Synthetic Context with LLM for Entity Linking from Scientific Tables

Yuji Oshima[1,2] **Hiroyuki Shindo**[1] **Hiroki Teranishi**[3,1]
Hiroki Ouchi[1,3] **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology [2]Mercari inc. [3]RIKEN
{oshima.yuji.ov6, shindo, hiroki.ouchi, taro}@is.naist.jp
hiroki.teranishi@riken.jp

Abstract

Tables in scientific papers contain crucial information, such as experimental results. Entity Linking (EL) is a promising task that analyses tables and associates them with a knowledge base. EL for table cells requires identifying the referent concept of each cell while understanding the context relevant to each cell in the paper. However, extracting the relevant context from the paper is challenging because the relevant parts are scattered in the main text and captions. This study defines a rule-based method for extracting broad context from the main text, including table captions and sentences that mention the table. Furthermore, we propose synthetic context as a more refined context generated by large language models (LLMs). In a synthetic context, contexts from the entire paper are refined by summarizing, injecting supplemental knowledge, and clarifying the referent concept. We observe this approach improves accuracy for EL by more than 10 points on the S2abEL dataset, and our qualitative analysis suggests potential future works.

1 Introduction

Information analysis of scientific papers has numerous applications in accelerating science, such as paper retrieval, reading assistance, and automatic knowledge base construction. In particular in information science, crucial information, such as experimental results, evaluation datasets, tasks, and evaluation metrics, is often recorded in tables within the papers. Thus, the analysis of table information is an important research field.

For table analysis, entity linking (EL) that associates table cells in scientific papers with a knowledge base (KB) is an important task, and various methods and datasets have been proposed for this purpose (Kardas et al., 2020; Yang et al., 2022; Lou et al., 2023). S2abEL (Lou et al., 2023) is a large-scale evaluation dataset for EL targeting tables in papers for the machine learning field. In

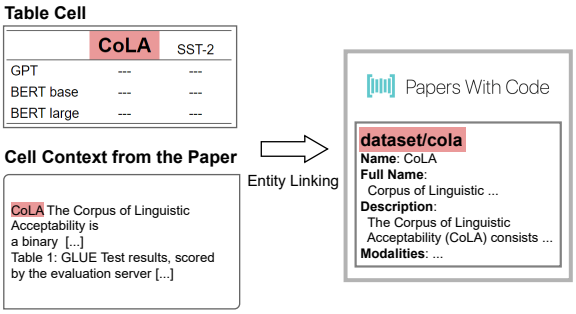

Figure 1: Example of entity linking for table cells in Devlin et al. (2019). Given a target table cell (e.g., "CoLA"), a model seeks to link it to the corresponding entity (e.g., "/dataset/cola") in *Papers with Code* by considering the contexts in the paper related to the cell.

the dataset, each table cell is linked to an entity defined in Papers With Code (PwC)[1], a free and open KB in the scientific domain, as illustrated in Figure 1. To correctly link the target table cell "CoLA" to the corresponding entity "/dataset/cola" in PwC, a model needs to understand the concept of CoLA from the contexts scattered in the main text, captions, and references.

However, extracting such contexts relevant to each cell from a paper has three technical challenges. (i) Relevant contexts for a cell text are scattered in an entire paper, and mentions are often abbreviated or paraphrased, (ii) The context or explanation for a referent concept of a cell can be insufficient, and (iii) General words such as "Ours," "Baseline," and "All" are often used in cell texts, and the referent is ambiguous. An example for the first, in the paper of *The Evolved Transformer* (So et al., 2019), a cell text "ET PERP" is interpreted as "the perplexity achieved by the Evolved Transformer", although the term "ET PERP" does not appear in the main text of the paper. Second, explanations for well-known methods such as LSTM are often omitted, and thus, sufficient contexts are

[1]https://paperswithcode.com/

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 202–214

unavailable in the paper itself. Third, the cell text "All" stands for the entire dataset. Identifying the dataset requires understanding the context of the main text. However, the word "All" is general and frequently used in irrelevant contexts in the paper.

To address these challenges, we propose a data synthesis method for providing an EL model with supplemental contexts for table cells by using large language models (LLMs) such as ChatGPT (OpenAI, 2023) and LLaMa2 (Touvron et al., 2023). LLMs, acquiring specialized knowledge through pre-training, can be utilized as knowledge bases (Taylor et al., 2022). They also demonstrate high performance in a zero-shot setting for abbreviation expansion (Gorman et al., 2021) and coreference resolution (Wei et al., 2022). Therefore, we expect that they can generate additional information for table cells, which was not mined by previous methods.

In experiments, we use context data for EL obtained by our refined rule-based method and LLM-based method and confirm consistent improvements by both methods over the baseline proposed in S2abEL, including a 10-point improvement in accuracy for EL. Furthermore, by utilizing synthetic context, we improved the accuracy of a subtask that extracts relevant papers for a given table cell by more than 12 points in terms of top-5 accuracy. This result demonstrates the effectiveness of our approach in linking table cells to entities without relying on existing knowledge bases. Our qualitative analysis also reveals that synthetic context data captures better supplemental information through context completion and knowledge completion by LLMs[2].

