特許中の画像とテキストを用いた手順オントロジーの構築

樊エイブン[1]　福田悟志[1]　難波英嗣[1]

概要：ある一連の動作に関する概念の集合を手順オントロジーと呼ぶ。新技術は典型的な手順と比べることで、はじめてその新規性を理解することができるため、典型的な手順に関する概念が記述されている手順オントロジーは、非常に重要な役割を果たすと考えられる。しかしながら、これまでにあらゆる技術分野を対象とした網羅的な手順オントロジーは構築されてこなかった。そこで、我々は、手順オントロジーの自動構築に関する研究を行っている。手順オントロジーを構築する際、特許要約と代表図面に着目した。これは、特許があらゆる技術分野をカバーしていることと、特許要約には手続きに関する発明を記述したものが存在するためである。本稿では、手順オントロジー構築の第一歩として、機械学習に基づく特許画像からのフローチャートの抽出と特許要約の構造解析手法について実験を行った。実験の結果、提案手法の有効性を確認することができた。

キーワード：フローチャート，画像認識，情報抽出，特許

Construction of Procedural Ontology Using Images and Texts in Patents

RUIWEN FAN[1]　SATOSHI FUKUDA[1]
HIDETSUGU NANBA[1]

Abstract: A procedural ontology is defined as a set of concepts on a series of actions conducted in a certain order. The ontology is crucial for understanding the state-of-the-art technologies, because typical procedural concepts are contained in the ontology, and we can recognize the novelty of each technology by comparing with typical procedural concepts. However, a procedural ontology that covers comprehensive technical fields have not been constructed. Therefore, we investigate automatic construction of a procedural ontology. In constructing the procedure ontology, we focused on patent abstracts and representative drawings. This is because patents cover all technical fields, and patent abstracts describe inventions related to procedures. In this paper, as a first step to construct a procedure ontology, we conducted several experiments on the extraction of flowcharts from representative drawings and the structural analysis method of patent abstracts based on machine learning. As a result of the experiments, we were able to confirm the effectiveness of our method.

Keywords: flowchart, image recognition, information extraction, patent

1. はじめに

ある特定の目的を達成するための一連の手続きを記したものを手順テキストと呼ぶとき、類似の手順テキストの集合から抽出された典型的な手順が手順オントロジーである。本研究では、この手順オントロジーを構築することを目指し、その第一歩として、特許画像からのフローチャートの抽出と特許要約の構造解析を行う。

一般にオントロジーの人手による構築は非常にコストがかかる。このため、自然言語処理技術を用いて、テキストデータベースからオントロジーを自動的に構築する様々な手法が提案されている。その多くは、上位下位関係や部分全体関係など、用語と用語の様々な関係の抽出を目的としたものである。例えば、用語の上位、下位関係を抽出する代表的な手法としては、「A などの B」などの定型表現に着目したものがあり、「パターン法」と呼ばれている[7, 8]. この場合、「などの」というパターンの前に出現する名詞句 A を後ろに出現する名詞句 B の下位語として抽出される。また、名詞句間の関係だけでなく、動作(事態)に着目した研究も存在する[11]。しかしながら、幅広い分野の一連の手続きに関する知識をテキストから自動抽出し、それらを体系化する試みはほとんどない。

本研究では、特許から手順オントロジーを自動的に構築する手法を提案する。特許では、新しい技術や発明を説明するために、それを実現する手順を記載することがしばしばある。図 1 は「対訳辞書作成装置」に関する日本国特許(特開 2017-091382)の要約であり、S11 から S16 までの手順から構成されていることがわかる。図 2 は、同じ特許の代表図面であり、要約と同じ内容がフローチャートとして表現されている。

ここで、個々の特許には新規性があるため、ひとつの特許だけからこれらの情報を抽出しても、それが対訳辞書作成装置の典型的な手順になっているとは限らない。そこで、日本国特許に付与されている分類コードのひとつである F タームに着目し、同一の F タームが付与されている複数の

