ファインチューニング済みChatGPTによる特許用語シソーラスの自動構築

Automatic Construction of a Patent Term Thesaurus with Fine-Tuned ChatGPT

中央大学　理工学部　ビジネスデータサイエンス学科　教授
難波　英嗣

2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。博士 (情報科学)。東京工業大学精密工学研究所助手、広島市立大学大学院情報科学研究科准教授等を経て、2019 年より中央大学理工学部教授。自然言語処理、テキストマイニングの研究に従事。

nanba@kc.chuo-u.ac.jp　TEL 03-3817-1883

① はじめに

特許文書は国際的な技術情報の宝庫である一方で、言語・表記・複合語のばらつきが検索や比較の大きな障壁となり、各国特許庁や企業は莫大なコストをかけてシソーラスや分類体系を維持している。既存の自動化研究では、Hearst Pattern[1] のような定型表現抽出や Word2Vec・BERT 系埋め込みを用いた教師ありモデルが主流だが、単言語前提やドメイン適応の不足により、専門用語の多言語対応と精緻な意味関係（上位下位・部分全体など）の把握は依然として限定的である。

本稿では、OpenAI ChatGPT-4o の汎用的な多言語知識を活かし、「任意の二つの特許用語を与えると、その意味関係を直接推定する」シンプルなフレームワークを提案する。英語については Google Patent Phrase Similarity Dataset[3] でファインチューニングを施し高精度を確保しつつ、他言語への拡張はモデル内に潜在するクロスリンガル表現を活用することで追加コーパスや翻訳資源を用いずに実現する。多言語面の検証には、日英特許から定型表現で抽出した上位下位候補（“A などの B”，“B such as A”）を用い、翻訳照合によって高信頼な対訳ペアを形成し、モデルが言語横断で一貫した関係を推論できるかを確認する設計とした。

② 関連研究

技術用語間の意味関係（同義関係、上位語関係、部分語関係など）の自動予測は、知識獲得や網羅的な特許検索のための基盤技術として長らく利用されてきた。従来の手法は、大きく 3 つの系統に分類される：記号的パターンルール、分布的または埋め込みに基づく手法、そして近年では大規模言語モデル（LLM）である。本稿では、それらの発展を年代順に概観し、特許に特化した研究を強調しつつ、本研究の差異を明らかにする。

初期の研究は、明示的な語彙・構文パターンに依拠していた。Hearst の先駆的な研究では、「X is a kind of Y（X は Y の一種である）」のようなテンプレートを導入し、極めて低コストで数千の上位語 – 下位語ペアを抽出している [1]。この考え方は後に日本語特許コーパスにも応用され、難波らは「A などの B（B such as A）」というパターンを抽出し、それを英語の同義語と対応付けることで、F1 値 0.78 を達成する日英シソーラスを構築した [4]。ただし、この手法は上位語 – 下位語関係に限定されており、対照的に本研究では、同義関係、部分語関係、段階的類似性など、より幅広い意味関係を多言語間で予測することを目指す。

分布的アプローチでは、大規模コーパスから連続的なベクトルを学習する。Jana らは、分布的シソーラスをこの空間に投影し、文脈的に類似した語をクラスタリングすることで同義語群の検出に成功した [5]。しかし、単なる類似性では、関係の種類（同義か上位語か）を判別することはできない。その後の研究では、分類器の訓練や制約の追加が行われ、Liu らは「X is a type of __」のようなマスク付きテンプレートを用いて BERT をプロンプトすることで、上位語の頑健な推定を実現した [6]。Transformer による事前学習により文脈モデル

の性能はさらに向上し、BERT[2] およびその変種である Sentence-BERT（SBERT）[7] は、意味的類似性で最先端の性能を達成した。しかし、ドメイン適応は不可欠であり、特許請求項コーパスで訓練された Patent-BERT は、BERT よりも特許関係のベンチマークにおいて大幅に優れた性能を示した。

LLM の登場により、関係性の直接的な推論が可能となった。ChatGPT-4 のようなモデルは、膨大な世界知識を内包し、わずかなプロンプトで定義、同義語、上位語などを生成できる。最近の報告では、ChatGPT-4 が多言語の文化用語に対して分類リンクを導出できることが示されており、これは従来のシステムにはなかった潜在的な言語間能力を示唆している。特許分野では、Peng と Yang が文脈エンコーダと引用由来のフレーズグラフを組み合わせ、自己教師あり手法により局所文脈を超えたグローバルな証拠を捉え、類似性の相関を 7 ポイント改善した [8]。このようなハイブリッド手法は精度を向上させるが、引用情報の収集やグラフ学習といった重厚なパイプラインを要し、かつ単一言語に限定されている。