2 Related Work

2.1 Entity Linking in Scientific Table

Table analysis is crucial for extracting experimental information and results in information extraction from scientific papers. For instance, Axcell (Kardas et al., 2020) extracts tables from the LaTeXsource of papers and performs linking of table cells to entities in a knowledge base. Similarly, S2abEL (Lou et al., 2023) constructs a dataset annotated with entities linked to cells, along with the type of information and the source references for that information, for comparable tasks. Axcell and S2abEL use features representing table cells, such as the cell's positional information and text from the main body that matches the cell's text. SciREX (Jain et al., 2020) and CitationIE (Viswanathan et al., 2021) aim to extract information from the entire paper, not only tables. Kostić et al. (2021) and Zhuang et al. (2022) perform entity extraction and relation extraction from both the text and tables. In these works, the entire document is converted into a feature. However, relevant descriptions of specific table cells are scattered throughout the document. Therefore, it is necessary to efficiently extract the contexts of the cells from the document.

2.2 Data Augmentation/Synthetic Data

Data augmentation and synthesis using LLMs are employed in various tasks. Lai et al. (2022) demonstrate the effectiveness of information completion using generative models, while Chen et al. (2023) show that information summarization is beneficial for entity linking. This study aims to generate sufficient and necessary information for linking by simultaneously performing information completion and summarization within the full context of papers. On the other hand, some studies leverage the asymmetry in task difficulty, where inverse problems are easier to solve than forward problems, to generate training data by solving inverse problems using LLMs. Wang et al. (2021) employ few-shot learning to create training data from labels, and Josifoski et al. (2023) generate synthetic training data for the general information extraction task. Although these are effective methods for problems that are difficult to solve directly with LLM's zero-shot or few-shot capabilities, synthesizing tables or papers from entities to be linked is challenging. Therefore, this study aims to enhance the learning efficiency of existing human-annotated data using synthetic data generated with LLMs.

3 Entity Linking in Scientific Tables

EL for scientific tables aims to map each table cell within a paper to an entity in a KB (PwC in our experiments) or "OutKB" if no corresponding entity is found in the KB. The baseline method *S2abEL* is proposed by Lou et al. (2023). They divide this task into the following subtasks:

1. **Attributed Source Matching (ASM)**: Identifying the *attributed source(s)* for a table cell within a paper. The attributed source(s) is the reference paper that originally proposed or introduced the concept that a target cell refers

[2]We will release our code and data to reproduce the experiments.

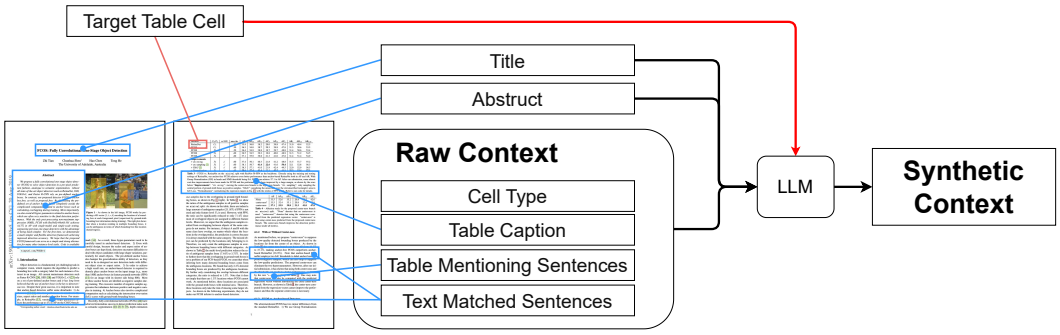

Figure 2: Generation of synthetic context: Sentences related to a particular cell (raw context) are provided to a Large Language Model (LLM). By having the LLM explain the content of the target cell, contextual information related to the target cell is extracted from the raw context.

to. This step aims to distinguish similar surface forms to find the correct referent entities in the subsequent subtasks. e.g., the paper by Warstadt et al. (2019) is identified as the attributed source of "CoLA".

2. **Entity Disambiguation (ED)**: Retrieving candidate entities from the KB that are likely to be linked to a target cell. Then, selecting the referent entity from the candidates for a given cell (or assign *OutKB* if none of them is appropriate). e.g., The entities associated with the paper by Warstadt et al. (2019) in PwC are added to the candidates for "CoLA" and "/dataset/cola" is selected from among them.

Note that Lou et al. (2023) perform cell type classification before the pipeline[3]. However, the classification of cell types has already exceeded the accuracy 90% in previous research, and replacing it with the correct types contributes less than a 1% improvement in the final EL accuracy. Therefore, we use the correct cell types to proceed to the subsequent subtasks.

