特許文書における複合語の意味関係解析

分野オントロジ構築支援システムと検索への応用

慶應義塾大学大学院政策・メディア研究科
特別研究教員准教授
内山　清子

PROFILE 1997年慶應義塾大学大学院政策・メディア研究科修士課程終了、1998年学術情報センターCOE研究員、2003年スタンフォード大学CSLI訪問研究員、2005年学術博士（慶應義塾大学）、2005年より現職

慶應義塾大学大学院政策・メディア研究科修士課程
粟飯原　俊介

PROFILE 2007年3月慶應義塾大学環境情報学部卒業、現在、慶應義塾大学大学院政策・メディア研究科修士課程在学中

慶應義塾大学環境情報学部教授
石崎　俊

PROFILE 1970年東京大学工学部計数工学科卒、同助手を経て、1972年通産省工技院電総研勤務、1985年推論システム研究室室長、自然言語研究室長を経て、1992年から慶應義塾大学環境情報学部教授、現在言語処理学会会長

1 はじめに

特許分野では知的財産戦略の推進に伴い、新しい特許や技術に関連した類似研究や先行研究の情報を効率的に検索する重要性が高まっている。特許文では特許に値する発明であることを説明するために、類似特許との差異や、新規性、進歩性を明確に記述する必要がある。

特許検索は、特許を申請する時だけでなく、新しい技術を開発する時に、技術情報や動向を調査するためにも利用される。その場合には、特許の専門家以外の人間にも使いやすい検索が求められる。広く一般的に利用することを前提として、現在の特許検索における問題点を自然言語処理の観点から考え、その問題点を解決する一手段を紹介する。

2 特許検索における問題点

ここでいう特許検索とは、特許庁が公開している特許電子図書館（IPDL）[※1]を利用した検索のことを指す。これは、IPDLのWebページから公報テキスト検索を利用して、公報種別を指定し、各検索項目（発明の名称、要約、請求の範囲、IPCなど）に含まれる検索キーワードを入力することで特許文書を500件まで表示し閲覧することができるシステムである。このシステムを利用した検索において、以下の2点に着目して問題点とその解決方法を述べる。

・検索キーワード（複合語）
・分類コード（Fターム）

※1 http://www.ipdl.inpit.go.jp/homepg.ipdl

2.1　検索キーワード（複合語）

特許文には専門用語や特許出願者が作った新しい複合語だけでなく、分野で共通に用いられる専門用語を名詞句や動詞句で表現したものなどが多く含まれる。たとえば、「機械翻訳」という用語を「自動翻訳」「機械による翻訳」や「機械で翻訳する」「機械が翻訳する」などの言い換えにより表現することができる。そのため、キーワード（専門用語）によるパターンマッチングでは、網羅的に検索することが難しい。 言い換えは同じ意味内容を伝達する言語表現が複数あることで、与えられた言語表現から様々な言い換えを自動生成する手法は自然言語処理の分野で盛んに研究されている [1]。複合語を中心とした専門用語の言い換え [2] を検索に利用するために、複合語を構成している語（以下、語基と呼ぶ）の間の関係を明記しておくことが必要である [2]。上記の例では「機械」と「翻訳」の意味関係を「機械」が「翻訳」の「動作主」あるいは「道具」の関係が成り立つことがわかれば、名詞句や動詞句に言い換えることができる。

更に、語基間の意味関係を利用して、オントロジ構築に応用可能である。たとえば、「機械翻訳」と「統計的機械翻訳」という用語がある場合、複合語のhead（中心となる語基）が両方とも「機械翻訳」で同じである。そして、「統計的」と「機械翻訳」には修飾関係が成り立つ。この場合には「機械翻訳」の下位（部分）概念として「統計的機械翻訳」を配置することが考えられる。重要度が高く、語基数が少ない複合語と、その複合語を一部（後項）に持つ複合語との関係を上位・下位関係と定義することができる（ここでは、上位・下位関係には部分・全体関係も含んでいる）。このように、語基間の関係を明記することにより、検索やオントロジ構築に応用することが可能である。

2.1　分類コード（Fターム）

まず、分類コードについては、国際特許分類（IPC）、FI（ファイルインデックス）、Fタームの3つがある。国際特許分類（IPC）は 日本を含む90ヶ国以上の国において共通の特許分類として、各国の特許文献に付与されるとともに特許文献の調査に利用されている。国内文献の検索キーであるFI（ファイルインデックス）はIPCの末尾に分冊識別記号（アルファベット1文字）や展開記号（3桁の数字）を追加して、IPCをさらに細分類した日本固有の特許分類である。Fタームは技術的観点から項目分けされているIPCとは異なり、複数の観点（発明の目的、用途、材料など）から細分類した特許分類で、特許の実態審査における先行技術調査を効率的に行うために開発された。

