LLM による合成文脈データを用いた表のエンティティリンキング

大嶋 悠司 [1,2] 進藤 裕之 [1] 寺西 裕紀 [3,1] 大内 啓樹 [1,3] 渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 株式会社メルカリ [3] 理化学研究所
{oshima.yuji.ov6, shindo, hiroki.ouchi, taro}@is.naist.jp
hiroki.teranishi@riken.jp

概要

論文の表には実験結果などの重要な情報が含まれるため，表を解析して知識ベースと紐づけるエンティティリンキングは有用な技術と期待されている．このタスクでは表のセルを読解して，それが指す知識ベースのエンティティと紐づけるには，幅広い文脈理解が必要になる．しかしながら，論文の本文からセルの文脈を適切に抽出する必要があるという課題があった．本研究では，大規模言語モデルを活用し，セルの文脈を生成する合成文脈を提案する．このアプローチにより，既存手法よりリンキング精度が 5 ポイント以上向上することを実証した．また，合成文脈は場合によって論文には記述されていない補助知識も補完されることを明らかにした．

1 はじめに

科学技術論文の情報解析は論文検索や読解補助，自動知識ベースの構築など科学を加速させる多くの応用が期待される分野である．特に情報科学技術分野の論文では，実験結果や実験に使用されたデータセット・タスク・評価指標などの重要な情報が論文中の表に記載されるため，表の情報解析が重要視される．そのため情報科学論文の表に記載される情報を知識ベースとリンキングするエンティティリンキング（EL）は有用な技術と期待され，手法やデータセットが提案されている [1, 2, 3]．

S2abEL[3] は，情報科学論文の表に対する EL のための大規模な評価データセットである．これは表のセルを入力として，その文字列が意味する Paper with Code（PwC）のエンティティを出力とするデータセットである．例えば表に Transformer という記述があった場合に，PwC 内の Transformer エンティティと紐付けることが正解となる．モデルは本文中

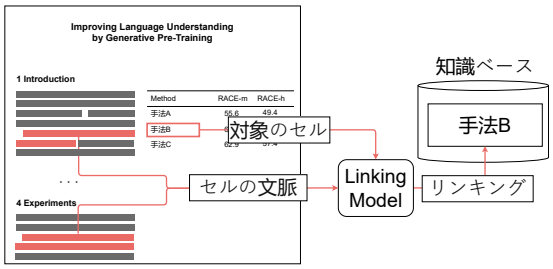

図 1 論文の表のセルに対するリンキング

や引用文献から Transformer の概念を理解し，知識ベース内を検索する必要がある．しかし，本文中からセルの文脈を抽出することが困難な場合がある．例えば，表のセルに手法の省略された名称のみ記載されるが，本文中にはその手法の正式名称が記載されている場合にそれらが同一のものであるとモデルが判断できない場合がある．また，例えば Transformer はよく知られた手法であるため，本文中にはその説明が記載されていない場合がある．しかし，これらの表のセルに関連する本文の記述の抽出や補助的な情報の追加を人手で行うことは高いコストがかかる．

ここで本研究では機械学習モデルを使ったデータ拡張手法の一つである合成データを用いてこの課題を解決することを提案する．特に ChatGPT[4] や LLaMa2[5] などの大規模言語モデル（LLM）には高い言語理解能力があり，様々なタスクのためにデータを生成できる．本研究では，LLM を利用して特定のセルに関連する論文中の文脈を抽出したものを合成文脈データと呼ぶ．LLM に論文の本文を与え，リンキングしたい単語を中心に論文の要約を行うことでその単語の文脈を抽出することが可能である．LLM による合成文脈データを用いることで，提案手法は既存手法と比較して 5 ポイント以上高い精度を達成した．また合成文脈データとルールベースによる特徴量を比較することで，合成文脈データが文

