特許請求項からの部分全体関係の抽出

Recognition of Meronymy Relations in Patent Claims

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

シソーラスは、文献を検索したり、高度な言語処理を行ったりするための有用な情報源として活用されている。しかし、シソーラスを人手で構築し、更新することは非常にコストがかかる。テキストデータベースからシソーラスを自動構築する様々な手法が提案されているものの、人手による構築作業に取って代わるレベルまでには至っていない。本研究では、用語間の関係のひとつである部分全体関係に着目し、特許請求項から部分全体関係を抽出する手法を提案する。

これまでに、テキストから部分全体関係を抽出する様々な手法が提案されているが、テキスト中の部分および全体を示す語を抽出するのが一般的であった。本稿では、従来手法で抽出された部分全体関係の候補対の中から不適切なものを除外することで、より高い精度で部分全体関係を抽出する手法を提案する。

2 関連研究

特許データベースから部分全体関係を抽出する手法がいくつか提案されている [1][2][6]。新森ら [1] は、手がかり句を用いて構成要素や手順を含む請求項を構造解析する手法を提案している。ただし、新森らの手法では、構成要素の言語単位が句よりも長い文節に近い定義となっているため、抽出された構成要素をそのまま部分全体関係としてシソーラスに組み入れるには適さないという問題がある。

難波 [2] は、請求項から主要部、構成要素、手順を示す文字列を抽出するためのデータセットを構築し、Bi-LSTM-CRF[3] および CRF[4] を用いてこれらを抽出する手法を提案している。難波の定義する主要部、構成要素、手順は単語または句で表現されるため、上述の問題は解決できる。一方で、Bi-LSTM-CRF や CRF では十分な精度で主要部、構成要素、手順を抽出できないという問題がある。この問題に対し、本稿では、BERT[5] および T5[6] を用いて抽出精度の向上を図る。

Aslanyan[7] らは、特許の用語間の関係に焦点を当てた Patents Phrase to Phrase Matching Dataset[1] を公開している。このデータセットは、2 つの特許用語を入力とし、用語間の意味的な類似度を出力するシステムの構築を想定している。このデータセットの特徴は、用語間の関係を類似度として数値で表現すると同時に、exact(同 一 語)、synonym(同 義 語)、hypernym/hyponym(上 位 下 位)、meronym/holonym(部 分 全

表 1　Patents Phrase to Phrase Matching Dataset の例
(文献 [7] より引用)

anchor	target	context	rating	score
acid absorption	absorption of acid	B08	exact	1.00
acid absorption	acid immersion	B08	synonym	0.75
acid absorption	chemically soaked	B08	domain	0.25
acid absorption	acid reflux	B08	not rel.	0.00
gasoline blend	petrol blend	C10	synonym	0.75
gasoline blend	fuel blend	C10	hypernym	0.50
gasoline blend	fruit blend	C10	not rel.	0.00
faucet assembly	water tap	A22	hyponym	0.50
faucet assembly	water supply	A22	holonym	0.25
faucet assembly	school assembly	A22	not rel.	0.00

1　https://www.kaggle.com/datasets/google/google-patent-phrase-similarity-dataset

体）、not rel.（無関係）といった用語間の関係性を示すラベルが付与されている。（表 1 および図 1 参照）

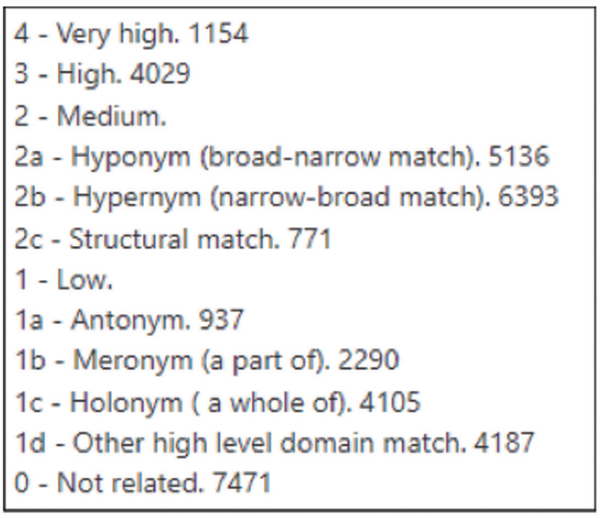

図 1　Patents Phrase to Phrase Matching Dataset の用語間の関係

本研究では、このデータセットを用いて、2 つの用語間の関係を推定する分類器を構築し、請求項から抽出された部分全体関係の候補に適用することで、不適切な候補を除外する。

