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

粟飯原 俊介‡　内山 清子†　石崎 俊‡
‡ 慶應義塾大学 環境情報学部
† 慶應義塾大学大学院 政策・メディア研究科
〒 252-8520 神奈川県藤沢市遠藤 5322
E-mail:{t03003sa,kiyoko,ishizaki}@sfc.keio.ac.jp

1　はじめに

本論文では, 分野オントロジー構築支援に向けた (1) 特許文書の重要語抽出に適した重要度計算法の検討,(2) 重要語に挟まれた助詞と動詞からなる定型表現の抽出と定型表現を用いた重要語間関係の自動獲得の手法を提案する. 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. 特許文書を検索するために国際特許分類 (IPC), 日本固有の分類である FI と, 発明の手法や目的などの複数の観点から再分類した F タームが用意されている. 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ーとして利用可能である. 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. 特定分野のオントロジー (F ターム分類) 構築支援として, 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複合語を抽出し, 重要複合語間の関係を自動獲得する手法が必要である.

重要語抽出は従来から様々な手法が検討され [1, 2], 良好な成果を挙げている. 特許分野において, 効率的に複合語を含めた重要語を抽出する手法を選別するために, 先行研究の手法を比較すると同時に, 分野の異なる特許文を用いることで, 特許分野に共通する特有の表現をフィルタリングする手法について検討する. 次に, テキストから語の関係や関連語を自動獲得する研究として, 一般用語では (1) 上位・下位関係を表す定型表現を用いる方法 [3, 4],(2) 言語的パターン,HTML タグ, 単語の出現に関する統計値を利用する方法 [5], 専門用語では (3) 特定の専門分野における専門用語と強く関連する語を収集する手法 [6],(4) テキストからの言い換え表現 (用語異形) の抽出に基づく方法 [7] が行われてきた. これらの研究は言語的パターン, たとえば,「A などの B」「A を B する」といった語の関係を表す定型表現を事前に用意し, テキスト中からそのパターンに該当する語のペアを抽出することにより, 語彙情報を獲得している. 従来の研究に用いられる定型表現は, 主に上位・下位, 同義・類義が中心であったが, 特許文書を複数の観点で分類するためには, 目的, 用途, 部分全体等の関係も抽出することが重要となってくる. 本論文では, 目的, 用途, 部分全体関係を決定する表現を特許文から抽出し, その定型表現と意味関係の対応付けをルール化し, 重要複合語間の関係を自動獲得する手法について検討する.

以下 2 章では, 重要語を抽出する複数の手法を比較することにより, 特許文書における効率的な重要度計算を検討し,3 章では特許文書のテキストから定型表現に利用して,2 章で抽出した重要複合語間の関係を獲得する方法を提案し, 最後に今後の課題を述べる.

2　特許文書群からの重要複合語抽出

2.1　複合語の抽出

特許文書からの重要複合語を抽出するために, 特許庁が公開している特許電子図書館の Web ページから公報テキスト検索を利用し, 公報種別が「公開特許公報」のうち, 要約・請求項に「機械翻訳装置」を含む 459 文書を収集し, 対象テキストとした. この文書を形態素解析器 Mecab[8] の Java 実装である sen[1] と IPA 品詞体系辞書[2] を用いて形態素解析を実行した. 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, 名詞-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, 非自立以外であるものの,

[1] http://ultimania.org/sen/
[2] http://chasen.naist.jp/stable/ipadic/

単独あるいは連続する語を抽出した．この時, 特許用語に多く使用される「上記, 前記, 当該, 該, 毎」等を排除した．その結果, 語基を 5370 種,2 語基以上から成る複合名詞を 16890 語抽出した．特許文書においては, 単名詞そのものが重要語となりうる場合は少ないと考えられるため, 語基自体は重要度の指標としては利用するが, スコアリングの対象からは除外した.

2.2 スコアリング

複合語に対する重要度の指標として多くの提案がなされているが, 今回は 2.2.1 に記す指標に関して実験を行い, 特許文書における重要複合語の抽出手段として適切だと思われる指標を選定する.

2.2.1 各手法ごとの比較と考察

正解セットは, 機械翻訳に関する二語基以上の用語を (1) 特許分野では F ターム (5B091) と FI ターム (G06F17/27-17/28@Z),(2) 研究分野では機械翻訳に関する文献 [10] から抜粋して作成した. 以下の各指標に関して, 算出した重要度順にソートした候補語の上位 n 件における完全一致数, 適合率, 再現率,F 値を求め, 評価を行った.

