大規模言語モデルを用いた専門用語間の関係性解析

岩熊耕平　難波英嗣　福田悟志
中央大学大学院理工学研究科

概要

用語間の関係は、情報検索や文書作成など、さまざまなタスクに活用できる重要な情報源である。従来の関係性解析手法として定型表現による抽出があるが、低コストである一方、抽出精度に課題があった。埋め込み表現を用いた教師あり学習による解析手法も提案されているが、多くは汎用的に使用される用語を対象とするデータセットを利用しており、特許などの専門的な文書に含まれる用語の特徴を十分に捉えることが困難である。そこで本研究では、大規模言語モデル（LLM）を活用した専門用語間の関係性解析手法を提案する。特許用語を対象とした英語データセットである Google Patent Phrase Similarity Dataset を利用し、専門用語の文脈的意味を捉えた関係性解析を行う。

1 はじめに

自然言語処理において、上位下位関係や部分全体関係をはじめとする用語間の関係性は、重要な情報源である。用語間の階層関係は、情報検索や文書作成など、さまざまな下流タスクに応用可能である。しかし、人手による用語関係の解析やオントロジーの構築には多大な時間と労力が必要であり、これらの自動化が強く求められている。テキストデータベースから上位下位関係を構築する代表的な手法として、定型表現を活用した手法が挙げられる。この手法では、「A 等の B」などの定型表現に着目し、B の下位概念として A を抽出する。この手法は上位下位関係を漏れなく抽出するのに有効であるものの、抽出結果に一定数の誤りが含まれることや、多義語に対応できないことが課題として指摘されている。分散表現を用いた教師あり学習による手法も提案されているが、従来のベンチマークでは一般的な用語を対象としたものが中心であり、専門的な用語を対象とした場合に十分な性能を発揮できないという課題が存在する。

近年、多言語コーパスを用いて学習された LLM（大規模言語モデル）が提案されている。これらは汎用的な語と同様に、専門的な用語に関する知識も有しているため、LLM を活用することで文脈的意味に基づいた用語分析が可能になることが期待される。そこで本研究では、LLM および埋め込みモデルを活用した専門用語間の関係性を自動で解析する手法を提案する。第一に、埋め込みモデルを用いて用語間の埋め込み表現を取得し、それらのベクトル類似度を算出することで用語間の意味的類似性を推測する。第二に、LLM を用いて用語間の意味的関係を出力することで、専門的知識を十分に考慮した関係性を直接的に推論する。さらに、LLM を Fine-tuning することで専門的用語に関する知識を補完し、関係性推論能力の向上を目指す。

2 関連研究

Hearst ら[1,2]は、用語の上位下位関係をテキストから自動的に取得できる低コストな語彙関係抽出手法を提案している。上位下位関係を表す語彙および構文パターンをリスト化することで、テキストコーパスから上位下位関係を持つ用語対を抽出する。安藤ら[3]は、定型表現を用いて新聞記事から名詞の下位概念を抽出する手法を提案している。「などの」をはじめとする定型表現を活用し、上位語となる単語を規定し、それに関連する下位語となる名詞を自動的に抽出する。これらのルールベースの自動抽出手法は、設計が容易で漏れなく抽出できるという利点があり、関係性推論において広く用いられている。しかし、抽出結果に誤った関係性が含まれる点が課題である。

これに対し、用語の埋め込み表現を取得することで、文脈的意味を考慮した関係性抽出を行う手法が提案されている。取得した埋め込み表現を用いて教師あり学習を行うことで、より正確な関係性の推測を目的としている。Jana ら[4]は、埋め込み表現を用いた同義語分類手法を提案している。近しい意味を持つ用語が共起しやすい傾向に着目し、共起回数に

