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―

Construction of an Ontology Based on a Patent Classification System

広島市立大学大学院情報科学研究科教授 **難波　英嗣**

PROFILE 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

筑波大学大学院システム情報工学研究科助教 **乾　孝司**

PROFILE 2004 年奈良先端科学技術大学院大学情報科学研究科博士課程修了。博士（工学）。日本学術振興会特別研究員、東京工業大学統合研究院特任助教等を経て、2009 年筑波大学大学院システム情報工学研究科助教。自然言語処理の研究に従事。

日立製作所中央研究所 / 東京工業大学精密工学研究所 **岩山　真**

PROFILE 1992 年（株）日立製作所入社。文書検索、自然言語処理の研究に従事。2009 年度より、特許版産業日本語委員会委員。

東京大学大学院情報理工学系研究科ソーシャル ICT 研究センター教授 **橋田　浩一**

PROFILE 1986 年電子技術総合研究所入所。2001 年産業技術総合研究所。2013 年より東京大学大学院情報理工学系研究科ソーシャル ICT 研究センター教授。

東京工業大学大学院情報理工学研究科准教授 **藤井　敦**

PROFILE 1998 年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009 年より東京工業大学大学院情報理工学研究科准教授。自然言語処理、情報検索、Web マイニングの研究に従事。

1　はじめに

本稿では、特許データベースからオントロジーを構築する手法について述べる。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。しかし、オントロジーを人手で構築し、更新することは非常にコストがかかる。一方で、テキストデータベースからシソーラスやオントロジー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。そこで、本稿では、最小減の労力で効率的にオントロジーを構築する枠組みについて述べる。

オントロジーを効率的に構築するため、我々は特許分類コード体系のひとつである F タームに着目する。F タームとは、特許を目的・利用分野・材料といった様々な観点から分類することを目的として日本国特許庁が構築した特許の分類体系のひとつである。F タームの詳細については次節で述べるが、実は、F タームの構造そのものがオントロジーに近い体系になっている。そこで、本研究では、F タームの体系をオントロジーの構築に流用する。これをベースに、人手で構築された他の知識体系や、テキストデータベースから自動的に構築された上位・下位概念辞書や同義語辞書 [Nanba 2007] などを

マッピングすることで、特許との親和性を保持しながら、学術論文など他のジャンルの文献にも利用可能なオントロジーの構築を目指す。

本論文の構成は以下のとおりである。次節では、Fタームに基づくオントロジー構築の枠組みについて述べる。3節では、現在著者らが取り組んでいる情報分野のオントロジー構築について述べる。最後に4節で本論文をまとめる。

2 Fタームに基づくオントロジーの構築

2.1 Fタームとは

Fタームは、前述のとおり、特許を目的・効果・構成などの様々な観点から分類することを目的とした分類体系であり、技術分野を示すテーマコードと観点の集合から構成される。ここでは、機械翻訳分野のFタームを例に説明する。機械翻訳には5B091という1つのテーマコードが、また「言語」(AA00)、「処理対象要素」(AB00)、「翻訳方式」(BA00)などの9個の観点が設けられている。ある機械翻訳システムについて考えた場合、そのシステムの対象言語は何か、どんな仕組みで翻訳するのか、などの属性が存在するが、これがそれぞれ「言語」(AA00)や「翻訳方式」(BA00)などの観点に当たると考えて良い。

Fタームでは、観点が階層化されており、例えば、「言語」(AA00)という観点には、この観点を具体的に示す「・多言語間」(AA01)や「・2言語間」(AA03)といったFタームコードが存在する。Fタームコード間で一般/具体関係がある時には、ドットレベル記法で表すことになっている。図1の例では、「翻訳方式」(BA11)の下位分類として「直接翻訳」(BA12)と「間接翻訳」(BA13)があり、さらに「間接翻訳」の下位分類には「トランスファー方式」(BA14)と「ピボット方式」(BA17)がある。

2.2 Fタームに基づくオントロジーの構築

Fタームに基づくオントロジーの構築は、以下の2つの手順から構成される。

(手順1) Fタームからの知識抽出

BA11　・翻訳方式
BA12　・・直接翻訳
BA13　・・間接翻訳
BA14　・・・トランスファー方式
BA15　・・・・意味解析
BA16　・・・・・文脈解析
BA17　・・・ピボット方式

図1　テーマ"5B091(機械翻訳)"のFタームコードの例(その1)

(手順2) 手順1で得られた知識への他の知識体系のマッピング

各手順について以下に述べる。

(手順1) Fタームからの知識抽出

本研究で構築するオントロジーでは、4種類の用語間の関係「同義」「上位・下位」「属性・定義域・値域」「全体・部分」を扱う。図1のドットレベル記法では明示されていないこれらの関係の一部を自動解析し、残りを人手で判断することで、最終的に図2を得ることが、Fタームからのオントロジー構築の目的である。その詳細手順については、3節で説明する。