特許に特化したリソースの登場により、分野横断的な評価も活性化している。Google Patent Phrase Similarity Dataset は、類似度や関係ラベル付きの 5 万の語句ペアを提供しており [3]、これを用いた Kaggle コンペでは、SBERT 系モデルが最強のベースラインであることが示され、特許特化型の事前学習の有効性も明らかとなった。ただし、ほとんどのエントリは英語のみに対応しており、シソーラス構築の自動化には至っていない。

本研究は、先行研究と以下の 3 点で異なる。第一に、軽量な埋め込みフィルタを保持しつつ、最小限のファインチューニングを施した ChatGPT-4o によって関係性を推定し、引用グラフや手作りのルール集合を用いない。第二に、パターン抽出によるバイリンガルのシードペアを用いることで、多言語一貫性を担保し、同一モデルによって日本語と英語の両方でシソーラスを構築できるようにする。第三に、LLM の出力を直接拡張可能なグラフに書き込むことで、関係推定を別個の後処理ステップとせず、即時的なシソーラス構築に転換する。このようにして、本研究は、従来手法が未解決であった多言語対応、専門知識獲得、パイプラインの複雑性といった課題に対処するものである。

③ 提案手法

本フレームワークは、2 つの関係推定戦略と 1 つの多言語検証ステップを通じて、多言語対応の特許用語シソーラスを構築する。(i) 埋め込みベースの類似度推定、(ii) 大規模言語モデル（LLM）による用語間関係解析、(iii) パターン駆動型の多言語拡張の 3 段階から構成される。ステージ (i) および (ii) は、いずれも用語ペア間の意味関係を予測するという点で共通するが、ステージ (i) は用語間の類似度を測るのに止まるのに対し、ステージ (ii) は用語間の関係を 10 種類に分類する。

3.1 埋め込みベースの類似度推定（ステージ (i)）

用語ペア（t_i , t_j）が与えられたとき、それぞれの語に対応するベクトルを、OpenAI Embeddings（d = 1536）または multilingual-e5-large（d = 1024）から取得する。これらのコサイン類似度を意味的関連性の近似スコアとして用いる。このスコアが所定の閾値 τ（OpenAI では 0.35、e5 では 0.30）を超えるペアは、同義関係または分類関係である可能性が高いものとして仮定的に関連ありと見なし、ステージ (iii) の多言語検証ステップに移る。この埋め込みに基づく手法は、高速かつ言語非依存な近似を提供し、ファインチューニングを必要としないという利点がある。

3.2 LLM による用語間関係性推論（ステージ (ii)）

3.1 節とは別の方法として、同じ用語ペアを、Google Patent Phrase Similarity Dataset でファインチューニングされた ChatGPT-4o mini に渡すこともできる。その際のプロンプトは以下のとおりである。

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.’

モデルは表 1 から 1 つのラベルを選択し、それを数値スコア {1.00, 0.75, 0.50, 0.25, 0.00} にマッピングする。ステージ (i) と比較すると、LLM はスカラー値による類似度ではなく、明示的な関係タイプ（例：下位語、部分語）を返す点が異なる。

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

3.3　パターン駆動型の多言語拡張（ステージ (iii)）

1. シード抽出

・日本語：「A などの B」に一致する語句を抽出
・英語：「B such as A」に一致する語句を抽出

これらのパターンにより、仮の下位語 (A) - 上位語 (B) ペアが生成される。

2. 翻訳整合

英語のペアを ChatGPT により日本語に機械翻訳し、日本語セットと突き合わせることで、高信頼な日英用語対を得る。

3. 多言語間検証

各対はステージ (i) または (ii) によって検証され、日本語と英語の推論結果が一致する場合のみ採用される。

4. シソーラスグラフの更新

採用されたペアは、用語ノード間のエッジ（関係タイプ）として追加される。新たなペアが到着するたびに、グラフは自動的に更新される。

本手法は、高速な埋め込み類似度による推論と、LLM による明示的なラベリングという 2 つの補完的推論経路を提供し、それらを言語間で統合する検証層を組み合わせることで、複雑な引用グラフや手作業のルールを用いることなく、最小限のファインチューニングで多言語対応を実現する。実験の詳細については次節で述べる。

4　実験

4.1　実験条件

データセット

英語タスクには、Google Patent Phrase Similarity Dataset[3] を採用し、36,473 ペアを訓練用、9,232 ペアを検証および評価用に使用する。

比較手法

・埋め込みモデル：Word2Vec、GloVe、BERT、SBERT、Patent-BERT（[3] によって報告されたベースライン）、OpenAI Embeddings（text-embedding-3-large）、multilingual-e5-large。
・グラフ＋エンコーダ：RA-Sim により公開されたフレーズグラフ埋め込み（[8] によって報告されたベースライン）。
・LLM：ChatGPT-4o および ChatGPT-4o mini の事前学習済みモデル、さらに英語の訓練データに対してファインチューニングしたバージョン。