1. 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

2. 複合語 W が単独に出現する文書数を指標とする DF

3. DF の逆数をとり、TF と積を取ったものを指標とする TF-IDF

4. 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[9]

5. 連接頻度 LR, 連接種類 LR の各指標に F を掛け合わせた FLR[1]

図 1 に完全一致語数の変化を示し, 図 2 に、F 値の変化を示す. また表 1 では, 候補語 500 語,1000 語,1500 語,2000 語ごとの, 各手法における, 適合率, 再現率,F 値を示す. この結果からは各手法においてそれほど顕著な差は見出せなかったが, 完全一致数と F 値ともに DF が比較的良好な値を示している. 完全一致数においては良好な値を示している連接頻度・連接種類 FLR は F 値ではあまりよい値が得られなかった.

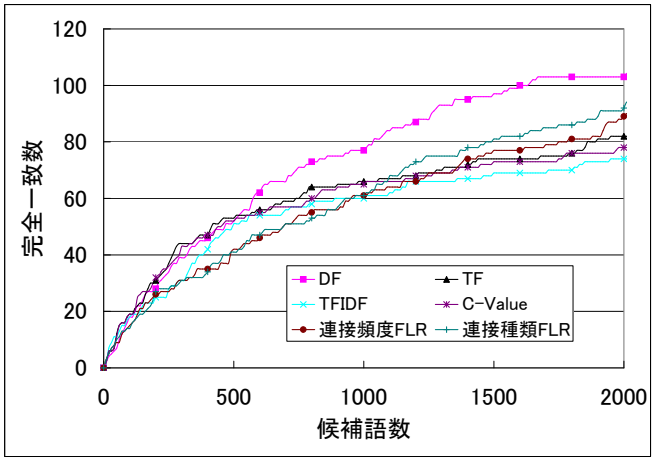

図 1: 抽出した重要語候補数と完全一致数の関係

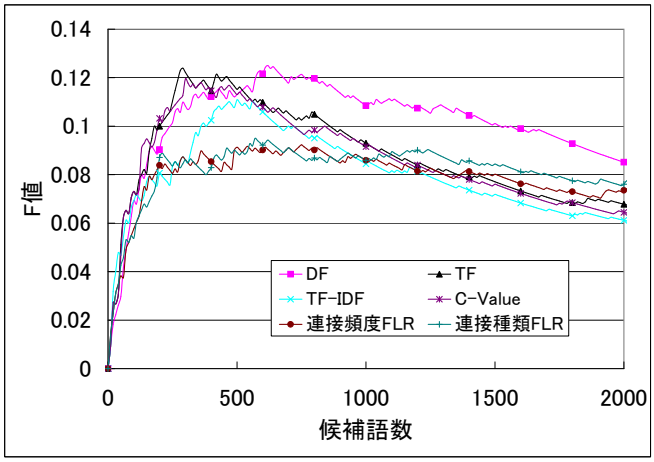

図 2: 抽出した重要語候補数と F 値の関係

表 1: 各手法ごとの適合率, 再現率,F 値

候補語数	DF	TF	TF-IDF	C-value	連接頻度 FLR	連接種類 FLR
	0.104	0.106	0.102	0.104	0.084	0.082
500	0.124	0.126	0.124	0.121	0.1	0.098
	0.113	0.115	0.111	0.113	0.091	0.089
	0.077	0.066	0.06	0.065	0.061	0.061
1000	0.183	0.157	0.155	0.143	0.145	0.145
	0.108	0.093	0.085	0.092	0.086	0.086
	0.065	0.049	0.046	0.049	0.051	0.054
1500	0.231	0.176	0.174	0.164	0.183	0.193
	0.101	0.077	0.072	0.076	0.080	0.084
	0.052	0.041	0.037	0.039	0.045	0.046
2000	0.245	0.195	0.186	0.176	0.212	0.219
	0.085	0.068	0.061	0.064	0.074	0.076

2.3 特許文書固有の表現の排除と分野重要語の決定

各手法において抽出された重要複合語候補の上位に, 形態素解析レベルでのストップワード定義において削除できなかった特許文固有の二語基以上の複合語が多く含まれている. 特許文書においてそのような偏在する複合語を除去をする必要がある. その際の指標としては情報エントロピーを用いた.

ある分野における語の (正規化した) 情報エントロピーは, 語が文書 i において出現する確率を P_i, 分野内

の文書数を N とすると以下のように定義できる.

$$H = -\frac{1}{\log N}\sum_{i=1}^{N} P_i \log P_i \qquad (1)$$

$logN$ で割ることにより $0 \leq H_i \leq 1$ に正規化され, 単語が各文書に等確率で出現するほど 1 に近い値となる. また一文書のみにしか出現しない語のエントロピーは 0 となる. この値が複数の特許分野において共通して高い語を特許文書の固有語として扱い, 重要複合語の候補内から排除する.