基づいて用語の埋め込み表現を取得し、得られた分布から用語間の距離に基づいて同義語の判定を行う。また、教師あり学習と定型表現を組み合わせた手法も提案されている。Liu ら[5]は、BERT を用いて定型表現の一部を MASK トークン化し、上位下位関係を抽出する教師あり学習手法を提案している。このような手法では、網羅的な抽出を行い、分類器を構築することで誤った表現を削除できる。これらの教師あり学習手法は、学習データとして意味的関係性や類似性に関するベンチマークを必要とする。汎用的な用語に関するデータセットはいくつか存在するが、専門的な用語を対象としたものは数が限られている。

Aslanyan ら[6]は、特許文書に含まれる技術フレーズに焦点を当てた意味的関係性データセットとして、Google Patent Phrase Similarity Dataset を提案している。このデータセットは、共同特許分類（CPC）による文脈的意味を持つ約50,000件の評価済み用語ペアで構成されている。対象言語は英語であり、二人の評価者によって作成された。同義語や上位下位関係、部分全体関係など、用語間の関係性を示す評価クラスに加え、関係性に基づく類似度スコアも含まれている。表 1 に、関係性クラスと類似度スコアの対応関係を示す。本研究では、このデータセットを用いて LLM の Fine-tuning を行い、特許用語を対象とした用語間の関係性を推測することを目的とする。

表 1　用語対の関係性・類似スコアの対応関係

関係性	類似スコア
Very Highly related	1.00
Highly related	0.75
Hyponym (broad-narrow match).	0.50
Hypernym (narrow-broad match).	0.50
Structural match.	0.50
Antonym.	0.25
Meronym (a part of).	0.25
Holonym (a whole of).	0.25
Other high level domain match.	0.25
Not related.	0.00

3 提案手法

本研究では、2 種類の手法を用いて用語間の関係性を解析する。3.1 節では、埋め込みモデルを用いて用語間の類似度を算出する手法を提案する。3.2 節では、LLM を用いた直接的な関係性推論手法を提案する。

3.1 埋め込みモデルによる類似度算出

本研究では、学習済みの埋め込みモデルを用いて用語間の意味的類似度を推測する。図 1 に類似度推論のシステム例を示す。まず、用語対を埋め込みモデルに入力し、各用語に対する埋め込み表現を取得する。その後、埋め込み表現のベクトル間のコサイン類似度を算出し、それを用語間の類似度として採用する。類似度は-1 から 1 の範囲で算出され、値が 1 に近づくほど用語対は意味的に類似していることを示している。

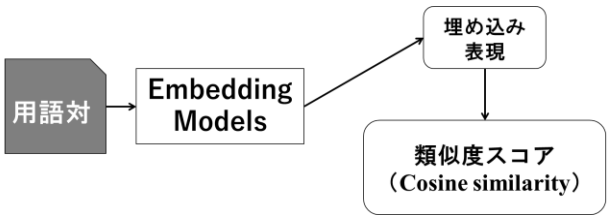

図 1　埋め込みモデルによる類似度推論

3.2 LLM による用語間関係性推論

3.1 節では、用語対の埋め込み表現がどの程度類似しているかに基づいて、意味的な類似性を推測している。それに対し、本節の手順では、LLM を用いることで用語対の関係性を直接的に推論する。図 2 に、LLM による関係性推測のシステム例を示す。

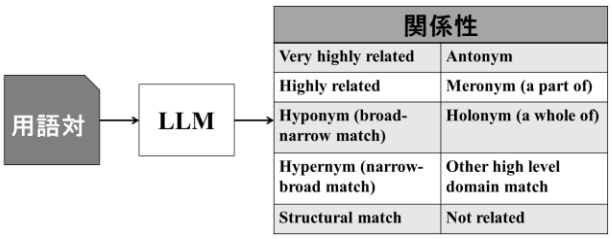

図 2　LLM による関係性推論

具体的には、用語対を LLM への入力として、該

当する関係性を推測する。出力される関係性は、図2に示される10種類を用いる。推論の際、出力する関係性は1つのみとする。図3に、LLMを用いて関係性を推測するためのプロンプト例を示す。

```
Based on ' reading machine', what is the relationship of ' photocopier'? Please choose the most appropriate option from the following:
1: 'Not related.'
2: 'Other high level domain match.'
3: 'Holonym (a whole of).'
4: 'Meronym (a part of).'
5: 'Antonym.'
6: 'Structural match.'
7: 'Hypernym (narrow-broad match).'
8: 'Hyponym (broad-narrow match).'
9: 'Highly related.'
10: 'Very highly related.'
```