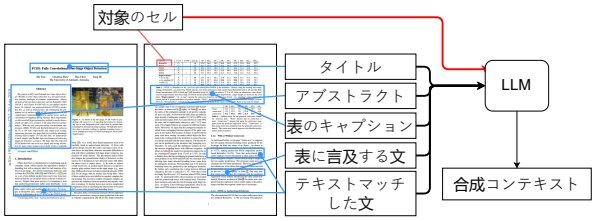

図 2　合成文脈データの生成：特定のセルに関する記述（本文文脈情報）が LLM に与えられ，対象セルの内容を LLM に説明させることで本文から対象セルに関するコンテキストが抽出される．

脈補完・知識の追加といった形でリンキングに必要な情報を効果的に抽出・補完していることを明らかにした．

2　エンティティリンキング

エンティティリンキング (EL) では，論文中の表における各セルを知識ベース（PwC）内のエンティティに紐づけることを目的とする．また該当するエンティティが存在しない場合は知識外を示す “ontKB” とする必要がある．S2abEL ではこのタスクを，以下の 4 つのサブタスクに分割している．(i) **セルの分類：**表のセルを method, dataset, metric, dataset&metric, other の 5 つのタイプに分類するタスク．(ii) **出典引用文献の抽出：**セルの出典となる引用文献を特定するタスク．(iii) **エンティティ候補の検索：**表のセルに関連するエンティティを知識ベースから検索し，リンキングされるエンティティの候補を抽出するタスク．(iv) **エンティティ選択：**エンティティの候補から正しいエンティティを選択もしくは outKB を推定するタスク．これらサブタスクのうちセルタイプ分類は，先行研究において分類精度が 90%を超えており，分類の推定を正解データに置き換えても EL の精度は 1 ポイント以下しか変化しない．そのため本研究では対象とせず，セルのタイプは正解データを用いて後続のサブタスクを行う．

3　手法

3.1　合成文脈データ生成

本研究では，合成文脈データを LLM で生成し，セルの文脈情報として利用することを提案する．これまでの LLM を使ったデータ生成アプローチは，疑似ラベルを生成したり [6, 7]，学習データを増加させる目的で使用されている [8]．しかし，本研究ではタスクを解くための重要な文脈を補完することで学習データの質を向上させることを目的とする．本研究における合成文脈データ生成のフローを図 2 に示す．

本文文脈情報　LLM への入力に用いられる本文中のセルに関連する記述を本文文脈情報と定義する．本文文脈情報はタイトル，アブストラクト，表のキャプション，本文中の表を引用する文，及び本文中のセルのテキストを含む文の 5 つで構成される．

この本文文脈情報を入力として，対象セルを中心に要約することで合成文脈を生成する．合成文脈では本文文脈の中で対象セルに関連する部分だけが抽出されることに加えて，本文文脈に不足の情報があり LLM がそれに関する知識を持つ場合にその情報が補完されることが期待される．本研究では LLM として OpenAI gpt3.5-turbo-16k を利用する．

3.2　リンキング手法

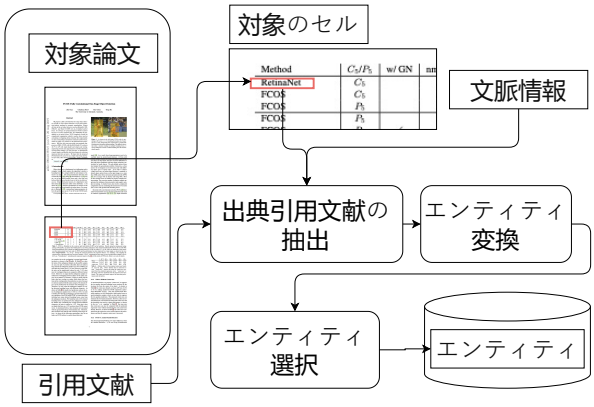

図 3　エンティティリンキング：論文本文，引用文献及び文脈情報を入力として特定の表のセルと紐づくエンティティを知識ベースから検索する．3 つのサブタスクからなる．

