様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １３ 日現在

機関番号：３２６４１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２１０１

研究課題名（和文）手順オントロジーの自動構築および情報検索への適用

研究課題名（英文）Automatic Construction of Procedural Ontology and Its Application to Information Retrieval

研究代表者
難波　英嗣（Nanba, Hidetsugu）

中央大学・理工学部・教授

研究者番号：５０３４５３７８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：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、本研究では、手順オントロジーの自動構築に関する研究を行った。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。本研究では、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。

研究成果の学術的意義や社会的意義
オントロジーは、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が、一般にオントロジーの人手での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。本研究は、特許の要約および代表図面から体系的な手順オントロジーを自動構築する知る限りはじめての試みであり、その成果は、特許検索や特許文書作成支援などへの応用が期待できる。

研究成果の概要（英文）：A procedural 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 investigated automatic construction of a procedural ontology. In constructing the procedure ontology, we focused on patent abstracts and representative drawings. In this research, 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.

研究分野：自然言語処理

キーワード：特許　構造解析　画像分類　オントロジー　手順　情報検索

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

オントロジーは、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が、一般にオントロジーの人手での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。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。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。

２．研究の目的

本研究では、一連の手順に関して体系化された知識を手順オントロジーと呼び、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。特許の請求項には発明の内容が記載されており、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。このうち、順序列挙形式の請求項に着目し、その構造を解析することで、一連の手順を抽出する。こうして作成した手順オントロジーを情報検索のタスクに適用し、従来の情報検索技術よりも検索性能を向上させることが可能であることを実証する。

３．研究の方法

３．１　手順オントロジーの構築手順

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)

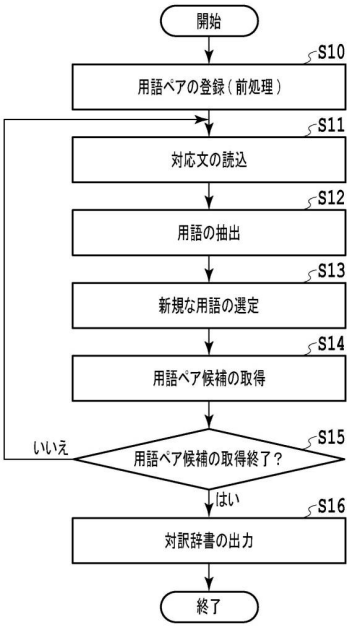

図 2：図 1 の特許要約に対応する代表図面

３．２　フローチャート画像の自動検出

ある画像がフローチャートであるかどうかを認識するタスクは、画像分類の一種と考えられる。その代表的な手法は、大量のラベル付き画像データを対象に畳み込みニューラルネットワーク(CNN)等のモデルを用いて学習をする。近年では、このような学習済モデルが公開されており、このモデルを用いてファインチューニングにより目的の画像分類器を構築する手法が一般的になっている。本研究でも、既存のモデルを利用してフローチャート画像の検出器を構築する。本研究では、ImageNet(https://www.image-net.org/)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(CNN)モデル(VGG16、VGG19、ResNet50、InceptionV3、DenseNet169、DenseNET121、MobileNet)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実験により検証する。また、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 CNN モデルを構築し、ファインチューニングによる手法と比較する。

３．３　要約からの手順情報の抽出

要約を入力とし、図 3 に示すような構造タグ付きの要約を出力することを目的とする。図 3 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、人手で構造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。

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 3：請求項へのタグ付与の例

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、東北大学が公開している Pretrained Japanese BERT models 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。

３．４　フローチャート画像からの手順情報の抽出

フローチャート画像からの手順情報の抽出は、コンピュータビジョン向けライブラリ OpenCV(https://opencv.org)を用い、1. 輪郭の抽出、2. 輪郭の近似・座標情報の抽出、3. 形状の認識、4. 面積によるクラスタリング、5. 文字認識、6. フローチャートと本文の対応付けの 6 つの手順から構成される。図 4 では、形状認識の際に形状ごとに色分けをしており、長方形は緑、楕円形は青、ひし形は赤で表している。解析結果からわかるとおり、矩形領域、条件分岐

を示すひし形、フローの最初と最後を示す楕円が正しく解析されている。一方で、条件分岐が連続する個所の右側の領域など、抽出すべきでない個所を誤って抽出してしまっている。