図 3　関係性推論のプロンプト例

また、本研究では特許に含まれるような専門的な用語を対象とするため、事前学習のみでは関係性推論においてLLMが十分な性能を発揮しないと考えられる。この問題に対して、本研究ではLLMのFine-tuningによって対処する。学習データには、図3に示したプロンプトに加え、正解となる用語対の関係性が含まれている。これにより、専門的な用語に関する知識を補完し、LLMを関係性推論のタスクに最適化させることで、より正確な意味的関係性の出力を目指す。

4 実験

4.1 実験条件

実験データ

本研究では、Google Patent Phrase Similarity Datasetに含まれる特許用語対を対象とする。データセットのうち36,473件を学習用データとし、9,232件を評価用データとする。

比較方法

本実験では、モデルが出力した類似スコアが、評価用データの類似スコアとどの程度相関しているかを比較する。以下の埋め込みモデルおよびLLMを用いて、類似スコアを算出した。

- OpenAI Embeddings (text-embedding-3-large)
- E5[7] (multilingual-e5-large)
- GPT-4o (gpt-4o-2024-08-06)
- GPT-4o mini (gpt-4o-mini-2024-07-18)

また、ベースラインとしてAslanyanらによる以下のモデルにおける評価結果と比較を行った。

- Word2Vec [8](TensorFlow HubのWiki-words-250を使用)
- BERT [9](TensorFlow HubのBERT-Largeを使用)
- Sentence-BERT [10](all-mpnet-base-v2 事前学習済みモデルを使用)

評価尺度

評価指標には、ピアソン積率相関係数およびスピアマン順位相関係数を用いる。なお、埋め込みモデルでは用語間の類似度を算出し、類似スコアとの相関を評価するが、LLMでは関係性を出力とするため、用語間のベクトル類似度を直接算出することはしない。実際には、出力された関係性に対して、表1に示した関係性と類似スコアの対応関係に基づき、用語間の類似スコアを算出する。その後、算出した類似スコアと評価用データに含まれる正しい類似スコアとの相関を評価する。

4.2 実験結果及び考察

評価結果を表2に示す。提案手法であるGPT-4oおよびGPT-4o miniのFine-tuningモデルは、既存の埋め込みモデルであるBERTやWord2Vecと比較して、ピアソン積率相関係数およびスピアマン順位相関係数の両方で上回る結果となった。このことから、GPT-4oおよびGPT-4o miniによる用語間の関係性推論は、意味的類似性を十分に考慮できることが示された。GPT-4oとGPT-4o miniでは、前者の方が優れた性能を発揮したが、両者の間に大きな差は見られなかった。また、Fine-tuning後のGPT-4oとGPT-4o miniの性能差は、事前学習段階における性能差と比較して大幅に縮小した。

提案手法の埋め込みモデルである OpenAI Embeddings および E5 は、既存手法の埋め込みモデルと比較して劇的な性能向上は見られなかった。原因としては、OpenAI Embeddings および E5 から取得した用語対の埋め込み表現の類似度が高い傾向にあり、類似スコアとの相関に影響を与えたと考えられる。

表 2　類似度スコアの評価結果

Model	**Pearson cor.**	**Spearman cor.**
Word2Vec [6]	0.437	0.483
BERT [6]	0.418	0.409
Patent-BERT [6]	0.528	0.535
Sentence-BERT [6]	0.598	0.535
GPT-4o	0.505	0.514
GPT-4o mini	0.371	0.403
OpenAI Embeddings	0.581	0.564
E5 Embeddings	0.574	0.546
GPT-4o (Fine-tuning)	**0.762**	**0.738**
GPT-4o mini (Fine-tuning)	0.742	0.718