4 Method

In this study, we aim to improve the overall accuracy of EL by enhancing the context of cells, which is the most important feature for both ASM and ED tasks in the baseline method. To this end, we first improve the rule-based context extraction method and then generate synthetic data to refine the context.

4.1 Supplementing Context Information

Meticulous cell context extraction: For EL, a model needs to interpret the concept that a cell text represents and extract appropriate contexts for it from the paper. As context information of a target cell, the prior research has utilized various features, including sentences retrieved by BM25 (Robertson and Zaragoza, 2009), the cell's position in the table, and the surrounding cells. However, the retrieval method can miss relevant sentences or extract irrelevant sentences due to text fluctuation (e.g., abbreviation, paraphrasing) or the use of general words, resulting in insufficient and erroneous information sourced from the main text. To alleviate this, we first collect text fragments covering broader contexts. Specifically, we use the following features as the contexts for a target cell: (i) The cell's text. (ii) The cell type. (iii) The table caption. (iv) Sentences referring to the table: Sentences that explicitly contain references to the table, such as "Table 1." (v) Sentences containing the cell's text. We refer to a set of the features as the *raw context* for a target cell.

For example, the raw context of a cell in the paper by Devlin et al. (2019) illustrated in Figure 1 is as follows: (i) *CoLA*. (ii) *dataset*. (iii) *"Table 1: GLUE Test results, scored by the evaluation server [...]"*. (iv) *"Results are presented in Table 1"*. (v) *"CoLA The Corpus of Linguistic Acceptability is a binary single-sentence [...]"*.

Synthetic context generation: To focus on essential information in a raw context and supplement it by injecting external knowledge, we employ an LLM to generate a description for a cell based on the corresponding raw context. We refer to a description generated by an LLM as a *synthetic context*. The process of synthetic context generation in this study is illustrated in Figure 2. In this research, we employ OpenAI's GPT-4 Turbo (1106) as the LLM.

[3]The cell types are *Method*, *Dataset*, *Metrics*, *Dataset&Metrics*, and *Other*.

For example, the synthetic context generated by the LLM for the example shown in Figure 1 is *"CoLA stands for Corpus of Linguistic Acceptability. It is a dataset used for a binary single-sentence classification task in natural language processing. [...]"*. This exhibits the LLM's capabilities of providing a synthetic context that summarizes adequate information for EL.

4.2 Subtasks of Entity Linking

We integrate the improved context extraction/generation methods into each subtask in the baseline pipeline as follows.

Attributed Source Matching: We follow the approach of S2abEL (Lou et al., 2023) for ASM. The potential attributed sources for a cell are all cited papers and the current document itself. Including the document itself is necessary for the case where the cell's referent concept is newly proposed in the document. To find the attributed source from the potential source, we calculate the relevance scores between the cell and each potential source. As features for a scoring model, we concatenate the title and abstract of a potential source and the cell's context. As the scoring model, S2abEL adopts SciBERT (Reimers and Gurevych, 2019). In this research, we employ SciBERT with raw context and GPT-2 (Radford et al., 2019) with synthetic context[4]. A scoring model is trained with binary cross-entropy loss.

Entity Disambiguation: To collect candidate entities for a target cell, we sort the attributed sources by the ASM score and then retrieve entities from the KB for each attributed source until we obtain k candidates for the cell. To select the most promising entity from the candidates, following the prior work, we employ a model to calculate a score for each entity candidate by feeding the concatenation of the entity name, its description, and the cell context into the model. We adopt SciBERT for the scoring model and train it with binary cross-entropy loss. The highest-scoring entity is linked to the cell when the score is greater than a predefined threshold. Otherwise, OutKB is assigned to the cell, representing being out of KB. We set the threshold to 0.5, the same as prior research.

[4]As the token length of synthetic contexts often exceeds the input token limit of SciBERT (512), we did not employ SciBERT with synthetic context.

5 Experiments

The experiments follow the setup of S2abEL, where training, validation, and test data are created from different topics. The results below represent the cross-validation average on the topics in S2abEL. This allows us to compare the generalization performance of the models without overfitting to specific topics.

5.1 End-to-end Entity Linking

In EL experiments, as explained in §3, the cell types are determined using the ground truth data.

5.1.1 Experimental Settings

In end-to-end entity linking experiments, we compare raw context and synthetic context against S2abEL (Lou et al., 2023) as the baseline. The baseline additionally leverages dense retrieval (DR) for ED to retrieve entity candidates from the KB directly, but we do not use it to see the sole effect when using raw context and synthetic context[5]. We apply the same context to both ASM and ED. We report three metrics: InKB accuracy, OutKB F1, and overall accuracy. InKB hit@1 accuracy shows the hit rate at the top when an entity to be linked is present. For OutKB entities, we report F1 score. The overall accuracy is determined as follows[6]:

1. For OutKB mentions: A cell is considered correct if predicted as an OutKB mention.

2. For InKB mentions: A cell is considered correct if ranked as the top prediction (@top1 hit).

The number of entity candidates k retrieved is set to $k = 50$ (the same as prior work) and $k = 20$. Training details are displayed in Appendix D.

