請求項の構造解析によるオントロジーの自動構築

Automatic Construction of an Ontology by Analyzing Structure of Patent Claims

広島市立大学大学院情報科学研究科准教授 **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等を経て、2010 年より広島市立大学大学院情報科学研究科准教授。自然言語処理、テキストマイニングの研究に従事。

nanba@hiroshima-cu.ac.jp　　TEL 082-830-1584

1 はじめに

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ため、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。筆者は、これまで特許データベースから特許の請求項を解析することにより、オントロジーを自動構築する様々な研究を行ってきたが [1, 2, 3, 4, 5]、本稿では特に手順の体系化や部分 - 全体関係を取り上げる。

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 [6] は、請求項の構造解析を修辞構造解析の一種と捉え、手がかり語に基づいた請求項構造解析手法を提案している。筆者も新森らと同様に請求項を構造解析する手法を提案するが、以下に述べる 2 点で異なる。第一に、請求項の構造を解析した後、解析結果を利用してオントロジーを構築することを視野に入れた請求項の構造を定義している点である。第二に、新森らが日本語の請求項のみを対象にしているのに対し、筆者は日本語と英語の両方を対象にする。

本稿の構成は以下のとおりである。次節では、請求項の構造を定義し、その解析方法を提案する。さらに、請求項の構造解析に関する実験について、その結果を報告する。3 節で本稿をまとめる。

2 特許請求項の構造解析

2.1 請求項の構造

図 1 と図 2 に、日本語および英語で記述された請求項に人手で構造タグを付与した例を示す。

半導体回路のレイアウトデータに対し、<block id=”1”link=”7”>シミュレーションを施したシミュレーション画像データのパターン線分を膨張する <comp> 膨張処理部 </comp></block> と、<block id=”2”link=”7”>当該膨張処理が施されたシミュレーション画像データの <comp> パターン内側領域 </comp></block>、及び <block id=”3”link=”7”>外側領域の一方にマスクを施す <comp> マスク処理実行部 </comp></block> と、<block id=”4”link=”7”>当該マスクが施された <comp> シミュレーション画像データ </comp></block> と、<block id=”5”link=”7”>荷電粒子線装置によって得られた画像を重畳させ、当該マスクが施された領域以外の領域について、輝度信号を検出する <comp> 輝度信号抽出部 </comp></block> と、<block id=”6”link=”7”>当該輝度変化が所定値を超えた部分の有無の判定、或いは当該所定値を超えた部分の位置情報を抽出する <comp> 欠陥抽出部 </comp></block> を備えたことを特徴とする <block id=”7”link=”-1”><head> 欠陥検査装置 </head></block>。

図 1　日本語請求項へのタグ付与の例

1. A method of <block id="1" link="-1"><head>optimizing the geometry of a femoral stem of a hip joint prosthesis, the femoral stem</head></block> comprising
<block id="2" link="1"><comp>a neck</comp></block>; and
<block id="3" link="1"><comp>an anchoring blade</comp> that is attached to the neck and that tapers towards a distal end with a lateral narrow side
comprising a distal straight portion and a proximal arcuate portion corresponding to a curve, a transition between the distal straight portion and said proximal
arcuate portion occurring at an outer lateral point</block>; and
<block id="4" link="1"><comp>said method</comp></block> comprising <block id="5 link="4"><proc>optimizing the profile of the curve of said proximal arcuate portion by a process of iterative modeling steps using a series of curves each defined by a path traced by the outer lateral point of the blade on withdrawal of a profile of the stem from a cavity of complementary shape to the stem.</block>

図 2　英語請求項へのタグ付与の例

図において、手順または部分を示す文字列の前後に block タグを付与している。各 block タグには id が順に付与されている。さらに、各ブロックとのリンク関係を block タグの link 属性として記載している。リンク関係について、図 1 を例に説明する。この請求項は「欠損検査装置」に関するもので、id=7 の block に、その記載がある。id=1 ～ 6 の block は、それぞれ欠損検査装置の構成要素となっているため、これらの block は link 属性の値が欠損検査装置の id=7 となっている。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。英語請求項に関しても、同様の形式でデータを人手で作成する。このデータを用いて、人間がタグを付与したのと同様のタグを付与するシステムを、機械学習ベースの手法で構築する。