本研究における LLM の Fine-tuning の有効性を検証するため、Fine-tuning による類似度スコアの変動を分析した。表 3 にその結果を示す。GPT-4o では 42%、GPT-4o mini では 52%の用語対に関して類似度スコアが改善された。Fine-tuning 前後で用語対の類似度算出分布を比較すると、正しい類似度スコアと判定されるものが増加していた。また、予測したスコアと実際のスコアの差が減少し、より近しい意味的類似性だと判定される用語対が増加したことが分かった。これにより、Fine-tuning を行うことで、LLM が持つ専門的な用語の知識を補完し、関係性推論のタスクに対してより高い性能を発揮することが示された。

表 3　Fine-tuning による類似度スコアの変動

	Improve	Same	impair
GPT-4o	3899 (0.422)	4335 (0.470)	998 (0.108)
GPT-4o mini	4806 (0.521)	3520 (0.381)	906 (0.098)

5 結論

本研究では、専門用語間の意味的関係性を解析するため、Google Patent Phrase Similarity Dataset に含まれる用語対に対して、埋め込みモデルを用いて意味的類似度を算出し、LLM によって関係性を推測する手法を提案した。埋め込みモデルとして OpenAI Embeddings および E5 を、LLM として GPT-4o および GPT-4o mini を用いた。

実験の結果、LLM の Fine-tuning 済みモデルは、既存の埋め込みモデルと比較して、類似度スコアのピアソン積率相関係数およびスピアマン順位相関係数ともに上回ることが示された。これにより、LLM による関係性推論は意味的類似性を十分に考慮しており、Fine-tuning が専門的な用語の知識を補完し、関係性推論をより容易にすることが分かった。

本研究では Google Patent Phrase Similarity Dataset の用語対を用いたが、日本語特許などの他言語にも同様に専門用語に対応することが求められる。定型表現による関係抽出などを活用し、複数言語を用いて関係性解析の精度を向上させることが今後の課題として挙げられる。

参考文献

1. Marti A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the 14th International Conference on Computational Linguistics, p. 539-545, 1992.
2. Stephen Roller, Douwe Kiela, and Maximilian Nickel, Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, p.358-363, 2018.
3. 安藤まや, 関根聡, 石崎俊, 定型表現を利用した新聞記事からの下位概念単語の自動抽出, 情報処理学会研究報告. 情報学基礎研究会報告 72, 77-82, 2003.
4. Abhik Jana, Nikhil Reddy Varimalla, and Pawan Goyal, Using Distributional Thesaurus Embedding for Co-hyponymy Detection, Proceedings of the Twelfth Language Resources and Evaluation Conference, p.5766-5771, 2020.
5. **Chunhua Liu, Trevor Cohn, and Lea Frermann**, Seeking Clozure: Robust Hypernym extraction from BERT with Anchored Prompts, Proceedings of the 12th Joint Conference on Lexical and Computational Semantics, p.193-206, 2023.
6. Grigor Aslanyan and Ian Wetherbee, Patents Phrase to Phrase Semantic Matching Dataset, arXiv:2208.01171 [cs.CL],2022.
7. Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei, Text Embeddings by Weakly-Supervised Contrastive Pre-training, arXiv:2212.03533 [cs.CL], 2022.
8. Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 [cs.CL], 2013.
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 4171–4186, 2019.
10. Nils Reimers and Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, p. 3982–3992, 2019.
11. 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山, 特許データベースからのシソーラスの自動構築, 言語学処理学会年次大会発表論文集 13th, p. 1113-1116, 2007.
12. 大石康智, 伊藤克亘, 武田一哉, 藤井敦, 単語の共起関係と構文情報を利用した単語の階層関係の統計的自動識別, 情報処理学会研究報告. SLP, 音声言語情報処理 61, p.25-30, 2006.