本研究での EL の手法を図 3 に示す．EL を構成するそれぞれのサブタスクにおける手法を説明する．

出典引用文献の抽出　：対象論文における対象のセルと引用文献との関連性を測定するために，セルの文脈情報と 1 つの引用文献のタイトル・アブストラクトを入力とする二値分類を行う．モデルとして GPT2 [9] に線形出力層を組み合わせて用いる．正解ラベルはセルの出典となる引用文献には 1，それ以外は 0 として，Binary Cross Entropy loss を用いて GPT2 を含むパラメータ全体を学習する．モデルの出力は入力された文献とセルの関連性を表すスコアとなる．またセルの内容が新規に提案された概念である場合もあるため，対象論文も同様にモデルに入

表 1　エンティティリンキングの実験結果

手法	overall	OutKB F1	InKB acc
S2abEL (再実装)	0.605	0.715	0.268
提案手法	**0.661**	**0.764**	**0.373**
S2abEL with DR([3])	0.58.2	0.714	0.334

力される．対象論文出典としてが選択された場合はセルの内容が論文中で提案された概念であると判断される．ただし，セルに引用文献番号などが付与されており，出典となる引用文献が明示的に示されている場合はその引用文献を選択する．

エンティティの変換　：出典引用文献抽出のスコアを用いて出典の可能性の高い引用文献の上位 k 個を得て，知識ベースを用いてそれぞれに紐づくエンティティに変換する．S2abEL では出典引用文献抽出の結果に加えて，Direct Retrieval(DR) というセルのテキストを知識ベースから直接検索した結果も利用する．しかし出典引用文献候補だけで理想的にはエンティティ候補として 90%以上の recall を得られることが S2abEL の実験結果から分かっているため，本研究では DR は用いない．

エンティティ選択　：エンティティ選択モデルの学習は S2abEL の手法を踏襲し，SciBERT を用いたクロスエンコーダアーキテクチャ [10] をファインチューニングする．訓練されたモデルは，セルがエンティティと紐づくスコアを出力し，最も高いスコアのエンティティのスコアが 0.5 未満である場合，そのセルは outKB とする．

4　実験

4.1　実験設定

S2abEL データセットを用いて EL の評価を行う．評価指標は，先行研究と同様に overall accuracy, OutKB F1, InKB accuracy を用いる．出典引用文献の抽出の学習に合成文脈データを用いて学習したものと S2abEL の再現実装，及び S2abEL 論文での結果を比較する．ただし S2abEL の再現実装では先行研究と異なり，DR を用いずに学習される．また ED の推定時のエンティティ候補数は，先行研究と同様に k=50 とする．

4.2　実験結果

表 1 に EL 実験の結果を示す．合成文脈を用いた提案手法は S2abEL の再現実装と比較しても精度が OutKB F1 スコアが 5 ポイント，InKB accuracy は 10 ポイント以上向上していることがわかる．また S2abEL の論文中の Direct Retrieval (DR) も追加した結果と比較しても，OutKB F1 スコアが 5 ポイント，InKB accuracy は 4 ポイント向上している．提案手法は特に InKB accuracy が向上しており，出典引用文献の抽出の精度が向上したことで正解のエンティティを候補に含めることに成功していることがわかる．

5　出典引用文献の抽出実験

5.1　タスク定義と手法

EL 実験から出典引用文献の抽出の精度向上がリンキング精度に寄与することが分かった．そこで合成文脈データの効果を直接測るため，EL のサブタスクである出典引用文献の抽出においてセルのコンテキスト情報の変化が精度に与える影響を観測する．ただし，EL 実験ではセルの出典が明示されている場合はその情報を利用したが，本実験ではその情報を利用せずすべてのセルの出典を推定する．

