論文用語の特許用語への自動変換

釜屋 英昭[1]　難波 英嗣[1]　竹澤寿幸[1]　奥村 学[2]

1 広島市立大学大学院 情報科学研究科　　2 東京工業大学 精密工学研究所

１．はじめに

近年，知的所有権に対する関心が高まり，企業はもちろん，個人が特許を取得するケースも増加してきている．特許出願の際には過去に同様の出願技術が存在していたかどうかの確認作業が必要不可欠である．特許庁の審査官や企業のサーチャーが行なうこの作業を無効資料調査と呼ぶ．無効資料調査では，特許と論文データベースの両方を個別に検索する必要がある．また，大学などの研究者においても，特許出願が研究活動のひとつとして重視されるようになってきており，研究者が特許と論文を検索する機会が増えつつある．しかし，特許では請求範囲をなるべく広く確保するため，一般性の高い特許用語を用いて記述する傾向がある．例えば「DRAM」は「半導体記憶装置」と記述される．このため，単純に表層的な単語の一致度を見るだけである従来の検索モデルでは，同じキーワードで特許データベースと論文データベースを検索しても，用語の使われ方の違いから，そのキーワードに関する論文や特許を十分に収集できるとは限らない．そこで本研究では，特許，論文間の引用関係及び用語間の上位，下位関係に着目し，論文用語から特許用語へ自動変換を行う．

本論文の構成は以下のとおりである．次節では，関連研究について述べる．3 節では，論文用語の特許用語への変換手法を提案する．4 節では，提案手法の有効性を調べるために行った実験について述べる．

２．関連研究

これまでにも特許を対象とした数多くの検索システムが構築されてきたが[Iwayama 2006][Fujii 2007]，近年では特許だけでなく，学術論文も横断的に検索できるシステムの開発やサービスの提供が始まっている．Thomson 社の ISI CrossSearch では，様々な分野の学術雑誌，国際会議の会議録，世界 40 ヶ国の特許発行機関から収集した特許データベースなどを検索することができる．富士ゼロックス社の DocuPat では，日米特許データ 1,800 万件と科学技術振興機構(JST)が提供する科学技術文献データ 2,000 万件を一つのインタフェースで検索することが可能である．しかし，これらのサービスでは特許と論文用語の変換機能は提供されていないため，あるテーマに関する特許と論文を網羅的に収集するには，ユーザ自身が特許と論文用語の違いの問題を解決する必要があった．

この問題に対し，我々はこれまでに，用語の変換とは別の側面から取り組んできた．近年，特許中で関連論文を，逆に論文において関連特許を引用するケースが増えているが，このような文書間の引用関係をたどれば，論文や特許と関連する文書を集めることができる．そこで我々は特許と論文間の引用関係の解析に取り組んできた [安善 2005, 2006]．ただ，現状では，特許中の引用文献の中で論文が占める割合と，論文中の引用文献の中で特許が占める割合は数パーセント程度であるため，あるテーマに関する特許と論文を網羅的に収集するのに，引用関係をたどるだけでは限界がある．そこで，特許，論文間の引用関係に加え，論文用語の特許用語への変換にも取り組み，特許，論文データの効率的な検索環境の構築を目指す．

３．論文用語の特許用語への自動変換

3.1 特許，論文間の引用関係を用いた用語変換

本研究では，安善らの手法[安善 2005, 2006]で得られた特許，論文間の引用関係データを用い，以下の手順で，論文用語を特許用語に変換する．なお，この手法を引用手法と呼ぶ．

（1）システムに論文用語を入力
（2）システムは，入力された用語を表題に含む論文をデータベースから検索
（3）手順２で検索された論文と引用関係にある特許を収集
（4）手順 3 で収集された特許から用語を抽出し，頻度順にならべ，出力

ここで，手順４において，特許中のどの個所から用語を抽出するのかを検討する必要がある．本研究では，特許から用語を抽出する際，請求項に着目する．請求項とは，「特許を受けようとする発明を特定するために，必要と認める事項のすべてを記載した項」のことであり，特許明細書の中で最も重要な個所である．また，請求範囲をなるべく広く確保するため，請求項では一般性の高い特許用語を用いて記述されるという特徴がある．