3　部分全体関係の抽出

本研究では、以下の手順で部分全体関係を抽出する。

1. 請求項からの部分全体関係候補の抽出

2. 部分全体関係候補からの不適切な対の除外

手順 1、2 について、3.1 節、3.2 節でそれぞれ説明する。

3.1　請求項からの部分全体関係候補の抽出

請求項中の部分語および全体語にそれぞれ comp、head タグを付与したコーパスを作成した。図 2 に例を示す。ここで、comp と head タグは構成要素と主要部を示し、部分と全体に対応する。

走行機体に対して <comp> 対地作業装置 </comp>（2）を <comp> 駆動手段 </comp>（4）により前後軸芯周りで駆動ローリング自在に連結するとともに、<comp> 対地作業装置 </comp>（2）の左右両側部に、泥面上に接地追従しながら前記 <comp> 対地作業装置 </comp>（2）の対地高さを検出する <comp> 接地センサ </comp>（5 R），（5 L）を配設し、各 <comp> 接地センサ </comp>（5 R），（5 L）からの検出信号の差が所定レベル内に収まるように前記 <comp> 駆動手段 </comp>（4）を駆動制御する <comp> ローリング制御手段 </comp>（A）を備えてある作業機の <head> ローリング制御装置 </head> であって、前記各 <comp> 接地センサ </comp>（5 R），（5 L）のいずれか一方の検出信号の変化が所定不感帯内にあり、かつ、他方の検出信号が所定短時間内に所定不感帯を越えて変化した場合には、当該検出信号に基づく前記 <comp> 駆動手段 </comp>（4）の駆動を牽制する <comp> 駆動牽制手段 </comp>（B）を備えてある作業機の <head> ローリング制御装置 </head>。

図 2　特開平 5-1 へのタグ付与の例

まず、このコーパスを学習用データに用いて BERT[5] および T5[6] で学習し、部分全体関係候補の抽出器を構築する。次に、この抽出器を用いて 1993 年から 2023 年までの公開公報に含まれるすべての請求項 77,728,548 件を解析し、部分全体関係候補を抽出する。

3.2　部分全体関係候補からの不適切な対の除外

3.1 節で抽出された部分全体関係候補の中から、次に述べる手法で不適切なものを除外する。2 節で述べた Patents Phrase to Phrase Matching Dataset を用いて、与えられた 2 つの用語間の関係を推定する分類器を構築する。分類器を構築するには、まず 2 つの用語間の類似度を計算し、次に類似度の値に応じて用語間の関係ラベルを割り当てる必要がある。これら 2 つの手順について、以下に説明する。

まず、2 つの用語間の類似度を計算する方法について述べる。2 つの用語間の類似度を計算するためのひとつの方法は、2 つの用語をそれぞれ埋め込み表現に変換し、その内積を計算することである。用語を埋め込み表現に変換する方法はこれまでに数多く提案されているが、本研究では E5(Embeddings from Endpoints, Entities, Environments, and Edges)[8] を用いる。その理由は、Patents Phrase to Phrase Matching Dataset は英語の特許用語を対象にしているのに対し、本研究で

は日本語の特許用語を対象にしているためである。E5は、様々なテキストや用語対を用い、対照学習によりテキストの埋め込み表現を学習している。このテキスト対には、例えば、日本語と英語の対訳関係にあるテキスト対も含まれるため、言語が異なっていても意味的に類似する２文であれば類似した埋め込み表現が獲得できる。従って、Patents Phrase to Phrase Matching Dataset を使って構築した英語用の用語間関係の分類器をそのまま日本語に適用することができる。

次に類似度の値に応じて用語間の関係ラベルを割り当てる方法について述べる。まず、Patents Phrase to Phrase Matching Dataset の訓練用データに含まれる２つの用語を E5 を用いて埋め込み表現に変換し、その内積を計算する。次に、２つの用語に付与された関係ラベルに着目し、以下の５つのカテゴリに分類できるよう、しきい値を決定する。

- [レベル 4] exact(同一語)
- [レベル 3] synonym(同義語)
- [レベル 2] hypernym/hyponym(上位下位)
- [レベル 1] meronym/holonym(部分全体)
- [レベル 0] not rel.(無関係)