5.1.2 Result

Table 1 shows that the raw context and synthetic context conditions have improved overall accuracy by over 10 points compared to the S2abEL baseline when $k = 50$. This indicates that the necessary context information for the EL task has been successfully extracted from the main text in the raw context. When comparing synthetic context to raw context, there is a slight improvement in overall

[5]The experimental results in S2abEL reported that the ASM-based retrieval method without using DR achieves over 90% recall when $k \geq 30$.

[6]In the S2abEL dataset, 42.8% cells are marked as OutKB mentions.

Method	Overall acc.	OutKB F1	InKB hit@1
	$k = 50$		
S2abEL (Lou et al., 2023)	58.2	71.4	33.4
S2abEL w/o DR	60.8	71.7	27.1
Raw Context	69.9	**76.5**	47.1
Synthetic Context	**70.5**	76.4	**53.1**
	$k = 20$		
S2abEL w/o DR	60.2	70.3	25.3
Raw Context	68.9	75.5	44.7
Synthetic Context	**70.8**	**76.6**	**52.2**

Table 1: Result of End-to-end Entity Linking

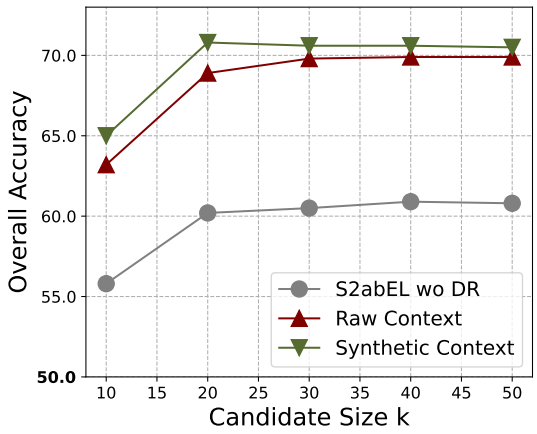

Figure 3: Evaluation of different Contexts on variation number of candidate entity.

accuracy and a 6-point increase in the InKB hit@1. This demonstrates that synthetic context effectively captures the appropriate information from raw context and is supplemented by the LLM.

In the results with smaller entity candidates $k = 20$, the difference between raw context and synthetic context has become more pronounced. This suggests that in the case of $k = 50$, there is almost no difference in the entity candidates extracted by raw context and synthetic context. We observe that the overall accuracy for synthetic context does not much degrade for smaller $k < 30$, compared with other conditions in Figure 3. In the synthetic context setting, the proposed EL pipeline achieves better accuracy with 20 candidates than with 50 candidates. This might be because, when the correct entity is successfully retrieved within the top 20 candidates ($k = 20$), increasing the number to 50 ($k = 50$) introduces noise, which leads to a decrease in accuracy.

5.2 Evaluating Method Combinations for ASM and ED in Entity Linking Tasks

5.2.1 Experimental Settings

To observe the effects of the proposed raw/synthetic contexts in ASM and ED subtasks, we compare the accuracy of EL with exhaustive combinations of contexts and subtasks[7]. The number of entity candidates is $k = 20$ since the difference of entity candidates is small in larger k.

5.2.2 Result

Table 2 shows that using either raw or synthetic context for only ASM or ED improves accuracy compared to using the context defined in S2abEL. When used only for ASM, the improvement in accuracy is about 1 or 2 points, while for ED, the improvement is around 6 to 7 points. When comparing the two proposed methods applied to ASM, it is observed that synthetic context improves the overall accuracy by 1 point, indicating that synthetic context is capable of enhancing ASM. When raw or synthetic context is applied only to ED, the improvements in InKB hit@1 are larger than when applied only to ASM. In any conditions, synthetic context consistently achieves higher accuracy than raw and S2abEL contexts. These results suggest that both raw context and synthetic context are effective for both ASM and ED tasks and that synthetic context is capable of representing more effective contexts than raw context.

5.3 Evaluating the Impact of Context in ASM

To directly observe the effect of the improved context, we evaluate using the precision of the ASM. This experiment is evaluated only on cells with attributed source papers. In the S2abEL dataset, some cells do not have an attributed source, and we filtered out these cells in this experiment. Furthermore, unlike when used as a subtask of EL, it is evaluated based on the accuracy of selecting the attributed source paper rather than the entity. The results are evaluated based on the accuracy of the top 1 and top 5 ranked by the score.

[7]To investigate the effects of context refinement solely, we do not use dense retrieval for collecting candidates in the pipeline.