これらの分類コードを正しく特許文に付与してあれば、関連特許や類似特許を効率的に検索可能である。特にFタームは特定分野の専門用語で構成され、詳細な技術情報の検索に有効であり、一種の分野オントロジとしても利用可能である。 特許庁では、Fタームを技術の進展に対応し適切な検索キーとして機能するように、必要に応じてテーマ統合・分割などを行っている。しかし新しい技術情報がFターム分類に反映されにくい、あるいは専門家でないとFタームを使いこなせないなどの問題点がある。 更に、Fタームは実際の特許文にはない用語から作成されている場合があり、Fタームの専門用語を特許の検索キーワードとして利用することが難しい。Fタームを出願特許文から自動的に生成する仕組みを作れば、Fタームの追加が容易になり、Fタームを検索キーワードとして利用することも可能となる。

3　分野オントロジ構築支援システム

特定分野オントロジ（Fターム分類）構築支援システムとして、特許文書のテキスト情報から、検索のキーワ

ードとなり得る重要語を抽出し、重要語の語基間の関係と重要語間の関係を取得する手法[3]を紹介し、その検索への応用について説明を行う。

3.1 重要語の抽出

重要語候補となる複合語を抽出するために、対象とする分野の文書を検索する。特許電子図書館の公報種別が「公開特許公報」のうち、検索項目が「要約と請求の範囲」に検索キーワードを入力し、そのキーワードを含む特許文書を検索する。

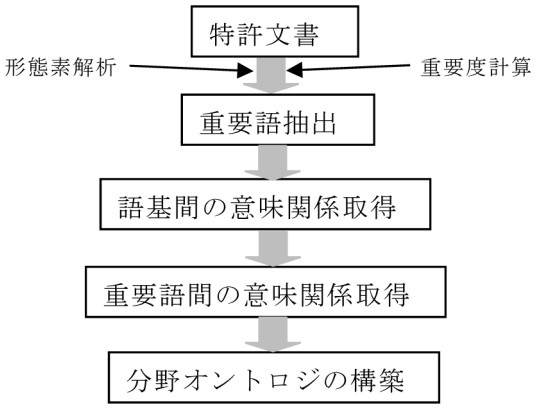

図1 分野オントロジ構築支援システム概要図

検索結果として得られた文書を形態素解析器Mecab※2のJava実装であるsen※3とIPA品詞体系辞書※4を用いて形態素解析を実行する。その中から解析時に付与された品詞が、IPA品詞体系の分類上名詞であるが、固有名詞-地域-国以外の固有名詞、代名詞、数、接頭詞、名詞接続、副詞可能、非自立は除外し、名詞が単独あるいは連続する語を抽出する。この時、名詞に連接する語で、特に特許用語に多く使用される語や接頭語として用いられる「上記、前記、当該、該、毎」等の語を複合語の構成語基には含まない語として除外することによって、効率的に重要語候補となる複合語を抽出することができる。

※2 http://mecab.sourceforge.net/.
※3 http://ultimania.org/sen/
※4 http://chasen.naist.jp/stable/ipadic/

重要度計算と特許固有表現の排除

複合語に対する重要度の算出方法は数多く提案されており、以下が代表的な指標である[4][5]。

- 複合語Wの単独出現頻度を指標とするF
- 複合語Wの単独出現頻度とWを部分文字列として含む総ての複合語の出現頻度を足し合わせたものを指標とするTF
- 複合語Wが単独に出現する文書数を指標とするDF
- DFの逆数をとり、TFと積を取ったものを指標とするTF-IDF
- TFに語基数と部分文字列の性質を取り入れた指標であるC-Value
- 連接頻度LR、連接種類LRの各指標にFを掛け合わせたFLR

上記の各手法において抽出された重要語候補の上位に、形態素解析レベル（一語基）では削除できなかった、特許文書に頻繁に用いられるが重要ではない語（特許固有表現）の二語基以上の複合語が多く含まれている。その例を表1に示す。特許固有表現、つまり特許文書全体において偏在する複合語を排除するために、情報エントロピーなどを用いてフィルタリングを行う。これにより、専門分野における重要語を効率的に抽出することができる。

表1 特許固有表現の例

請求項	特許請求	実施形態
特許文献	公報記載	使用例
ブロック図	フローチャート図	機能ブロック図

3.2 複合語の意味関係

複合語を言い換え形式と関連づけるためには複合語の構成語基間の意味関係を明示する必要がある。そこで、複合語の語基間の意味関係を表現するために、概念記述言語（CDL: Concept Description Language）を採