1 中央大学
Chuo University

特許から手順情報を抽出し、それらの共通項を検出することで、対訳辞書作成装置の典型的な処理手順に関する知識を自動獲得する。

対訳コーパスから複数の対応文を読み込みＳ１１、複数の対応文から用語を抽出しＳ１２、抽出された用語が用語ペアテーブルに登録されている用語ペアを構成する用語以外である場合には、当該用語を、新規な用語として選定するＳ１３。複数の対応文のマッチングに基づいて、新規な用語のペアを用語ペア候補として取得しＳ１４、用語ペア候補の出現頻度に応じて、当該用語ペア候補を構成する新規な用語ペアを対訳辞書として出力するステップＳ１６。取得するステップでは、複数の対応文の順序をランダムに変更して前記マッチングを繰り返し行う。

図 1：特許要約における手順の記載例(特開 2017-091382)

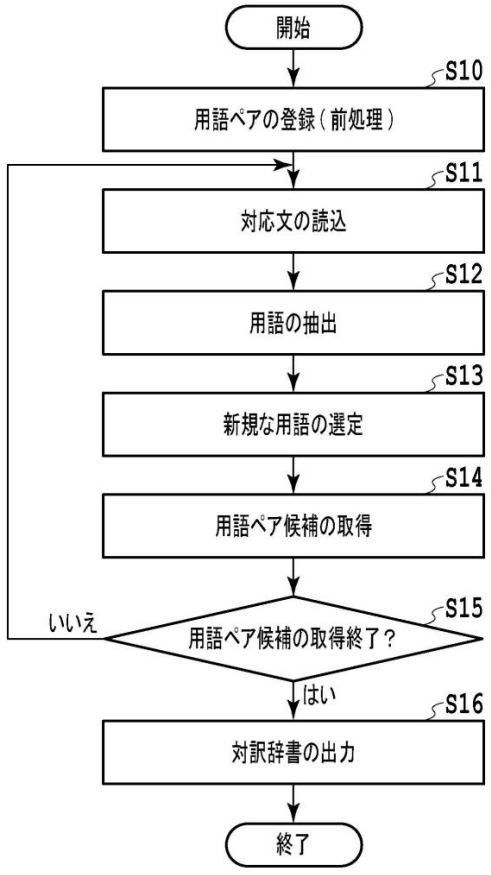

図 2 : 図 1 の特許要約に対応する代表図面の例 (特開 2017-091382)

2. 関連研究

2.1 手順情報の抽出

近年、複数の類似した手順テキストから、共通手順を抽出する研究が行われるようになってきている。山肩ら[17]は、「肉じゃが」や「カルボナーラ」などのクエリを用いて検索した料理レシピ集合に対し、各レシピをその調理手順を表したフローチャートに変換・統合することで、典型的な調理手順(レシピツリー)を導出する手法を提案している。さらに、典型的なレシピツリーと個々のレシピを比較することで、個々のレシピの特徴を抽出している。

料理レシピを対象にしたこの他の研究に、瀧本ら[13]のものがある。瀧本らは、複数の類似レシピから、その共通手順を抽出するタスクを、施設配置問題と捉えている。

高木ら[14]は、「バジルの育て方」などが記載された複数の手順テキストから、その類似点と相違点を検出し、それをひとつのフローチャートとして自動的にまとめ、出力する手法を提案している。

フローチャートを対象とした関連研究もある。近年では、myExperiment や SHIWA など、フローチャートを共有するサービスがはじまっており、これに伴い、あるフローチャートと類似するものを検索する技術の需要が出てきている。Starlinger ら[9]は、あるフローチャートと別のフローチャートがどの程度似ているのかを算出するため、2 つのフローチャート間の対応関係を取る様々な手法について検討している。

新森ら[12]は請求項の構造解析を修辞構造解析の一種と捉え、手掛り語に基づいた請求項構造解析手法を提案している。日本語の請求項には、一般に「～し、～し、～した」のように処理を順序的に記述する順序列挙形式や、「～と、～と、～とからなる、～」のように、構成要素を列挙する形で記述する構成要素列挙形式など、いくつかの特許固有の記述スタイルが存在する。新森らは、手がかり語と文脈自由文法を用いたルールを使い、日本語の特許請求項の解析を実現している。

これに対し、本研究では、機械学習を導入した請求項の構造解析を目指す。近年では、自然言語処理の様々なタスクにおいて深層学習が導入され、その有効性が確認されている。本研究でも、深層学習を用いた請求項の構造解析を試みる。

2.2 フローチャート画像の解析

論文や特許などの図表画像を解析する研究がこれまでにいくつか行われている。Shindo らは、論文から図表画像を抽出し、折れ線グラフを解析することで数値情報を抽出している[5]。Oka らは、論文中の表画像からポリマーの物性情報を抽出している[6]。