評価指標

英語タスクでは、予測された類似度スコアと正解スコアとの間の Pearson 相関係数および Spearman 順位相関係数を報告する。

4.2　実験結果

結果は表 2 に示す通りである。ファインチューニング済みの ChatGPT-4o は最も高い相関（Pearson 0.762）を達成し、グラフ拡張型の RA-Sim に対して Pearson 相関で 0.14 ポイント、Spearman 相関で 0.09 ポイント上回る性能を示した。

表２　類似度スコアの評価結果

タイプ	モデル	Pearson	Spearman
埋め込みモデル	Word2Vec[3]	0.437	0.483
	Glove[3]	0.429	0.444
	BERT[3]	0.418	0.409
	SBERT[3]	0.598	0.535
	Patent-BERT[3]	0.528	0.535
	OpenAI Embeddings	0.581	0.564
	Multilingual-e5-large	0.574	0.546
グラフ	RA-Sim[8]	0.622	0.652
LLM	ChatGPT-4o	0.505	0.514
	ChatGPT-4o mini	0.371	0.403
	ChatGPT-4o（fine-tuned）	**0.762**	**0.738**
	ChatGPT-4o mini（fine-tuned）	0.742	0.718

５　多言語シソーラスの自動構築

1993 年から 2023 年までに公開された日本および米国の特許全文を対象として、多言語対応のシソーラスを自動的に構築する。主たる目的は、上位語－下位語関係の抽出であるが、その過程でその他の意味関係も同時に抽出される。本手法の手順は以下の通りである。

1. 日本語の「A などの B」および英語の “B such as A” という表現を用いて、候補ペアとして日本語 613,251 件、英語 518,166 件を抽出し、ChatGPT による翻訳整合処理を経て 42,784 件のバイリンガルペアを保持した。
2. ChatGPT-4o mini（ファインチューニング済み）を用いて日本語・英語の両言語に対して関係性を推定し、両者のラベルが一致したペアのみを残した（21,673 件）。

ステップ２では、大量のデータを処理するには非常に高コストであることから、表２で最高値を示した ChatGPT-4o に匹敵する性能を持つ ChatGPT-4o mini（fine-tuned）を採用した。

図１は、本研究で提案した手法により構築された多言語シソーラスの一部を示したものである。ランダムに 200 項目を抽出して手動で評価を行った結果、194 件（97%）が正しいと判定された。

誤って抽出されたエントリの中には、完全に誤りとは言い切れず、文脈に依存するものが５件含まれていた。例えば、「materials（Anchor）- metals（Targe）- Hyponym」　や、「information（Anchor）- time（Target）- Hyponym」のような事例である。また、完全に誤っていたものとして、「materials（Anchor）- combinations（Target）- Homonym」の１件が確認された。

Anchor（英語）	Target（英語）	Anchor（日本語）	Target（日本語）	関係ラベル
network	Internet	ネットワーク	インターネット	Hyponym
metal	aluminum	金属	アルミニウム	Hyponym
liquid	water	液体	水	Hyponym
vehicle	automobile	車両	自動車	Highly related

図１　予測された関係ラベルが一致した日本語－英語用語対の例

６　結論

埋め込みベースの類似度推定、ファインチューニング済み ChatGPT-4o、およびパターン駆動型の多言語拡張を組み合わせた３段階のパイプラインを導入し、多言語特許シソーラスを構築した。Google Patent Phrase Similarity Dataset における実験では、提案手法の LLM が埋め込みベースラインおよび最新のグラフ拡張型モデル RA-Sim（Pearson 0.762 対 0.622）を上回る性能を示した。42,784 件の自動整合された日英上位下位用語対に対して、パターン＋ LLM 戦略は 97%の正解率を達成した。

参考文献

[1] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of COLING ' 92, 1992, pp. 539–545.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL 2019, 2019, pp. 4171–4186.

[3] G. Aslanyan, I. Wetherbee, Patents phrase to phrase semantic matching dataset, arXiv:2208.01171, 2022.

[4] H. Nanba, S. Mayumi, T. Takezawa, Automatic construction of a bilingual thesaurus using citation analysis, in: Proceedings of the PalR' 11 Workshop, 2011, pp. 1–8.

[5] A. Jana, N. R. Varimalla, P. Goyal, Using distributional thesaurus embedding for co-hyponymy detection, in: Proceedings of LREC 2020, 2020, pp. 5766–5771.

[6] C. Liu, T. Cohn, L. Frermann, Seeking clozure: Robust hypernym extraction from bert with anchored prompts, in: Proceedings of *SEM 2023, 2023, pp. 193–206.

[7] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of EMNLP–IJCNLP 2019, 2019, pp. 3982–3992.

[8] Z. Peng, Y. Yang, Connecting the dots: Inferring patent phrase similarity with retrieved phrase graphs, in: Proceedings of NAACL 2024, 2024, pp. 1877–1890.