ASM Method	ED Method	Overall acc.	OutKB F1	InKB hit@1
S2abEL	S2abEL	60.2	70.3	25.3
Raw Context	S2abEL	61.0	70.8	27.2
Synthetic Context	S2abEL	62.2	72.0	28.8
S2abEL	Raw Context	66.8	73.8	41.4
S2abEL	Synthetic Context	67.4	73.3	46.7
Raw Context	Raw Context	68.9	75.5	44.7
Synthetic Context	Synthetic Context	**70.8**	**76.6**	**52.2**

Table 2: Performance Comparison of ASM and ED Method Combinations in Entity Linking Tasks

5.3.1 Variation of Context and Scoring model

In the ASM task, we calculate scores for all cited references and the target paper to select the attributed source for the cell. The scoring model is given the cell's context and the title and abstract of each attributed source candidate. We compare variations of this scoring model and the cell context.

GPT4 Zeroshot We leverage GPT4-Turbo (1106) to ASM in a zero-shot setting. The raw contexts, along with the titles and abstracts of all cited references, are given to GPT4-Turbo, and it infers the attributed source paper directly. For GPT4 Zeroshot, the evaluation is based solely on the top 1 accuracy since it directly selects the cited reference without scoring.

Cell Context Three types of context are to be compared. S2abEL is the context defined in S2abEL paper (Lou et al., 2023), raw context, and synthetic context.

Scoring Model SciBERT and GPT2 are trained with each context. The cell context can be longer than the maximum input token length of BERT (512). Hence, GPT2, which allows more input tokens (1024), is used to capture all context. The detailed statistics of the token number of contexts are provided in Appendix C.

5.3.2 Result

Table 3 shows that the model fine-tuned with synthetic context demonstrates the highest performance in both @top1 and @top5 metrics compared to other conditions. The performance of GPT4 Turbo in zero-shot learning is lower than those of other fine-tuned models. When using the raw context with SciBERT, the @top1 accuracy is lower than the S2abEL, but the @top5 accuracy outperforms it. To compare SciBERT and GPT2 with raw context, GPT2 performs worse than SciBERT despite having larger parameters.

5.3.3 Error Analysis and Discussion

GPT4 zero-shot A drastic tendency was observed when GPT4 Turbo was prompted in a zero-shot manner to simultaneously choose attributed sources from both the cited references and the target papers. Depending on the prompt text, the GPT4 Turbo chooses them only from the target papers or only from cited references. To mitigate this issue, the prompt text was adjusted to choose the attributed source from the cited references first and then determine whether the concept was newly proposed. If the model determines it as a new concept, the attributed source from cited references is discarded, and the target paper is selected. Concrete examples of each are provided in Appendix B.

SciBERT vs GPT2 The raw context aims to extract sufficient information from the main text, which may include sentences with little relevance to the cell. This may explain why GPT2, which can use all the context, did not contribute to accuracy.

Synthetic Context Analysis The comparison of synthetic and raw context revealed that employing synthetic context improved accuracy in two ways: context completion and knowledge completion.

Context Completion Only the abbreviated name of a method or dataset is mentioned in the cell, but the full name and description are provided in the main text. For instance, a cell in a table is "ET Perp," and it refers to the "Perplexity achieved by the Evolved Transformer." However, the expression does not appear in the main text. Thus,

Cell Context	Scoring Model	accuracy@top1			accuracy@top5		
		all	method	dataset	all	method	dataset
GPT4 Turbo zero-shot	None	22.3	30.0	1.0	-	-	-
S2abEL (Lou et al., 2023)	SciBERT	49.3	55.3	29.8	63.2	64.9	53.7
Raw Context	SciBERT	45.0	45.3	44.7	67.3	65.6	69.1
Raw Context	GPT2	41.1	43.8	35.4	62.8	63.3	60.7
Synthetic Context	GPT2	**55.6**	**56.5**	**51.8**	**75.7**	**75.7**	**70.8**

Table 3: ASM Result of varying Cell Contexts and Scoring Model

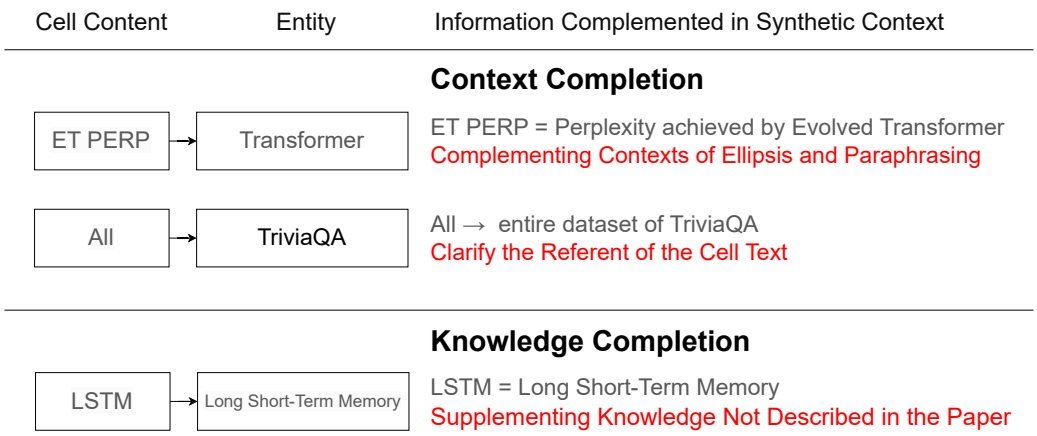

Figure 4: Completions in the synthetic contexts: There are two types of completions in synthetic contexts. **Context Completion**: LLM understands that the cell context is the abbreviated name and explains it in the synthetic context. Clarify the referent of the cell when the referent is ambiguous. **Knowledge Completion**: If the cell's content is not explained in the paper, LLM adds supplemental information.