5.2　実験設定

本実験では以下の 6 つの手法を比較する．

ベースライン　：LLM の zero shot learning によって出典となる引用文献を抽出する．モデルは OpenAI gpt-4-1106-preview を用いる．セルの生コンテキスト情報と全ての引用文献のタイトル・アブストラクトが与えられ，出典となる引用文献を選択する．

S2abEL　：S2abEL の再現実装を行い，出典引用文献の抽出タスクのみ行う．

合成文脈　：LLM による合成文脈データを入力としてモデルが学習される．

ChatGPT による知識補完　：gpt3.5-turbo-16k に文脈情報を与えず，セルのテキストのみを入力し gpt3.5 にセルの内容を説明させることで LLM の関連知識を抽出し，それを本文文脈情報に加える．

本文文脈　：本文文脈情報すべてをモデルの入力として与える．

結果は top1 accuracy 及び top5 accuracy で評価する．ただし GPT4 zeroshot はスコア計算を行わない

表 2　出典引用文献の抽出の実験結果

手法	accuracy@top1			accuracy@top5		
	all	method	dataset	all	method	dataset
GPT4 zeroshot	0.223	0.300	0.010	-	-	-
S2abel(再実装)	0.419	0.505	0.218	0.551	0.606	0.420
合成文脈	**0.536**	**0.565**	**0.465**	**0.736**	**0.742**	**0.702**
ChatGPT による知識補完	0.434	0.456	0.386	0.639	0.637	0.635
本文文脈	0.403	0.431	0.334	0.606	0.603	0.603

ため，top1 accuracy のみで評価する．またセルには method, dataset の 2 種類があるため，それぞれについても accuracy を計算する．

5.3　実験結果

出典引用文献の抽出実験の結果を表 2 に示す．結果から合成文脈を用いた手法が top1, top5 いずれにおいても最も高い結果であることがわかる．合成文脈は本文文脈，S2abEL と比較して all@top1 ともに all@top5 で 10 ポイント以上スコアが向上している．GPT4 zeroshot は学習を行ったいずれの手法よりも著しくスコアが低く，zeroshot では困難な課題であることがわかる．知識補完による手法は本文文脈と比較して all@top1, all@top5 ともに 3 ポイント程度精度向上しており，外部知識の LLM による補完が有効であることがわかる．また合成文脈が知識補完よりも高い精度であったことから，LLM のコンテキスト情報を要約する能力も重要であることがわかる．

6　合成文脈データの効果分析

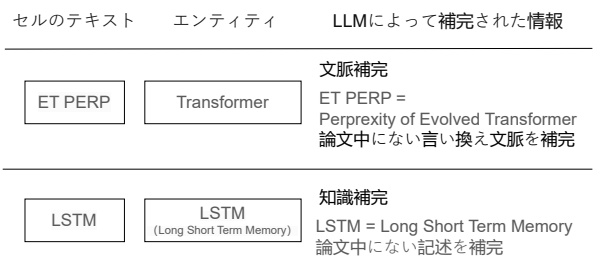

図 4　合成文脈データ分析：文脈補完・知識補完によってリンキング精度向上に寄与

合成文脈データとを用いることで精度が改善されたサンプルの具体例を示す．**文脈補完**：セルには手法やデータセットの省略名称のみ記載されるが，本文中に正式名称を用いてその説明が記載される．例えばある表のセルに “ET Perp” と記載される．これは “Perprexity of Evolved Transformer” という意味であるが，“ET Prep” という表現自体は本文中に出てこない．そのため既存手法ではセルを正しく Evolved Transformer に結び付けられなかった．合成文脈データでは，LLM がセルの内容が省略名称であることを理解し，正式名称，省略名称および概念の説明をコンテキストとして全て記載したためリンキングに成功した．**知識補完**：著名な手法やデータセットの場合，本文中に十分な説明がないことがある．例えばセルの内容が LSTM の場合，本文中に LSTM の記述は存在するがその概念の説明がない場合がある．そのため既存手法では LSTM を新規概念だと判断してしまった．合成文脈データでは，LSTM は Long Short-Term Memory を意味するモデルの一種であることが LLM によって補完されたため，正しい引用文献に紐づけることができた．