フローチャート画像の解析における関連研究プロジェクトとして CLEF-IP がある[3]。CLEF(Conference and Labs of the Evaluation Forum)とは、ヨーロッパを中心に行われている情報検索に関するワークショップであり、CLEF-IP は特許を対象としたタスクのことを指す。このタスクは実験レベルだけではなく、現実の課題に即した検索タスクのためのデータセットを提供することで、多言語及びマルチモーダル特許検索タスクの研究の促進を図っている。CLEF-IP では図形を認識し、フローチャートの要素となるテキス

ト、エッジ、ノードを検出しフローチャートの認識を行っている。CLEF-IP の基本的な課題は、本研究と共通するが、CLEF-IP が実施された 2013 年当時と比べ、画像解析技術が大幅に向上している点、また、本研究では、画像データだけでなく、画像に付随するキャプションや本文のデータがセットになっているという点が異なる。

フローチャート画像を認識するこの他の研究として、Herrera-Cámara のものがある[1]。この研究では、手書きフローチャート画像を解析し、C 言語のソースとして出力する手法を提案している。Sethi らは、深層学習関連の論文中の図表画像からフローチャートを識別し、さらにフローチャートを解析することで、Keras と Caffe でソースを出力するシステムを構築している[4]。著者らは、過去の研究において、フローチャートを解析する手法を提案しているが[16]、本研究では、解析対象となるフローチャート画像を大量の特許画像から検出する手法を提案する。

3. 特許からの手順オントロジーの構築

3.1 手順オントロジーの構築手順

本研究では、ひとつの特許から、その代表画像と要約をそれぞれ解析し、手順情報を抽出する。それらを、特許に付与されている F タームごとに集計することで、手順オントロジーを構築する。以下、図 3 は、機械翻訳(5B091)というテーマコードの F タームの例である。

テーマコード：5B091（機械翻訳）
AA
AA00　言語
AA01　・多言語間
AA11　・1 言語間
AA12　・・方言・標準語間
…
AB
AB00　処理対象要素
AB01　・記号、数字、数式
AB11　・複合語、熟語、イディオム
…

図 3：F タームの例(5B091:機械翻訳分野)

この図において、AA00「言語」や AA01「・多言語間」が観点と呼ばれている。観点は階層的な構造をしており、各説明語の前に記述されている「・」が階層の深さを示している。この場合、AA00 の下層に AA01 や AA11 が存在し、さらに、AA11 の下に AA12 が存在する。A で始まる観点は、「翻訳の対象」という観点からの分類体系となっている。一方、BA ではじまる観点はいずれも「翻訳の方式」に関するものになっている。

図 1 と図 2 には、5B091AA01 という F タームが付与されている。同じ F タームが付与された特許を収集し、それらから手順情報を抽出すれば、多言語間の機械翻訳に関する手順情報の収集が実現できる。3.2 節では特許要約の解析について、3.3 節では特許画像の解析について、それぞれ述べる。

3.2 要約の構造解析

本研究では、特許要約を入力とし、図 4 に示すような構造タグ付きの請求項を出力することを目的とする。図 4 において、comp タグ、proc タグ、head タグはそれぞれ、構成要素、手順、主題を示す。こうしたタグを自動的に付与するシステムを構築するため、人間がタグを付与したデータを準備し、それを教師データとして用いることで、機械学習ベースの構造解析器を構築する。著者らの過去の研究では、請求項の構造解析を行う手法を提案している[15]。今回は、請求項ではなく、要約が対象となるが、要約は、第一請求項と類似した構造を持つことが多いため、要約の構造解析用に新たにタグ付きコーパスを作成するのではなく、請求項の構造解析用タグ付きコーパスで構造解析器を構築した後、そのシステムを要約に適用する。

半導体基板上に、<proc>半導体膜を形成する工程</proc>と、
前記半導体膜の所定の領域に、<proc>ドーパント不純物を導入する工程</proc>と、
前記<proc>半導体膜をパターニングする</proc>ことにより、前記ドーパント不純物が導入された前記半導体膜からなる抵抗素子と、前記ドーパント不純物が導入されていない前記半導体膜からなるゲート電極とを形成する工程とを有することを特徴とする<head>半導体装置の製造方法</head>。