baseline methods failed to link the cell to the Transformer correctly. In the synthetic context, the abbreviation context is explained by the LLM, and the cell is successfully linked to the correct entity.

We found that when the referent of a cell is ambiguous on its own, such as "baseline" and "All", the context supplements this by clarifying the referent in the synthetic context. Table 3 in the paper of Clark and Gardner (2018) contains a cell named "All." The synthetic context for this cell is *"[...]the term 'All' likely refers to the entire dataset of TriviaQA, which [...]"*, making the referent of the word "All" clear.

Knowledge Completion There may not be sufficient descriptions for well-known methods or datasets in the main text. For example, the concept of cell content "LSTM" might not be adequately explained despite many mentions of LSTM in the main text. Hence, previous methods misinterpret LSTM as a new concept. In the synthetic context data, LLM complemented the fact that LSTM stands for Long Short-Term Memory. And it allows the cell to successfully associate with the correct entity.

Errors in Synthetic Context On the other hand, errors were observed due to the misunderstanding injected in synthetic contexts. Specifically, in the paper of Bauer et al. (2018), the cell labeled *"Dev"* in Table 3 refers to a development set for evaluation. However, it is identified as a person name *"Devi Parikh,"* by LLM. The name is the author of a reference and included in the raw context but unrelated to the experiments in Table 3. This implies the raw context is not sufficient to identify essential information accurately.

5.4 Experimental Investigation of LLM for Synthetic Context

In this experiment, we measure the impact on task scores by various LLMs that generate synthetic contexts.

5.4.1 Experimental Settings

In previous experiments, GPT4 Turbo was used as the LLM. The quality of the synthetic context is believed to be influenced by the language com-

Generation Model	accuracy@top1			accuracy@top5		
	all	method	dataset	all	method	dataset
GPT4 Turbo	**55.6**	**56.5**	**51.8**	**75.7**	**75.7**	70.8
GPT3.5-16k	51.4	54.3	44.8	71.3	73.4	66.7
TULU v2 70B+DPO	54.0	55.8	49.2	74.5	74.6	**72.9**

Table 4: ASM Result of varying LLM models for synthetic context generation

prehension ability and specialized knowledge of relevant the domain of the LLMs. Therefore, in this experiment, we evaluate synthetic contexts generated by the following two models in addition to GPT4 Turbo. **GPT3.5 Turbo**: The training data of GPT3.5 Turbo is assumed to be similar to GPT4 Turbo. However, as the parameter size of it is smaller than GPT4-turbo, it has lower language comprehension and language refinement abilities. **TULU2 70B+DPO** (Ivison et al., 2023): An open-sourced model that continues learning from LLAMA2 70B using the instruction dataset and the direct preference optimization (DPO) algorithm (Rafailov et al., 2023). It demonstrates performance equivalent to GPT3.5-turbo in MT-Bench and AlpacaEval. Scientific documents in the machine learning field are included in the training dataset, such as SciERC (Luan et al., 2018) and Qasper (Dasigi et al., 2021). As a result, we expect it to have enough knowledge to perform knowledge completion.

5.4.2 Result

Table 4 shows that TULU v2 70B+DPO demonstrates higher accuracy than GPT3.5-16k and exhibits competitive performance with GPT4-Turbo. Regardless of the type of LLM, there is a consistent trend that method type cells have higher accuracy than dataset type cells. However, this trend is more pronounced in GPT3.5-16k.

5.4.3 Error Analysis and Discussion

We confirmed that TULU2 70B+DPO has knowledge about famous methods, datasets, and evaluation metrics and performs appropriate knowledge completion. This indicates that when the pretraining dataset for an LLM includes data related to the domain, knowledge completion can be expected to be effective in supplementing concepts, which are often abbreviated or omitted in a paper. Although an LLM used in our pipeline needs to be updated to learn newly introduced methods and concepts, frequent updates would not be necessarily required, as it takes some time for such concepts to become prevalent. We expect the knowledge update for LLMs can be partly resolved with retrieval-augmented generation.

Although we confirmed the knowledge of LLMs is effective for entity linking, we observed the outputs that are not based on facts, known as hallucinations (Ji et al., 2023). For example, a table cell in the paper of Zhong et al. (2019) is "CFC (ours)", and its attributed source is the paper itself. However, a part of the synthetic context generated by TULU2 70B+DPO is

"*The term "CFC (ours)" in the context of the scientific paper titled "modelname for Multi-evidence Question Answering" refers to a new question answering model[...]*", which refers to a non-existent paper title. Suppressing such hallucinations while leveraging the knowledge of LLMs is a challenge for future work.

