

[12] Viacheslav Paramonov, Alexey Shigarov, and Varvara Vetrova. 2020. Table Header Correction Algorithm Based on Heuristics for Improving Spreadsheet Data Extraction. In *Information and Software Technologies*, Audrius Lopata, Rita Butkienė, Daina Gudonienė, and Vilma Sukackė (Eds.). Springer International Publishing, Cham, 147–158.

[13] Vaibhav Rupapara, Manideep Narra, Naresh Kumar Gonda, and Kaushika Thipparthy. 2020. Relevant Data Node Extraction:A Web Data Extraction Method for Non Contagious Data. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. 500–505. https://doi.org/10.1109/ICCES48766.2020.9137897

[14] Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4634–4642.

[15] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-Shot Information Extraction via Chatting with ChatGPT. arXiv:2302.10205 [cs.CL]

A EXAMPLE PROMPT

This chapter offers an example of a prompt for few-shot text classification method using ChatGPT, as detailed in Subsection 3.2. Note that the provided prompt has been translated into English for reference purposes, as the original prompt is in Japanese.

Example Prompt (reference translation)

Below are item classifications and actual text excerpts from annual security report tables, which we intend to use for the creation of a database.

Classification class
Header, Attribute, Data

Class definitions
Header: Column name strings
Attribute: Primary key strings for each record
Data: Strings in each record that are not the primary key

Specific examples
Header: Revenue, Total, Amount (Million), (Name), Percentage (%), Name of segment, Position, Breakdown
Attribute: Company name, Person's name, Title, Security name, Range, Period
Data: (Note), Minato-ku, Tokyo, 100 shares

Input text
{The text of a cell}

Task
Classify the input text into a classification class and output the corresponding class name.

Output text