4 実験

３節で述べた手法の有効性を確認するため、実験を行った。請求項からの部分全体関係候補の抽出は 4.1 節で、部分全体関係候補からの不適切な対の除外は 4.2 節でそれぞれ述べる。

4.1 請求項からの部分全体関係候補の抽出

実験データ

2019 年の請求項に対して、人手で comp および head タグを付与した 800 件のデータを用いる。

評価方法

Recall、Precision、F 値により評価する。モデルが抽出した文字列が人手によるタグと完全一致した場合を正解と判定する。

実験結果と考察

実験結果を表２に示す。Recall、Precision、F 値全てにおいて T5 よりも BERT の方が高くなった。したがって、部分全体関係候補からの不適切な対の除外では、BERT による抽出器を用いる。

表２　請求項からの部分全体関係候補の抽出の結果

	Recall	Precision	F値
BERT	0.926	0.941	0.933
T5	0.799	0.843	0.933

4.2 部分全体関係候補からの不適切な対の除外

実験データ

1993 年から 2023 年までの公開公報に含まれるすべての請求項に対し、BERT を用いた抽出器により部分全体関係候補を抽出する。抽出された候補を頻度順に並べ、その中の 1284 件を人手で部分全体関係にあるかどうかを判定した結果を評価に使用する。なお、1284 件中 820 件が人手により部分全体関係であると判定された。

評価方法

抽出された用語対に対し precision@n で評価する。

比較手法

●提案手法：3.2 節で述べた手順に従って Patents Phrase to Phrase Matching Dataset を用いて学習し、用語間の類似度が 0.81 以上 0.9 未満のものを部分全体関係と判定し、この範囲にあるものを類似度の高い順に並べる。

●ベースライン手法：1993 年から 2023 年までの公開公報に含まれるすべての請求項から抽出された部分全体関係候補を頻度順に並べる。

実験結果

実験結果を図３に示す。図３より提案手法がベースライン手法よりも高い精度で部分全体関係を抽出できることがわかった。

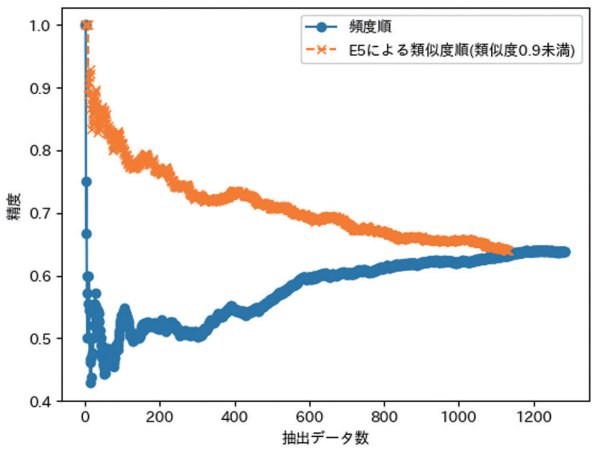

図 3:　頻度順で抽出した場合の部分 - 全体関係の抽出精度

5 おわりに

本研究では特許請求項から BERT を用いて部分全体関係の用語対を抽出し、さらにこれらの用語対を E5 を用いて埋め込み表現に変換し、その類似度が 0.81 ～ 0.9 の範囲にあるものを抽出した時、高い精度で部分全体関係にある用語対が抽出できることがわかった。

参考文献

[1] 新森昭宏，奥村学，丸山雄三，岩山真，手がかり句を用いた特許請求項の構造解析，情報処理学会論文誌，Vol.45, No.3, pp.891-905, 2004.

[2] 難波英嗣，手順オントロジー構築のための特許請求項の構造解析，情報処理学会第 138 回情報基礎とアクセス技術研究発表会 (IFAT)，2020.

[3] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, Neural Architectures for Named Entity Recognition, arXiv:1603.0136v3 [cs.CL], 2016.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of the Eighteenth International Conference on Machine Learning, pp.282-289, 2001.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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, 2019.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research. Vol.21, No.140, pp.1-67, 2020.

[7] Grigor Aslanyan and Ian Wetherbee, Patents Phrase to Phrase Semantic Matching Dataset, arXiv:2208.01171 [cs.CL],2022.

[8] 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.