6 Conclusion

In this study, we proposed new context extraction methods from the main text for entity linking of table cells of scientific papers. First, we propose a rule-based context extraction method (raw context) to collect broad context from a paper. Then, we introduce the synthesized data using an LLM to refine the raw context (synthetic context). By employing raw context and synthetic context, we improved the accuracy of entity linking by more than 10 points. In the qualitative analysis, we observe the LLM refines raw context by supplementing context and completing information.

7 Limitations

Application scope. Our entity linking method in this work depends on an existing KB. Thus, it cannot be applied to fields where KB does not exist. However, the @top1 accuracy for ASM tasks in section 5.3 can be considered linking table cells to papers, which allows linking to unknown concepts without depending on an existing KB. Therefore,

there is potential for application as the automatic construction of knowledge bases in domains where organized KB does not exist yet.

Practicality. Our proposed method outperforms the existing method and achieves 53.1% @top1 accuracy in EL, 55.6% for @top1, and 75.7% for @top5 in ASM. These results suggest that fully automating the linking of table cells in papers is still challenging; however, it could potentially be used to assist manual annotation for table cells.

Model bias. Synthetic context depends on LLMs' generative capabilities and knowledge, making it susceptible to the model's bias. This study targets only English-language papers in the machine learning domain, which may limit generalization to other languages and fields.

Model availability. The experiments in this study were conducted using OpenAI's GPT4 Turbo 1106, GPT3.5 16k, and TULU2 70B+DPO. GPT4 Turbo and GPT3.5 are accessible via the OpenAI API, but access may be lost in the future due to model version updates. Currently, these models are supported by the Azure OpenAI API.

Acknowledgements

This work was supported by JST Moonshot R&D Grant Number JPMJMS2236. This research is also supported by the Mercari R4D PhD Support Program.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*.

Yulin Chen, Zhenran Xu, Baotian Hu, and Min Zhang. 2023. Revisiting sparse retrieval for few-shot entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12801–12806, Singapore. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *ArXiv*, abs/2105.03011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Kyle Gorman, Christo Kirov, Brian Roark, and Richard Sproat. 2021. Structured abbreviation expansion in context. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 995–1005, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online. Association for Computational Linguistics.

Bogdan Kostić, Julian Risch, and Timo Möller. 2021. Multi-modal retrieval of tables and texts using tri-encoder models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 82–91, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2022. Improving candidate retrieval with entity profile generation for Wikidata entity linking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3696–3711, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik, and Doug Downey. 2023. S2abEL: A dataset for entity linking from scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3089–3101, Singapore. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt. `https://chat.openai.com`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *International conference on machine learning*, pages 5877–5886. PMLR.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the citation graph for scientific information extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–731, Online. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Sean Yang, Chris Tensmeyer, and Curtis Wigington. 2022. TELIN: Table entity LINker for extracting leaderboards from machine learning publications. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 20–25, Online. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, Nitish Shirish Keskar, and Richard Socher. 2019. Coarse-grain fine-grain coattention network for multi-evidence question answering. *ArXiv*, abs/1901.00603.

Yuchen Zhuang, Yinghao Li, Junyang Zhang, Yue Yu, Yingjun Mou, Xiang Chen, Le Song, and Chao Zhang. 2022. ReSel: N-ary relation extraction from scientific text and tables by learning to retrieve and select. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 730–744, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Prompt for Synthetic Context generation

Generate Synthetic Context as a summarization task in a zero-shot setting with an LLM. Synthetic context data generation was performed using GPT4-Turbo-1106-preview, GPT3.5-turbo-16k, and TULU2-70B DPO. The parameters during generation, temperature, and top-k were set to 0 to stabilize the generation. The input to the GPT4-Turbo-1106-preview and GPT3.5-turbo-16k models follows a format that embeds the text of the cell (CELL_CONTENT), the paper title (PAPER_TITLE), the abstract (PAPER_ABSTRACT), and the paper context information (PAPER_CONTEXT). The total number of prompt tokens to generate all synthetic context is 1182k, and for completion tokens, it is 1695k. Hence, generating all synthetic context with GPT4-Turbo 1106 costs about $170.

prompts

system_prompt : *You are a researcher in the field of machine learning. You are provided with a word that appears in a certain paper and information in the paper related to that word. Please explain the word based on the information provided.*

user_prompt : *Please explain the word {CELL_CONTENT}. The title of the paper in which this word appears is "{PAPER_TITLE}", and the abstract is "{PAPER_ABSTRUCT}". The category of this word is {CELL_TYPE}. The relevant descriptions in the text are written below. {PAPER_CONTEXT} Please provide your answer as concisely as possible.*