重要語のスコアリングに関して比較的良好な値を示した DF を, 本タスクにおいての重要度の指標として採用した. 機械翻訳分野において抽出した語と, 公報種別が「公開特許公報」のうち, 要約・請求項に「ロータリエンジン」ないしロータリーエンジン」が含まれている特許文 242 文書における語と比較を行い, 情報エントロピーがともに 0.01 以上の値を持つ語 132 語を除去し,DF を用いて重要度順に並べた際に上位にくるものを機械翻訳分野における重要複合語とした. 除去した語の一部を表 2 に示す.

表 2: フィルタリングした語例

解決策	構成図	フローチャート図	特許請求
請求項	特許文献	ブロック図	数値計算
使用例	実施形態	利用分野	公報記載

3 重要語間関係の獲得

3.1 意味関係の種類

重要複合語間の意味関係を表現するために, 概念記述言語 (CDL: Concept Description Language)[12] の CDL.nl にある関係概念を使用した. 本論文では 45 個の関係概念のうち,(1)Agt (agent:動作), (2)Obj (affected thing:対象), (3)Met (method or mean:方法),(4)Equ(equivalent:同義) (5)Cnt (content, namely:内容), (6)Pur (purpose or objective:目的),(7)Icl(included/a kind of:上位),(8)Pof(part-of:部分) を選択した. 但し,Equ(同義) には類義関係を含め,UsedFor/By(用途) を定義する関係概念がないため,Pur(目的) に含めることとした. この CDL の関係概念は文を構成する単語間の関係を記述するものであるが, 重要複合語間や重要複合語の語基間の定義にも有効であると考え [13, 14], 今後はそれらの概念の再定義も検討している.

3.2 パターンルールの作成

パターンルールは 3 ステップ (1) 上位オントロジー用語の選定,(2) 語の関係を特定できる定型表現の抽出,(3) 定型表現の選定と意味関係の付与, に基づいて作成した. まず 2 章で抽出した重要複合語は, 機械翻訳分野における特許文書に広く分布している語のため, オントロジーの上位層に位置する語であると考え, 上位 200 語を上位オントロジー用語候補とした. 次に, 重要複合語で挟まれた語のパターンを抽出するために, 従来の助詞や機能語 (「A などの B」「A としての B」) の他に, 助詞＋名詞 (サ変接続) ＋動詞 (サ変スル) ＋助動詞 (基本形) の形式を加えて検索することにした. 上位 200 語が前記の定型表現を挟んでいる重要複合語のペア (重要複合語＋定型表現＋重要複合語) を収集した. 最後に頻出する定型表現の中から, 語の関係を一意に且つ明確に表現している定型表現を 25 形式選定した. 各定型表現に対応する意味関係を人手で判定しルール化を行った. ルールの例を表 3.2 に示す.

表 3: 定型表現と意味関係のパターン

定型表現パターン	関係	関係方向	例
A からなる B	pof	B → A	単語列からなる出力文
A としての B	equ	B → A	目的言語としての英語文
A に変換する B	ins	B → A	目的言語に変換する機械翻訳方法
A に翻訳する B	ins	B → A	英語文に翻訳する翻訳プログラム
A を格納した B	pof	B → A	語彙情報を格納した単語辞書
A を含む B	pof	B → A	構文解析を含む機械翻訳装置
A を記憶する B	pof	B → A	単語情報を記憶する記憶領域
A を備えた B	pof	B → A	翻訳処理を備えた翻訳装置

また, 本論文におけるオントロジーという用語は, 特定分野の用語 (専門用語) 間に成り立つ意味関係を記述した用語集合を指すため, シソーラスに近い構造を持つと考え, 上記のルールに基づいた重要複合語間の関係を有向グラフの形式に変換した一部を図 3 に示す.

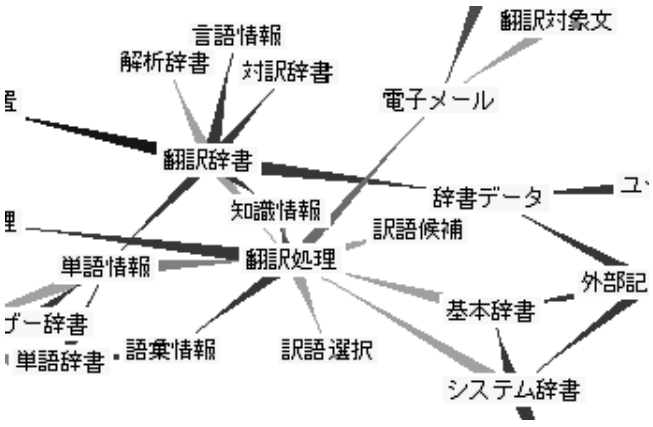

図 3: 重要語間関係の有向グラフ

3.3 ルールの適用と考察

上位層の用語間の関係定義に基づいて, 前節で作成したルールを利用して, 上位語 2000 語までの用語との関係について調べた. ルール化で使用した 154 パターン