用した[6]．CDLは人間・社会とコンピュータとの間で意味を共有し、人間・社会とコンピュータにおける情報処理をつなぐ新しい技術の実現する手段として共通のインタフェースになるものである．XML（eXtended Markup Language）上で開発されるコンテンツ（ドキュメントを含む広義のコンテンツ）の意味構造（概念構造）を記述するための言語で、RDF（Resource Description Framework）・OWL（Web Ontology Language）と相補的関係にある．

Concept Description Language for Natural Languages（CDL.nl）に記載された「関係」概念を、複合名詞を構成する語基間の意味関係の解析に用いる．CDLの「関係」概念は、基本的に文中の単語間の関係を表現するものであるが、複合名詞の語基間の意味関係を表現することにも有効だと考えた．「関係」概念の記述力や種類を検討し、CDL仕様にフィードバックし、将来的には共通の記述仕様に従って日本語だけでなく、多言語に対応することが可能である．

CDLにおける「関係」概念は、事象内関係（格関係）、実体間関係、限定・修飾関係の3つで、その中に更に45の関係が設定されている．ここでは、45の関係の中から10を語基間および語間の意味関係とした．意味関係の判別がしやすいように言い換えパターンを用意し、関係概念とともに表2に示す。

表2　語基、語間関係の意味関係

言い換えパターン	関係概念
AがBする	Agt（agt:動作主）
AをBする	Obj（affected thing:対象）
AでBする	Met（method or means:方法）
	Ins（instrument:道具）
AなB、A（的）にBする	Mod（modification:限定）
AのB、AするB	Cnt（content、namely:内容）
A（する）ためのB	Pur（purpose or objective:目的）
Aと同様のB	Equ（equivalent:同義・類義）
AなどのB	Icl（included/a kind of:上位）
Aを含むB	Pof（part-of:部分）

3.3　語基間の意味関係

複合語の語基間の意味関係を決定するためには、言語情報として、語基の文法的属性と意味属性が重要な手がかりとなる[7]。まず文法的属性については、語基間の関係の手がかりとなる品詞情報は、漢語は従来の品詞体系の枠組みにおける名詞（サ変名詞）やナ形容詞語幹に分類される。しかし、名詞の中でも文中において主語や目的語として用いられる語と修飾的な役割を持つ語は、文法的属性が異なる。この異なる文法的属性が意味関係の決定に重要な役割を果たすと考えられる。そこで、枠組みや定義において重要となる文法情報の整理を行った。接続関係として、従来の品詞枠における名詞が文中で用いられる時に取り得る情報を以下に示す。

接続情報

(a)格助詞「が、を」

(b)サ変動詞化語尾「する」

(c)形容詞性活用語尾「な」

(d)接尾辞「的」

(a)については格助詞「が」「を」を後続可能な語は主語、目的語になることができ、また(b)の「する」を後続する語はサ変動詞になって述語として機能するなど、文の主要な役割を果たす語の情報として、また複合名詞の内部構造で格関係を成立させるために必要な要素となっている。(c)(d)は、ナ形容詞性活用語尾「な」や「的」「性」などの接尾辞との接続関係により修飾用法を持つ語について調べることができる。表3に具体的な語基の接続関係について、毎日新聞記事5年分から(a)から(d)の連接頻度を数え、各連接頻度の合計から連接比率を算出している。

表3　語基の接続情報との連接比率

単語	(a)	(b)	(c)	(d)
言語	94%	0%	0%	6%
生成	41%	59%	0%	0%
特殊	0%	0%	100%	0%
基本	25%	0%	0%	75%

この特徴を用いた意味関係解析のルールとして以下が考えられるが、連接比率を素性として機械学習させることにより自動化することも可能である。

1. a（言語）obj b （生成）言語を生成する
2. c（特殊）mod a（言語）特殊な言語
3. c（基本）mod a（言語）基本的な言語

また、先行研究[7]において、文法的属性だけでは意味関係を決定できないことを確認しているため、既存の意味体系（EDR電子化辞書、日本語語彙大系、分類語彙表など）を用いて、意味情報を含めることにより、意味関係取得の精度が向上すると考えられる。

3.4 重要語間の意味関係

重要語間の意味関係を判別するための手がかりとして、重要語に挟まれている表現（定型表現）を利用する。定型表現を用いた方法は[8]、日本語では「AなどのB」「AのようなB」と表現されるパターンを抽出することによって上位下位関係を判別することである。特許文書においてもこの方法が利用できると考えられるが、定型表現のパターンは以下の3つに分けられる。