図 4：請求項へのタグ付与の例

本研究では、近年様々な自然言語処理タスクにおいて、その有効性が確認されている言語モデル BERT を用いて、請求項の構造を解析する。BERT の入力層に請求項を入力し、出力層側で各単語(トークン)に対応するタグを出力するよう学習する。なお、BERT のモデルは、Pretrained Japanese BERT models[a]をそのまま用いた場合と、事前に大量の特許データを用いてファインチューニングをしたモデルの 2 種類で実験を行う。ファインチューニングには、壹岐らから提供されたモジュールを用いる[10]。この他、CRF でも実験を行う。CRF では、ターゲットとなる単語から前後 4 単語のユニグラム、バイグラム、トライグラムを素性とした。

3.3 フローチャート画像の自動検出

CNN を用いて、特許中の画像から構成要素情報の抽出を行う。本手法では「ImageNet」と呼ばれる大規模画像デー

a https://github.com/cl-tohoku/bert-japanese

タセットで学習された 7 つの畳み込みニューラルネットワーク(CNN)モデルを用いてファインチューニングによる学習モデルを構築し、その有効性について、次節で述べる実験により検証する。

4. 実験

提案手法の有効性を確認するため、請求項の構造解析およびフローチャート画像の自動検出に関する実験を行った。それぞれ、4.1 節および 4.2 節で報告する。

4.1 請求項の構造解析

実験データ

日本国特許の請求項 2456 件に対し、人手で head, proc, comp タグを付与し、さらにブロック間の依存関係を付与したデータを用いる。表 1 に、請求項 2456 件中の各タグ数の内訳を示す。

表 1：請求項の構造タグ数の内訳

head	comp	proc
2298	6088	581

実験方法

人手で作成したタグ付きデータのうち、3/4 を訓練用とし、残りの 1/4 を評価用に用いた。評価には、精度(P)、再現率(R)、F 値(F)を用いた。

比較手法

以下の 3 種類の手法で実験を行った。

- BERT(事前学習なし)：Pretrained Japanese BERT models をそのまま利用
- BERT(事前学習あり)：公開特許公報から任意に選択した 350 万文を用いて Masked Language Model タスクにより事前学習したモデルを利用
- CRF(ベースライン手法)：前後 4 単語のユニグラム、バイグラム、トライグラムを素性として利用

実験結果

実験結果を表 2 に示す。この結果より、CRF と比べ、BERT(事前学習なし)が再現率を 0.09 以上向上させることができた。BERT(事前学習あり)は、BERT(事前学習あり)の精度を若干向上させることができたものの、F 値では BERT(事前学習なし)とほぼ同じ値となった。特許を用いた事前学習がそれほど解析精度の向上に大きく貢献しなかった理由のひとつは、事前学習に用いた文の多くは請求項ではなく明細から抽出してきたからであろうと思われる。このため、文体が異なる請求項に対して、大きな改善が見込めなかったためと考えられる。この点については、事前学習を、請求項のみを用いることでさらなる改善が期待できる。

表 2：請求項の構造解析精度

手法	精度	再現率	F 値
BERT(事前学習なし)	0.773	**0.867**	**0.817**
BERT(事前学習あり)	0.791	0.837	0.814
CRF(ベースライン手法)	**0.816**	0.776	0.795

4.2 フローチャート画像の自動検出

実験データ

日本国特許公開公報 2018 年から抽出した画像 7,099 件を利用して、フローチャートか否かを人手で判定した。表 3 にデータの内訳を示す

表 3：フローチャート画像の自動検出実験用データの内訳

図表種別	件数
フローチャート	1,120
その他	5,979
計	7,099

実験手法

Baseline として、深層学習ライブラリである Keras を用いて Conv2D が 3 層、MaxPooling2D が 2 層の畳み込みニューラルネットワーク学習モデル(Baseline)を構築した。比較手法として、「ImageNet」と呼ばれる大規模画像データセットで学習された 7 つの畳み込みニューラルネットワークモデルを用いて、ファインチューニングによる学習モデルを構築し、それらと比較した。評価は精度、再現率、F 値を用いた。