2.2 請求項の構造解析

前節で述べたデータは現在作成中であるが、このデータを用いて構築したシステムでオントロジーの構築が可能であるか確認するため、日本語請求項へのタグ付与データの一部 (144 請求項) を用いて簡単な実験を行った。本来は、システムに請求項が入力されると、それらにまず block タグを付与し、次に block タグ間の link 関係を解析し、最後に各 block 内の文字列に対し comp または proc タグを付与するが、今回は問題設定を簡略化し、入力された請求項に直接 comp または proc タグを付与する、いわゆる系列ラベリング問題として捉え、機械学習の一種である Conditional Random Field(CRF) を用いてシステムを構築した。機械学習の素性には、前後 3 単語および品詞のユニグラム、バイグラム、トライグラムを用いた。評価尺度は再現率、精度、F 値を用いた。また、2 分割交差検定で実験を行った。実験結果を表 1 に示す。

表 1　日本語請求項への head、comp および proc タグ付与結果

タグ	精度	再現率	F 値
comp	0.760	0.608	0.671
proc	0.389	0.194	0.259
head	0.843	0.784	0.812

表 1 から、proc タグについてはタグ付与精度が低いものの、head と comp に関しては、比較的良好な結果が得られた。proc タグについては、head や comp と比べ、記述方法が多様であるため抽出に失敗したケースが多かったが、今回の実験では、単語と品詞以外に特別な素性は使っていないため、今度、精度向上の余地は十分にあると考えられる。

この自動タグ付与システムをプリンタ関連 [1] の日本国特許の請求項に適用し、comp と head タグの対を抽出することで、部分 - 全体関係の構築を試みた。抽出結果を表 2 に示す。

1　F タームのテーマコードが 2C および 2H からはじまるもので、2004 年～ 2014 年の公開公報を対象とした。

表２　プリンタ関連特許から部分 - 全体関係を抽出した結果

全体	部分	抽出件数
画像形成装置	像担持体	11622
画像形成装置	制御手段	9555
画像形成装置	転写手段	7032
画像形成装置	現像手段	6797
画像形成装置	現像装置	5945
画像形成装置	画像形成部	5682
画像形成装置	画像形成手段	5389
画像形成装置	帯電手段	4109

表２より、部分 - 全体関係の抽出に関しては、比較的良好な結果が得られていることが分かる。

3 おわりに

本稿では、請求項の構造を解析し、部分 - 全体関係および手順に関するオントロジーを構築する手法について述べた。今後は、日英データを整備し、block タグの付与、link 関係の解析についても取り組む予定である。

参考文献

[1] 難波英嗣、竹澤寿幸. (2015) “複数手順テキストからの手順オントロジーの自動構築” 電子情報通信学会データ工学研究会.

[2] 福田悟志、難波英嗣、竹澤寿幸、乾孝司、若山真、橋田浩一、藤井敦. (2015) “F タームに基づいたオントロジーの構築” 言語処理学会 第 21 回年次大会

[3] 難波英嗣、乾孝司、岩山真、櫻井孝、橋田浩一、藤井敦. (2014) “特許分類コード体系に基づくオントロジーの構築 － 情報分野におけるケーススタディー－” 言語処理学会 第 20 回年次大会.

[4] Hashida, K., Nanba, H., Inui, T., Iwayama, M., Hashimoto, T., and Fujii, A. (2011) “Circulation of Collective Intelligence through Patents: An Early Progress Report” . Procedia - Social and Behavioral Sciences, Vol. 27, 113-121.

[5] Nanba, H. (2007) “Query Expansion using an Automatically Constructed Thesaurus” . In Proceedings of the 6th NTCIR Workshop, 414-419.

[6] 新森昭宏、奥村学、丸山雄三、岩山真. (2004) “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.