関係1	属性：方式 定義域：機械翻訳 値域：直接翻訳，間接翻訳
関係2	上位：間接翻訳 下位：トランスファー方式
関係3	上位：間接翻訳 下位：ピボット方式
関係4	属性：利用技術 定義域：トランスファー方式 値域：意味解析
関係5	属性：利用技術 定義域：意味解析 値域：文脈解析

図2　図1から得られる知識

(手順2) 手順1で得られた知識への他の知識体系のマッピング

手順1で抽出された知識に対し、様々な知識体系のデータをマッピングする。知識体系として利用可能なデータとして、特許データベースや特許検索履歴データから自動構築した上位・下位概念辞書 [Nanba 2007] [難波 2011] や関連語辞書 [乾 2011] などが挙げら

れる。マッピングは、例えば、図 1 の関係 2 と 3 について、「間接翻訳」の下位語として、自動構築した上位・下位概念辞書の間接翻訳の下位語を収集したり、「トランスファー方式 ピボット方式」といったような下位語集合から関連語辞書を用いて関連語を収集したりして下位語の候補を人間に提示し、その中から人手で適切な下位語を選択するという方法が考えられる。

3 情報分野のオントロジーの構築

本節では、2 節で述べた方針に基づき、著者らが現在進めている情報分野のオントロジーの構築について述べる。

3.1 情報分野の特許からの上位・下位概念辞書および属性辞書の自動構築

著者らは一般財団法人工業所有権協力センター（IPCC）から、情報分野の特許検索履歴データの提供を受けている。このデータに高頻度で出現する国際特許分類（IPC）コード G06F、G11C、G06K、G06T を情報分野と考えた。このコードが筆頭 IPC として付与されている公開年が 1993 ～ 2012 年の特許 396,532 件から、上位・下位概念辞書および属性辞書を構築した。上位・下位概念辞書の構築には、Hearst[Hearst 1992] の定型表現法を用いた。その結果、のべ 219,462 語から構成される情報分野の上位・下位概念辞書が構築された。図 3 に、その一部を示す。図において、各行の数値は、396,532 件の特許中での出現頻度、「A ＞ B」は A が B の上位語であることを示している。

6864 入力装置 ＞ キーボード
5542 記録媒体 ＞ CD
5052 入力装置 ＞ マウス
3357 ネットワーク ＞ LAN
3039 記憶媒体 ＞ CD
2906 情報処理装置 ＞ パーソナルコンピュータ
2507 通信ネットワーク ＞ インターネット
2381 入力手段 ＞ キーボード
（のべ 219,462 語）

図 3　情報分野の特許から抽出された上位・下位概念の一部

次に属性辞書の構築について述べる。一般に、ある用語 A と属性の関係にある B は、特許中で「A の B」と表記されることがある。そこで、上述の特許 396,532 件から「[名詞句] の [名詞句]」にマッチする個所をすべて抽出し、それを属性辞書の候補とした。なお、「A の B」と表現できる用語 A と B には、属性以外の関係も成立しうる。例えば「パソコンのキーボード」の場合、「パソコン」と「キーボード」は全体・部分関係にある。ここで、属性になりうる用語（目的、対象など）は、それほど種類が多いとは考えにくいため、「A の B」で表現される用語のうち、高頻度で B に出現するものは属性である可能性が高い。そこで、属性辞書から頻度の高い B を順に見て、さらに F タームコードが 00 で終わるもの（ドットがつかないテーマ中で最上位のコード）と照合し、属性となりうる用語を選定した。図 4 にその一部を示す。

目的・用途，利用分野・応用・適用分野，手段，効果 [1]，方式，対象，時期，入力・出力，構成・構造・機構・形状・利用技術・配置，機能

図 4　情報分野の特許から抽出された属性の一部

3.2 情報分野の F タームからの知識抽出

上述の情報分野に関連する IPC コード G06F、G06K、G06T、G11C を F タームのテーマの範囲に含むものを F タームリストから抽出した。その結果、F ターム全 2790 テーマ中 102 テーマが選定された。次に、これらの 102 テーマに関連する F タームコード 11,842 個を抽出した。これらのコードに対し、図 2 に示すような知識を抽出する手法について、以下に述べる。

（手法 1）上位・下位関係の推測

図 5 は、テーマ 5B091（機械翻訳）の F タームコードの例であり、この中で「形態素解析」に着目する。3.1 節で述べた自動構築された上位・下位概念辞書を用い、「形態素解析」の上位語を調べると、図 6 に示す上位語

1　「機械翻訳システムの効果」と言う場合、「効果」は、厳密には「機械翻訳システム」そのものの属性ではなく、機械翻訳システムに関する発明の属性であるが、オントロジーを構築する上で、今回はそこまで厳密な区別はしない。