7　おわりに

今研究では科学技術論文の表セルに対する EL タスクに対し，合成文脈データを適用した．合成文脈データはセルの情報を中心に論文を要約することでセルに関連する文脈を抽出する手法である．この適用により既存研究と比較して精度が 5 ポイント以上改善した．また詳細な分析により，本文の内容を要約するだけでなく LLM が知識を保管することによる精度改善に寄与していることも明らかにした．本実験では LLM として ChatGPT3.5-turbo-16k を用いたが，より大規模なモデルや特定のドメインに特化したモデルでの評価は今後の課題である．

謝辞

本研究は，社会人博士支援制度「mercari R4D PhD Support Program」の支援により実施しています．
また本研究は，JST ムーンショット型研究開発事業 (JPMJMS2236) の支援を受けたものです．

参考文献

[1] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. AxCell: Automatic extraction of results from machine learning papers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8580–8594, Online, November 2020. Association for Computational Linguistics.

[2] Sean Yang, Chris Tensmeyer, and Curtis Wigington. TELIN: Table entity LINker for extracting leaderboards from machine learning publications. In Tirthankar Ghosal, Sergi Blanco-Cuaresma, Alberto Accomazzi, Robert M. Patton, Felix Grezes, and Thomas Allen, editors, **Proceedings of the first Workshop on Information Extraction from Scientific Publications**, pp. 20–25, Online, November 2022. Association for Computational Linguistics.

[3] Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik, and Doug Downey. S2abEL: A dataset for entity linking from scientific tables. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3089–3101, Singapore, December 2023. Association for Computational Linguistics.

[4] OpenAI. Chatgpt, 2023. https://chat.openai.com.

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

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

[7] Chenxi Whitehouse, Monojit Choudhury, and Alham Aji. LLM-powered data augmentation for enhanced cross-lingual performance. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 671–686, Singapore, December 2023. Association for Computational Linguistics.

[8] Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 1555–1574, Singapore, December 2023. Association for Computational Linguistics.

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

A　合成文脈データ生成

合成文脈データ生成は ChatGPT3.5-turbo-16k を用いて行った．モデルへの入力は以下のフォーマットにセルのテキスト（CELL_CONTENT），論文タイトル（PAPER_TITLE), アブストラクト (PAPER_ABSTRUCT)，及び本文脈情報（PAPER_CONTEXT）を埋め込む．

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　モデルハイパーパラメータ

実験に利用した出典引用文献の抽出モデルのハイパーパラメータを記載する．

表 3　出典引用文献の抽出モデルのハイパーパラメータ

parameter	value
pretrained model	gpt2
learning rate	2e-5
batch size	16
max token length	1024

C　ChatGPT zero shot-learning

5 章の実験において，出典引用文献の抽出タスクに対して GPT4 を zero-shot で用いた．具体的には以下の文章に論文の本文情報と全引用文献のタイトル・アブストラクトを埋め込んで入力する．出力として出典となる引用文献の id と，もしそれが論文中で提案された新規概念かどうかのフラグを取得する．論文中で提案されたフラグだと判断される場合は引用文献 id を使わず SourcePaper を出典とする．

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.*

user prompt *:*

- *Cell Text: {CELL_CONTENT}*
- *Concept Type: {CELL_TYPE}*
- *Table Caption: {TABLE_CAPTION}*
- *Paper Contents related to the Table: {table_context}*

Please make a selection from the following options.
Source Paper: "{PAPER_TITLE}", "{PAPER_ABSTRUCT}"
Cited Papers: ID: "{PAPER_TITLE}", "{PAPER_ABSTRUCT}"
⋮