実験結果と考察

実験結果を表 4 に示す。表 4 より、今回比較した手法の中では、精度では DenseNet121 が最もフローチャートの検出精度が高いことが分かった。

表 4：８つのモデルによるフローチャート検出精度

	精度	再現率	F 値
Baseline	0.8508	0.8902	0.8701
VGG16	0.8750	**0.9711**	0.9205
VGG19	0.9227	0.9653	0.9435
ResNet50	0.8698	0.9653	0.9151
InceptionV3	0.9422	0.9422	0.9422
MobileNet	0.9326	0.9595	0.9459
DenseNet169	0.9593	0.9538	**0.9565**
DenseNet121	**0.9645**	0.9422	0.9532

5. おわりに

本研究では、日本国特許の要約および代表図面から手順情報を抽出し、さらに、特許に付与されている F タームと組み合わせることで、手順オントロジーを構築する手法を

提案した。実験の結果、請求項の構造解析では BERT(事前学習なし)手法により F 値 0.817 が得られた。また、フローチャート画像の検出では、DenseNet121 を用いてファインチューニングしたモデルで精度 0.9645 を達成した。

謝辞 本研究は JSPS 科研費 20H04210 の助成を受けたものである。

参考文献

[1] Herrera-Cámara J.I.. FLOW2CODE - From Hand-drawn Flowchart to Code Execution, Master Thesis, Texas A&M University, 2017.
[2] Ester, M., Kriegel, H.P., Sander, J., and Xu, X.. A density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD-96 Proceedings, 1996, p.226-231.
[3] Piroi, F., Lupu, M. and Hanbury, A.. Overview of CLEF-IP 2013 Lab Information Retrieval in the Patent Domain, Information Access Evaluation. Multilinguality, Multimodality, and Visualization. CLEF 2013. Lecture Notes in Computer Science, vol. 8138. Springer, Berlin, Heidelberg, 2013.
[4] Sethi, A., Sankaran, A., Panwar, N., Khare, S., and Mani, S.. DLPaper2Code: Auto-generation of Code from Deep Learning Research Papers, Proceedings of the 32th AAAI Conference on Artificial Intelligence, 2018.
[5] Shindo, H. and Matsumoto, Y.. Automatic Reading of Tables and Figures in Scientific Papers, CBI Annual Meeting, 2019.
[6] Oka, H., Shindo, H., Goto, K., Matsumoto, Y., Yoshizawa, A., Kuwajima, I., and Ishii, M.. Automatic Extraction of Polymer Data from Tables in Xml, Proceedings of the 3rd International Workshop on Scientific Document Analysis (SCIDOCA 2018), 2018.
[7] Hearst, M. A., Automatic Acquisition of Hyponyms from Large Text Corpora, in Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545, 1992.
[8] Roller, S., Kiela, D., and Mickel, M., Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.358-363, 2018.
[9] Starlinger, J., Brancotte, B., Cohen-Boulakia, S., and Leser, S., Similarity Search for Scientific Workflows, Proceedings of the VLDB Endowment, Vol. 7, No. 12, pp.1143-1154, 2014.
[10] 壹岐太一，金沢輝一，相澤彰子，学術分野に特化した事前学習済み日本語言語モデルの構築，情報処理学会第 139 回情報基礎とアクセス技術研究発表会，2020.
[11] 乾健太郎，事態オントロジー：言語に基づく推論のためのコトに関する基本知識，言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」論文集, pp.27-30, 2007.
[12] 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.
[13] 瀧本洋喜，笹野遼平，高村大也，奥村学，施設配置問題に基づく同一料理のレシピ集合からの基本手順の抽出，言語処理学会第 21 回年次大会発表論文集，pp.1092-1095，2015.
[14] 高木優，藤井敦，手順テキストを対象とした比較対象要約，言語処理学会第 21 回年次大会発表論文集，pp.573-576，2015.
[15] 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，2020.
[16] 樊エイブン，橘本勇太郎，難波英嗣，特許検索のためのフローチャート画像の解析，情報処理学会第 142 回情報基礎とアクセス技術研究発表会・第 120 回ドキュメントコミュニケーション研究会，2021.
[17] 山肩洋子，今堀慎治，杉山祐一，田中克己，レシピフローグラフを介したレシピ集合の要約と特徴抽出，電子情報通信学会技術研究報告, DE 研第 1 種研究会 データ工学と食メディア，Vol. 113, No. 214, DE2013-36, pp.43-48，2013.