を除いて 233 パターンを抽出し, 意味関係の正誤を判定した. 233 パターン中 4 パターンのみ不正解であった. 原因としては定型表現の曖昧性が考えられる. たとえば,「定型文としての辞書登録」における定型文と辞書登録の関係は, ルールにより同義 (equ) と判定されたが,「定型文を辞書登録する」という意味のため, 対象関係 (obj) となる. 今回, 同義表現抽出に用いられる「という」「といった」は曖昧性により意味を一意に決定できなかったため, ルールに含めなかった.「としての」はルール作成段階では一意に決定できたが, 多くの事例を参照すると曖昧性を生じる可能性が高くなる機能語だと考えられる. 助詞と動詞を含む定型表現から重要複合語間の意味関係を特定するルールを作成したが, 機械翻訳という限定された分野を対象とした定型表現に過ぎない. つまり, 異なる分野において今回利用した定型表現が現れる可能性は低いと考えられる. しかし定型表現の動詞に着目すると, 数種のカテゴリーに分類できる. たとえば「生成する」「変換する」「翻訳する」は, ある状態から別の状態に変化させる意味を持ち,「有する」「具備する」「備える」「含む」は所有を表す語である. カテゴリ毎に定型表現に用いられる動詞を用意しておけば, 異なる分野においてもルールを汎用的に利用できると推測する.

今回のルールは関係表現を絞り込んでいるため, 同じ重要複合語ペアで異なる定型表現（出力手段を備えた機械翻訳装置, 出力手段を有する機械翻訳装置など）があった場合でも, 全て意味関係が一致していた. しかし, ルールを拡張する段階で, 複数の関係が成立する可能性がある. 関係性における優先度も今後考慮していきたい.

4 おわりに

本論文では, 特許文書を対象として分野オントロジーの上位語候補となり得る重要複合語を複数の手法を検討して抽出した. 抽出した重要複合語の間に含まれる助詞と動詞からなる定型表現に着目し, 語の関係を一意に決定可能な定型表現を決定し, ルール化した. 上位 200 語で作成したルールを 2000 語までの用語に適用したところ, 重要複合語間関係の自動獲得において高い精度を得た. 今後は獲得したい意味関係と動詞との関連を分析し, 定型表現の拡張に基づいた汎用的ルールを用いて, 複数の分野を対象として実験を行い, 本手法の有効性を検証していきたい. また, 語の関係定義の客観性と信頼性を高めるために, 獲得した語の関係を複数の被験者に提示して評価する方法を検討していく.

謝辞

この研究は総務省委託研究「戦略的情報通信研究開発推進制度 (SCOPE)」により実施しました.

参考文献

[1] 中川裕志, 森辰則, 湯本紘彰 (2001). 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.

[2] 辻河亨, 吉田稔, 中川裕志 (2004). 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.

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

[4] 安藤まや, 関根聡, 石崎俊 (2003). 定型表現を利用した新聞記事からの下位概念単語の自動抽出. 情報処理学会研究報告, 2003-NL-157, pp77-82.

[5] 徳永耕亮, 風間淳一, 鳥澤健太郎 (2006). 属性語の Web 文書からの自動発見と人手評価のための基準. 自然言語処理,Vol.13, No.4,pp.49-67.

[6] 佐々木靖弘, 佐藤理史, 宇津呂武仁 (2006). 関連用語収集問題とその解法. 自然言語処理,Vol.13 No.3,pp.151-175.

[7] Yoshikane, F., Tsuji, K., Kageura, K. and Jacquemin, C. (2003). "Morpho-syntactic Rules for Detecting Japanese Term Variation: Establishment and Evaluation". Journal of Natural Language Processing, Vol.10, No.4, pp.3-32.

[8] 工 藤 拓, MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.sourceforge.net/.

[9] Frantzi, K. and Ananiadou, S. (1996). Extracting nested collocations. COLING '96, pp. 41 . 46, 1996.

[10] 長尾真, 黒橋禎夫, 佐藤理史, 池原悟, 中野洋 (2004). シリーズ 言語の科学 9 言語情報処理. 岩波書店.

[11] 北研二 (1999). 言語と計算 4 確率的言語モデル. 東京大学出版会.

[12] Zhu, M., Uchida, H. and Yokoi, T. (2006). Concept Description Language Specifications and Simple Syntax of CDL.nl. ISeC Technical Report.

[13] 内山清子, 石崎俊 (2006). 特許文に含まれる複合名詞の解析. 言語処理学会第 12 回年次大会発表論文集, pp.1107-1110.

[14] 内山清子, 石崎俊 (2006). 特許文に含まれる複合語の処理と検索への応用. 特許情報活用の時代の検索と機械翻訳技術, Japio 年誌 (財団法人 日本特許情報機構), pp.90-93.