B GPT4 Turbo zero shot-learning prompt

In §5.3.1 of the experiment, GPT4 Turbo was evaluated directly in a zero-shot setting for Attributed Source Matching. Specifically, the main text information of the paper and the titles and abstracts of all cited references were embedded into the following template as input. As output, the ID of the cited reference that serves as the source and a flag indicating whether it represents a novel concept proposed in the paper were obtained. If it is determined that the flag represents a concept proposed in the paper, the source is cited as SourcePaper without using the cited reference ID.

prompts

system prompt : *You are tasked with identifying the source reference of the concept indicated by the cell text in a table within a machine learning academic paper. This paper is referred to as the "Source Paper" and its cited literature as "Reference Papers". The concept indicated by the cell text in the table is either a dataset or a method, which was proposed either in the cited literature. Your task is to estimate the paper in which this concept was proposed. For making your estimation, you will be provided with the cell text of the table, the type of concept that the cell text of the table is indicating, the caption of the respective table, and descriptions in the "SourcePaper" that are relevant to the respective table. You will also be presented with potential choices which include the title and abstract each of the cited literature. Please make a selection from these options. Your response should be in the following JSON format: { "estimate_result": "ID of a ReferencePaper", "is_source": "True or False" } Please input that ReferencePaper's ID into the estimate_result field. Also, if you believe that the content indicated by the cell text in the table is something newly proposed in the SourcePaper, please enter True in the is_source field.*

C Cell Contexts Statistics

We compare the statistics of the number of tokens for the input to the model used in S2abEL and the raw context and synthetic context used in this study. Table 5 shows the mean, standard deviation, maximum, and minimum number of tokens for the entire data and the mean and standard deviation of the number of tokens when the Cell type is method or dataset. Comparing the features of S2abEL and raw context, the raw context tends to have a smaller average number of tokens and a larger standard deviation. This is because, in S2abEL, information about the position of the table and surrounding cells was used as input. In contrast, in the raw context of this study, sentences that mention the table or cells in the captions or main text are added. As a result, the number of tokens varies significantly depending on the mention in the main text, leading to a larger standard deviation. The synthetic context summarizes and complements the raw context and

Name	all				method		dataset	
	average	std	max	min	average	std	average	std
S2abEL	522.6	221.1	1292	49	569.2	220.9	434.1	192.6
Raw Context	510.2	282.7	5044	29	499.4	268.8	530.9	306.5
Synthetic Context	385.6	98.6	2121	65	394.1	98.2	369.3	97.2

Table 5: Cell Contexts Token number Statistics

consistently has fewer tokens. Furthermore, when there are many mentions in the main text, only necessary information is extracted, and when there are no mentions, information is supplemented, leading to a significantly smaller standard deviation.

D Training Details

We trained all models using the AdamW optimizer (Loshchilov and Hutter, 2019) with linear decay warm-up. All models were trained using a single 48Gb NVIDIA A6000 GPU. To train GPT-2 for ASM, We added a short prompt before cell contexts. The prompt is "Given a table cell text from an academic paper in the field of Machine Learning, classify whether the information in the cell originates from provided cited literature or other. The reference information is". For evaluation, cross-validation is conducted on 10 paper categories within the S2abEL dataset. From S2abEL dataset, a category is selected for testing, another category is randomly selected for validation, and other categories are used for training.

parameter	GPT-2	SciBERT
learning rate	2e-5	2e-5
batch size	16	32
max token length	1024	512
epoch size	2	2
warm-up ratio	10%	10%

Table 6: Training Hyperparameters

E Detailed Entity Linking Result

Cross-validation is conducted on 10 categories of paper within the S2abEL dataset in the Entity Linking experiments. The results across all folds are presented.

Test fold	Overall acc.	OutKB F1	InKB hit@1	Overall acc.	OutKB F1	InKB hit@1	Overall acc.	OutKB F1	InKB hit@1
img_gen	48.3	55.6	26.7	53.8	57.1	37.4	48.6	46.9	43.1
misc	71.3	83.2	1.2	80.1	87.8	37.8	86.5	92.0	74.4
mt	49.2	60.6	22.5	50.0	59.9	21.2	62.2	68.8	39.4
nli	61.4	73.4	26.6	66.4	76.9	36.6	64.8	72.7	52.3
object_det	31.2	36.8	15.8	64.7	60.5	58.7	65.1	73.6	59.5
pose_estim	65.8	77.3	29.6	84.2	95.8	67.6	70.7	82.7	41.7
qa	73.7	84.4	22.6	82.3	90.2	52.2	82.5	89.6	51.7
sem_seg	63.2	73.3	49.5	67.7	66.3	55.0	76.7	73.7	68.8
speech_rec	69.0	79.3	28.1	67.2	78.0	35.7	76.9	83.0	50.2
text_class	68.8	79.0	30.6	73.0	82.6	45.1	73.8	83.0	41.4
average	60.2	70.3	25.3	68.9	75.5	44.7	70.8	76.6	52.2