操作手段によりアクチュエータを駆動して所望の作業を行う**作業機**において，前記作業の作業機構に作成する負荷を検出する負荷検出手段と，・・・省略・・・，この変調手段の出力に応じて振動を発生する振動発生手段とを設けたことを特徴とする**作業機の操作用仮想振動生成装置**

図 1 請求項の例(特開平 10-011111 より引用，強調，省略および下線筆者)

図 1 は，請求項の一例であるが，この例から分かるように，請求項は慣例的に長い1文で記載されるため，請求項すべてから用語の抽出を行うと，その中に不要な語が多く含まれてしまう．

そこで，新森らの提案する請求項の構造解析手法[新森 2004]を用い，請求項の主要箇所を特定し，そこから用語を抽出する．それらを，予め用意しておいた不要語句リストと照合し，一致するものを出力候補から削除し，残ったものを頻度順で出力することで与えられた論文用語に対する特許用語の変換を実現する．

この他，特許中の請求項間の関係にも着目する．特許中には，複数の独立請求項(他の請求項を引用しない請求項)と，各独立請求項を引用する従属請求項が存在する．これまでの研究において，独立請求項とその従属請求項を使った時に最も高い精度が得られた[釜屋 2006]．そこで，今回は，独立請求項として第一請求項(特許中にある複数の請求項の中で，最初に記載されているもの)とその従属請求項を用いる．

3.2 用語間の上位下位関係を考慮した用語変換

3.1 節でも述べたように，特許では，請求範囲をなるべく広く確保するため，一般性の高い特許用語を用いて記述される．つまり，特許用語の多くは論文用語の上位用語であると考えられる．そこで引用手法とは別に，特許シソーラスを用いた上位語の収集による手法を提案する．このシソーラスは「A や B などの C」という，「などの」「等の」の 2 種類の定型表現に着目し，特許公開公報(1993～2002 年)から，これらの表現を含む文を収集している．収集してきた文から上位・下位関係は出現頻度で重みをつけ，約 700 万件得ることができた[難波 2007]．

以下の手順で上位下位シソーラスを用いた特許用語の収集を行う．なお，この手法をシソーラス手法と呼ぶ．

(1) ユーザが専門用語を入力
(2) システムは，入力された用語を下位語としている用語を上位下位シソーラスから収集
(3) 2 で得られた用語セットを頻度で並べ，出力

3.3 Mase 手法を用いた提案手法の改良

特許明細書の「符号の説明」という項目には，「磁気記憶装置(フロッピーディスク)」といった記述が数多く存在する．Mase ら[Mase 2005]は，このような記述から，「磁気記憶装置」と「フロッピーディスク」といった関連用語対を抽出している．この手法は，本研究においても有効であると考えられる．そこで，Mase 手法を実装して調べた結果，いくつかの入力用語に対しては，引用手法よりも高い精度で変換できることが確認されたが，全く用語が見つからないといった場合も多数あった．

そこで，入力された用語が Mase 手法により変換された場合において，その用語を考慮することで提案手法の改良を行う．Mase 手法によって，例えば「磁気記憶装置」や「リムーバブル記憶装置」といった出力が得られた場合，用語の末尾の名詞に着目すれば，入力用語は何らかの「装置」に関する用語ではないかと推測される．このような場合には，提案手法で得られた結果の中で用語の最後が「装置」のものは，他の用語よりもスコアを上げて用語の出力順序を変えることで，提案手法の改良を行う．

以前の研究では，Mase 手法の単純な頻度を加味するだけで，各提案手法に対する Mase 手法の比重を考慮していなかった[釜屋 2007]．そこで，今回は，引用手法，シソーラス手法それぞれに対し，最適となる Mase 比重を求める．

4．評価実験と結果

実験に用いるデータ

実験には特許公開公報(1993～2002 年)を用いる．特許，論文間の引用関係データは，安善の手法[安善 2005, 2006]を用いて抽出した特許中の引用論文の書誌情報約 85,000 件を用いる．

正解データセット

正解データセットは以下の手順で作成した．

1. 特許中で引用されている論文の書誌情報 85,000 件中から名詞句を抽出し，頻度順に並べる．
2. その中から論文用語 25 語を人手で選択する．
3. 論文用語毎に「比較手法」で述べた手法 3 を用いて請求項中のすべての名詞句を抽出し，頻度順に出力する．
4. その中から人手で正解判定を行う．

手順 2 で選択された論文用語の一部を以下に示す.

