手順オントロジー構築のための特許請求項の構造解析

Structure Analysis of Patent Claims for Construction of Procedural Ontology

中央大学理工学部経営システム工学科教授
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

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、特許請求項の構造解析を行う。

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[1, 3]。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[5]。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 および図 2 は「fabricating a semiconductor device（半導体装置の製造方法）」に関する米国および日本国特許の一例 であり、いずれも「(1) forming a semiconductor film（半導体膜を形成する工程）」、「(2) implanting a dopant（ドーパント不純物を導入する工程）」、「(3) patterning the semiconductor film（半導体膜をパターニングする工程）」の３手順から構成されることがわかる。

> A method for fabricating a semiconductor device comprising the steps of:
> forming a semiconductor film(1) over a semiconductor substrate;
> implanting a dopant(2) into a first region of the semiconductor film where a resistance element is to be formed, a second region of the semiconductor film where a gate electrode is to formed being not implanted with the dopant; and
> patterning the semiconductor film(3) to form the resistance element of the semiconductor film with the dopant implanted, and the gate electrode of the semiconductor film with the dopant not implanted, before any step of implanting any dopant to be implanted into the gate electrode.

図 1　特許における手続きの記載例（米国特許）

半導体基板上に、半導体膜を形成する工程(1)と、
前記半導体膜の所定の領域に、ドーパント不純物を導入する工程(2)と、
前記半導体膜をパターニングする(3)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半導体装置の製造方法。

図２　特許における手続きの記載例（日本国特許）

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、それが半導体装置の製造方法の典型的な手順になっているとは限らない。そこで，半導体装置の製造方法に関する複数の請求項を集め、各請求項から手順を抽出し、それらの共通項を検出することで、半導体装置の製造方法の典型的な処理手順に関する知識を自動獲得する必要がある。このような一連の処理を行うための第一歩として、本稿では、特許請求項の構造解析を行う。

2 関連研究

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[9]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[7]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[8]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[4]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[6]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。さらに、新森らは、日本語の特許請求項を解析対象としていたが、本研究では、日本語と英語の両方を対象とする。

3 特許請求項の構造解析

3.1 請求項の構造

本研究では、図１や図２に示すような請求項をシステムの入力とし、図３や図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３および図４において、工程の手順または装置の構成要素を示す文字列の前後に block タグと各 block タグの識別番号を示す id が付与される。ブロック間には依存関係があり、あるブロックの依存先は、block タグの link 属性として記述される。各 block タグ内の主要な内容を示す個所に comp または proc タグを付与する。comp タグは構成要素を、proc タグは手順を、それぞれ示す。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、機械学習ベースの請求項構造解析器を構築する。

1. A method for <block id="1" link="-1"><head>fabricating a semiconductor device</head> comprising the steps of:
<block id="2" link="1"><proc>forming a semiconductor film</proc> over a semiconductor substrate;
<block id="3" link="1"><proc>implanting a dopant</proc> into a first region of the semiconductor film where a resistance element is to be formed, a second region of the semiconductor film where a gate electrode is to formed being not implanted with the dopant; and
<block id="4" link="1"><proc>patterning the semiconductor film</proc> to form the resistance element of the semiconductor film with the dopant implanted, and the gate electrode of the semiconductor film with the dopant not implanted, before any step of implanting any dopant to be implanted into the gate electrode.

図３　米国特許請求項へのタグ付与の例

<block id="2" link="1">半導体基板上に、<proc>半導体膜を形成する工程</proc>と、
<block id="3" link="1">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<block id="4" link="1"><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<block id="1" link="-1"><head>半導体装置の製造方法</head>。

図４　日本国特許請求項へのタグ付与の例

3.2　請求項の構造解析手法の提案

請求項の構造解析は以下の３つの手順から構成される。なお、手順１は米国特許請求項のみに適用する。日本国特許については、手順２から実施する。

（手順１）請求項の複数ブロックへの分割

米国特許請求項を対象に、簡単なルールを用いて、請求項をブロックに分割する。米国特許の請求項のほとんどは、“：”や“；”や“；and”の前後でブロックに分割することができる。

（手順２）各ブロックからの主要個所の抽出及び種類の判定

各ブロック内における主要な個所（多くの場合、名詞句）に対し、head, proc, および comp タグを付与する。本研究では、Lample ら[2] の提唱する Bi-directional LSTM-CRF を用いて自動タグ付与を実現する。図５を用いて、Bi-directional LSTM-CRF による米国特許請求項へのタグ付与の過程を説明する。入力として単語埋め込みを LSTM に与え、LSTM の記憶セルで離れた距離の依存関係をモデル化することで、各単語に対して固有表現タグ（head, proc, comp）を付与する。またこの際、単に先頭の単語から順方向に解析する LSTM だけでなく、最後の単語から逆方向に解析する LSTM を合わせた双方向 LSTM を使用することで高精度での固有表現抽出を行う。本研究では、Bi-directional LSTM-CRF の実装のひとつである anaGo[1] を利用し、各ブロック中の単語に対して head, comp, proc タグを自動付与する。

なお、請求項内の単語を単語埋め込みに置き換える際、米国特許については、国際特許（WO）1998 ～ 2017 年の英語概要 2,560,234 件（1.3GB）を word2vec で学習したモデルを利用する。また、日本国特許については、日本国特許 2004 ～ 2016 年の日本語概要 11,318,066 件（3.3GB）を用いて学習したモデルを利用する。