リストが得られる。これらの上位語のうち、図５の一行目にある「言語処理技術」の部分文字列である「言語処理」と「技術」が図６にも出現しており、この結果から、図５の「言語処理技術」と「形態素解析」の間には上位･下位関係が成立していると推測できる。また、図６からわかるように、「A ＞ B」と「B ＞ C」が成立する時、「A ＞ C」も成立するため、これを逆に適用すると、図５において「言語処理技術 ＞ 文解析」と「文解析 ＞ 形態素解析」が成立する可能性がある、と推測できる。

言語処理技術
・文解析
・・形態素解析

図５　テーマ“5B091(機械翻訳)”のＦタームコードの例(その２)

65 解析処理 ＞ 形態素解析
52 手法 ＞ 形態素解析
41 解析 ＞ 形態素解析
40 処理 ＞ 形態素解析
23 自然言語処理 ＞ 形態素解析
20 構文解析 ＞ 形態素解析
20 言語処理 ＞ 形態素解析
17 技術 ＞ 形態素解析
14 構文解析処理 ＞ 形態素解析
12 言語解析 ＞ 形態素解析
10 単語切り出し技術 ＞ 形態素解析
9　自然言語解析 ＞ 形態素解析

図６　自動抽出された「形態素解析」の上位語

(手法２) 属性関係の推測

次に、属性関係の推測について述べる。一般的な属性については、Ｆタームコードが 00 で終わるものについて、その説明語句の中に、図４に示す用語の一部が含まれていれば、そのＦタームコードは、テーマコードに対して属性関係にあると推測できる。

(手法３) テーマコードリストからの推測

今、手法１を図７に示すＦタームコードに適用した結果、BA01 と BA02 の間に上位・下位関係が成立していることがわかったとする。この時､BA02 と隣接し、同じドットレベルの BA03 についても、BA01 との間に上位・下位関係が成立している可能性がある。

BA01　・X
BA02　・・Y ※ BA01 の下位概念
BA03　・・Z
BA04　・・・W

図７　Ｆタームコードの例

3.3　情報分野のＦタームから抽出された知識へのマッピング

ここでは特に属性語のひとつである「効果」に焦点を当てる。一般に、「効果」には「信頼性の向上」「耐久性の向上」などの表現が下位概念に該当する。筆者らは、これまでに特許の「発明の効果」の項目から効果に関する表現を抽出するシステムを開発している[福田 2013]。例えば「PM 磁束制御用コイルを設けて閉ループフィードバック制御を施すため、電力損失を最小化できる。」という文が入力されると、図８に示すように、要素技術と効果を示す個所に、それぞれ“TECHNOLOGY”および“EFFECT”タグを自動的に付与する。ここで、“EFFECT”タグの中には、さらに“ATTRIBUTE”と“VALUE”という２種類のタグを自動的に付与する。

PM 磁束制御用コイルを設けて <TECHNOLOGY> 閉ループフィードバック制御 </TECHNOLOGY> を施すため, <EFFECT><ATTRIBUTE> 電力損失 </ATTRIBUTE> を <VALUE> 最小化 </VALUE></EFFECT> できる.

図８　特許への要素技術と効果に関するタグ付与の例

このシステムを用いて 10 年分の公開公報を解析し、自動的に抽出された属性（ATTRIBUTE）と属性値（VALUE）の対を「効果表現リスト」と呼ぶ。図９は、その一例である。各行、「頻度　属性　属性値」を示しており、2,599,368 対（異なり）が抽出されている。

22393	信頼性	向上
21713	信頼性	高い
17362	構成	簡単
16870	生産性	向上
11283	作業性	向上
10522	操作性	向上
10376	コスト	低減
10175	製造コスト	低減

図９　効果表現リストの一例

4　おわりに

本稿では、特許の分類体系のひとつであるＦタームを用い、これに様々な知識体系をマッピングすることでオントロジーを構築する手法を提案した。現在、情報分野のオントロジーを構築中であり、完成後は一般公開する予定である。

謝辞

本研究を実施するにあたり、一般財団法人工業所有権協力センターから、特許検索履歴データを提供して頂きました。深く感謝致します。

参考文献

[福田 2013] 福田悟志，難波英嗣，竹澤寿幸．(2013) “論文と特許からの技術動向情報の抽出と可視化” 情報処理学会論文誌データベース，Vol.6，No.2，16-29.

[Hearst 1992] Hearst, M.A. (1992) “Automatic Acquisition of Hyponyms from Large Text Corpora” In Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545.

[乾 2011] 乾孝司，難波英嗣，橋本泰一，藤井敦，岩山真，橋田浩一．(2011) “最大クリーク探索に基づく特許検索履歴の統合” 言語処理学会 第 17 回年次大会，pp. 1059-1062.

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

[難波 2011] 難波英嗣，竹澤寿幸，乾孝司，岩山真，橋田浩一，橋本泰一，藤井敦．(2011) “特許検索履歴を用いたシソーラスの自動構築” 言語処理学会 第 17 回年次大会，pp. 900-903.