図19

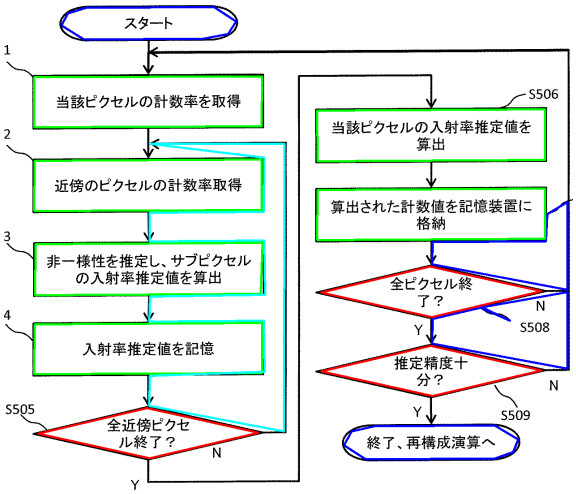

図４：手順情報の抽出結果

３．５　類似内容の特許請求項の自動検出

二つの請求項の同一性を自動判定する手法を提案する。まず、3.1 節で述べた手法を用いて、各請求項の構造を解析し、次に、要素間を対応付け、二つの請求項の主題部および構成要素か手順が一つ以上同一の場合、二つの請求項が同一であると判定する。本研究では、特許庁における審査において、審査対象となった請求項と、この請求項を拒絶する根拠となった請求項を類似内容の請求項対として実験に用いる。

４．研究成果

提案手法の有効性を確認するため、フローチャート画像の自動検出、請求項の構造解析、フローチャート画像の解析に関する実験を行った。それぞれ、4.1 節、4.2 節、4.3 節で報告する。

４．１　フローチャート画像の自動検出

実験データ

2018 年の公開公報から抽出した 7,099 画像に対し、人手で(1)フローチャート、(2)表、(3)構成図、(4)その他に分類したデータを用いる。

実験手法

3.2 節で述べたとおり、ImageNet で学習された 7 つの CNN モデルを用いて、ファインチューニングによるフローチャート画像の検出器を構築した。Baseline として、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を構築した。評価は精度、再現率、F-measure を用いた。

実験結果

実験結果を表 1 に示す。表より、今回構築した学習モデルの中では、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１：８つのモデルによるフローチャートの検出精度

	精度	再現率	F-measure
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

４．２　要約からの手順情報の抽出

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。

実験方法

人手による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、再現率、F-measure 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 MLM 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

結果を表 2 に示す。表より、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。一方、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-measure では BERT(事前学習なし)とほぼ同値となった。

表 2:請求項の構造解析精度

手法	精度	再現率	F-measure
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

４．３　フローチャート画像からの手順情報の抽出

実験データ

2018 年の公開公報から抽出した 37 画像(いずれもフローチャート)を用いる。これらの画像に対し、人手で 502 ノードの座標を判定した。なお、各座標はオブジェクトの形状に関係なく、すべて矩形領域の座標として近似的に判定している。

評価方法

出力結果と比較し性能を評価する。ノードの座標の一致度の評価尺度として精度と再現率を用いる。なお、システムの出力と人手のノードを比較する際、領域の重なる面積が 90%を超えた時、2 つのノードが一致したと判定する。また、比較手法としてクラスタリングをしない場合の精度、再現率を求める。

比較手法

3.1 節で述べた 6 つの手順のうち、手順 4 においてクラスタリングを行った場合(提案手法)とクラスタリングを行わず、すべてのノードを抽出した場合(ベースライン手法)とで比較した。

実験結果

結果を表 3 に示す。表より、提案手法と比較手法を比べると再現率を低下させることなく高い精度を得ることができたため、クラスタリングは有効であり、提案手法の方が優れているという結果が得られた。

表 3:フローチャート画像からの手順情報の抽出結果

	精度	再現率
(1) 提案手法	0.753	0.673
(2) ベースライン手法	0.693	0.673

４．４　類似内容の特許請求項の自動検出

実験データ

国内引用文献マスタから抽出した 1,554 対の請求項に対し、新森手法および難波手法を用いて構造解析をしておき、それらの構造を用いて、請求項対が同一の内容であるかどうかを判定する。

比較手法

以下の 2 種類の方法で実験を行う。

- 単語:4.2 節の手法で請求項の構造解析をした後、単語頻度+コサイン距離で請求項の同一性を判定
- BERT:4.2 節の手法で請求項の構造解析をした後、BERT+コサイン距離で請求項の同一性を判定

実験結果

実験の結果、単語手法では 0.337、BERT 手法では 0.646 の割合で請求項対の同一性が検出できることがわかった。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 難波英嗣
２．発表標題 特許出願技術動向調査報告書の自動更新に向けて
３．学会等名 2020年度人工知能学会全国大会
４．発表年 2020年

１．発表者名 難波英嗣
２．発表標題 類似内容の特許請求項の自動対応付け
３．学会等名 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
４．発表年 2021年

１．発表者名 難波英嗣
２．発表標題 手順オントロジー構築のための特許請求項の構造解析
３．学会等名 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)
４．発表年 2020年

１．発表者名 樊エイブン，福田悟志，難波英嗣
２．発表標題 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
３．学会等名 情報処理学会第145回 情報基礎とアクセス技術研究発表会(IFAT)
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関