・助詞と機能語　「などの」「といった」
・助詞と和語動詞「を含んだ」「を用いた」
・助詞とサ変動詞「を具備した」「から構成される」

語の関係を一意に且つ明確に表現している定型表現を選定し、各定型表現に対応する意関係を人手で判定・定義した例を表4に示す。パターン例には、CDLの記述方式と定型表現を含んだパターンを記述している。

「機械翻訳装置」に関する特許文を抽出し、定型表現を用いて重要語間の上位・下位（部分・全体）関係を定義した結果を有向グラフの形式に変換した一部を図2に示す。

表4　定型表現とその意味関係

定型表現	関係	パターン例
としての	上位・下位／部分・全体	前処理 >icl スペルチェック 前処理としてのスペルチェック機能
からなる	部分・全体／上位・下位	出力文 >pof単語列 単語列からなる出力文
を備えた	部分・全体／上位・下位	翻訳装置 >pof 翻訳処理 翻訳処理を備えた翻訳装置
を用いた	方法	機械翻訳装置 >met中間言語方式 中間言語方式を用いた機械翻訳装置
を記述した	内容／部分・全体	単語辞書 >cnt 意味情報 意味情報を記述した単語辞書

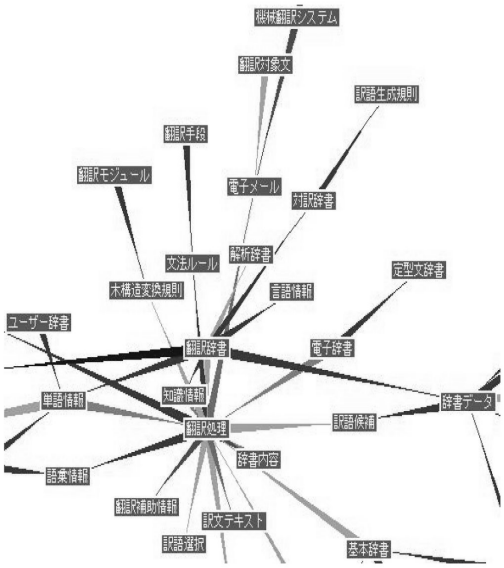

図2　重要語間関係の有向グラフ

4 特許検索への応用

分野オントロジ（Fターム）構築支援システムとして、重要語の抽出から、語基および重要語間の意味関係を取得する方法の紹介をした。これを検索へ応用するためには、検索キーワードの選択を、グラフ形式で表現した重要語間の関係を利用すれば、視覚的に操作を行うことができる。また、検索キーワードを名詞句や動詞句に展開することによって、網羅的に特許文書を検索できると考えられる。今後は、人手による分類コードの付与や、Fターム、特許マップの作成作業を、申請済みの特許文書からの知識獲得により自動化を推進し、効率的な特許検索に結びつけていくことが重要である。

参考文献

[1] 乾健太郎,藤田篤,"言い換え技術に関する研究動向",自然言語処理Vol.11, NO.5, pp.151-198. 2004.

[2] Fuyuki Yoshikane, Keita Tsuji, Kyo Kageura and Christian Jacquemin, "Morpho-syntactic Rules for Detecting Japanese Term Variation: Establishment and Evaluation", Journal of Natural Language Processing, Volume 10, NO.4, 2003.

[3] 栗飯原俊介,内山清子,石崎俊,"特許文における分野オントロジー構築のための重要複合語の抽出と重要複合語間関係の定義",言語処理学会第13回年次大会, pp.871-874(2007).

[4] 中川裕志, 森辰則, 湯本紘彰出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理研究会報告 2001-NL-145, 情報処理学会, pp.111-118 (2001).

[5] 辻河亨, 吉田稔, 中川裕志. 語彙空間の構造に基づく専門用語抽出. 情報処理学会NL研究会159, 1/2004, pp.155-162 (2004)

[6] Institute of Semantic Computing(ISeC), Concept Description Language CDL.core Specifications version1.0, ISeC Technical Report:2007-1-29(2007).

[7] 内山清子,竹内孔一,吉岡真治,影浦峡,小山照夫,"専門分野における複合名詞解析のための名詞文法属性の分類について",計量国語学会,第23巻1号,pp.1-24 (2001).

[8] Hearst, M.A. "Automatic Acquisition of Hyponyms from Large Text Corpora." In proceedings of the 14th International Conference on Computational Linguistics (Coling'92), pp.539-545(1992).