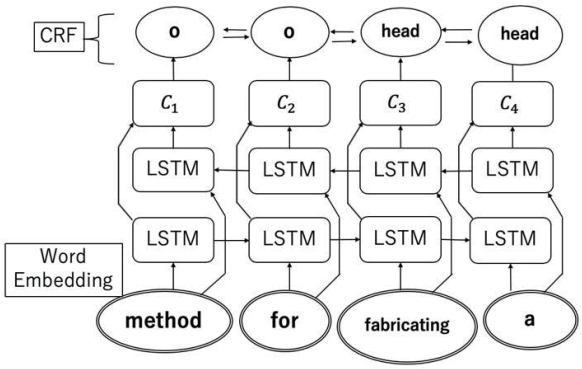

図５　Bi-directional LSTM-CRF を用いた米国特許請求項へのタグ付与の過程

この他、CRF でも実験を行う。CRF では、ターゲットと成る単語および品詞から前後４単語（品詞）のユニグラム、バイグラム、トライグラムを素性とした。

（手順３）ブロック間の依存関係の解析

head タグを含むブロック以降、comp および proc タグが隣接するブロックで連続して出現する場合、それらはすべて head タグにリンクする。もし異なる種類のタグが出現する場合、そのブロックは、タグの種類が変

1　https://github.com/Hironsan/anago

わる直前のブロックにリンクする。

なお、手順３について、本来ならば、依存関係の解析も機械学習を用いるべきであるが、comp および proc タグが head に依存するケースが全体の 94.0%（5580/5940）を占めており、機械学習の適用が難しいと判断されたため、今回は以下に示すルールを用いてブロック間の依存関係を解析する。

（ルール 1）head タグ以降、同じタグが連続する場合には、それらのタグは head タグに係る。

（ルール 2）comp タグの直後に proc タグが出現した場合には、その proc タグを含むブロックは直前の comp タグを含むブロックに係る。proc タグの直後に comp タグが出現した場合も同様。

4 実験

提案手法の有効性を確認するため、上記手順２に関する実験を行った。

4.1 実験条件

実験データ

米国特許および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　タグ数の内訳

タグ	head	comp	proc
米国	1557	6061	583
日本国	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２種類の手法で実験を行った。

・Bi-directional LSTM-CRF[2]

・CRF

4.2 実験結果および考察

米国特許および日本国特許を用いた実験結果を表２および表３に示す。表に示すとおり、Bi-LSTM-CRF と CRF では、米国特許でも日本国特許でも、CRF の方が再現率、精度ともに高い結果となった。

Bi-LSTM-CRF と CRF はいずれも、proc タグの精度、再現率、F 値が head や comp タグのものと比べて低いが、これは訓練データの数が少ないことに起因すると思われる。米国特許と比べ、日本国特許の場合、proc タグの付与精度が極端に低いが、これは言語固有の書き方の問題であると思われる。米国特許の場合、proc タグの先頭単語は動詞の ing 形で記載されることが多い。実際に表３では、proc タグの直後の単語は forming, implanting, patterning と、いずれも動詞の ing 形である。これに対し、日本国特許の場合は米国特許よりも表現が多様である。日本国特許では、例えば新森ら[6]が日本国特許請求項の解析の手がかりとして用いている「～し、～し、～した」といった動詞の連続が出現すれば解析できる。しかし、図４のように proc タグの末尾が「工程」の場合、この単語は動詞でもサ変名詞でもないため、この単語以前の文字列に proc タグを付与するのは容易ではない。

proc のタグ付与精度が極端に低くなったこの他の原因として、単語埋め込み表現の学習に用いたデータの問題がある。3.2 節手順２で述べたとおり、日本語の単語埋め込みは日本語概要から学習している。しかし、請求項は特許の中でもとりわけ特徴的な記述方式が用いられる。従って、単語埋め込みの学習に請求項だけを利用することで、日本国特許請求項のタグ付与精度が向上する可能性がある。

表２　head, proc, および comp タグ付与精度（米国特許）

	Bi-LSTM-CRF			CRF		
	P	R	F	P	R	F
head	0.85	0.86	0.85	0.89	0.89	0.89
comp	0.76	0.73	0.74	0.83	0.77	0.80
proc	0.52	0.37	0.43	0.64	0.39	0.48
平均	0.76	0.73	0.74	0.83	0.77	0.80

表３　head, proc, および comp タグ付与精度（日本国特許）

	Bi-LSTM-CRF			CRF		
	P	R	F	P	R	F
head	0.84	0.85	0.84	0.86	0.87	0.87
comp	0.77	0.78	0.78	0.82	0.79	0.81
proc	0.37	0.29	0.32	0.37	0.23	0.27
平均	0.76	0.77	0.77	0.82	0.78	0.79

5 おわりに

本研究では、米国特許および日本国特許の請求項を対象に Bi-directional LSTM-CRF および CRF により、請求項の主要部（head）、構成要素（comp）、手順（proc）を抽出する実験を行った。実験の結果、米国および日本国特許の両方において、CRF の方が Bi-directional LSTM-CRF よりも再現率、精度ともに高い結果となった。個別のタグに着目すると、日本国特許は proc タグの付与精度が極端に低かった。この点については、例えば、単語埋め込みの学習に用いるデータを請求項のみに限定することで改善できる可能性がある。

謝辞

本研究は、科研費（19K12101）の助成を受けたものである。

参考文献

[1] 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.

[2] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C., Neural Architectures for Named Entity Recognition, arXiv:1603.0136v3 [cs.CL], 2016.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.