> CPU, 半導体レーザ, DRAM, メモリセル, ワードプロセッサ, ノボラック樹脂, CD

なお, 正解判定を行う際, 以下の点を考慮した.

[基準 1] 概念的に最も近い用語のみ正解

例えば,「ワードプロセッサ」という論文用語に対して,「文書編集装置」を正解とし, ワードプロセッサの構成要素である「表示装置」は不正解とした.

[基準 2] 特許データベース中の文書頻度

ある用語の文書頻度が特許データベース中で極端に低い場合は, その用語は特許検索を行う上で有用でないと考え, 不正解とした.

[基準 3] 基準 1 で選択されたものとの比較

ある用語が基準 2 を満たさない場合でも, その用語が基準 1 で選択されたものと概念的にほぼ等しいと判断される場合, 低頻度でも正解とした. 例えば,「ワードプロセッサ」に対して,「文書編集装置」と概念的にほぼ等しい「文書作成装置」も正解である.「レーザ」と「レーザー」のような表記のゆれについても, 一方が正解と判定されていれば, もう一方も正解とした.

評価尺度

評価には, 以下に定義される ε という尺度を用いる[清田 2004]. この尺度では, システムが出力結果の上位に数多くの正解を出力すれば, 1に近い評価値が得られる. この他に再現率と精度でも評価を行う.

$$\varepsilon = \frac{\sum_{i \in R} \frac{1}{i}}{\sum_{j \in \{1,2,\cdots n\}} \frac{1}{j}}$$

i : システムが抽出した正解の順位番号
j : 正解の数
R : システムの出力

比較手法

(1) 引用手法を用いて用語を抽出
(2) (1)に Mase 手法を用いて改良し, 用語を抽出
(3) シソーラス手法を用いて用語を抽出
(4) (3)に Mase 手法を用いて改良し, 用語を抽出
(5) (2), (4)を合成して, 用語を抽出
(6) 入力された用語と共起する用語を抽出(GETA)
(7) 同義語抽出手法で用語を抽出(synonym)
(8) JST シソーラス[1]を用いて用語を抽出

[1] http://jois.jst.go.jp/JOIS/html/thesaurus_index.htm

手法(1)～(5)が提案手法であり, (6)～(8)がベースラインである. なお, (2), (4)の事前実験である Mase 比重実験, (5)の事前実験である合成比率実験については, 紙面の都合上省略している. これらの事前実験より, (2)の Mase 比重 0.8, (4)の Mase 比重 0.2, (2)と(4)の合成比率 3 対 7 という結果が得られた. 各比較手法における ε の評価を表1に示す. また, 表2に, 提案手法で結果の良かった(2),(4),(5), ベースラインで結果の良かった(7), さらに, Mase 手法単独の結果と, システムが理想的な出力を行った場合の数値を比較した. なお, 理想値の精度が 1 となっていないのは, 論文用語に対する正解の平均が 2.93 用語であるからである. 出力上位 5～20 での ε , 再現率, 精度を示す.

表1 ε による各手法の評価

提案手法					ベースライン		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0.14	0.17	0.23	0.24	0.30	0.01	0.06	0.05

表2 ε , 再現率, 精度による各手法の評価

			top5	top10	top15	top20
提案手法	(2)	ε	0.153	0.150	0.156	0.157
		recall	0.169	0.242	0.297	0.311
		precision	0.115	0.073	0.058	0.047
	(4)	ε	0.213	0.235	0.239	0.240
		recall	0.274	0.362	0.393	0.399
		precision	0.145	0.104	0.078	0.061
	(5)	ε	**0.261**	**0.286**	**0.292**	**0.298**
		recall	**0.309**	**0.421**	**0.459**	**0.533**
		precision	**0.170**	**0.121**	**0.092**	**0.076**
ベースライン	(7)	ε	0.053	0.055	0.057	0.058
		recall	0.079	0.087	0.101	0.104
		precision	0.053	0.038	0.037	0.035
	Mase	ε	0.083	0.097	0.106	0.107
		recall	0.108	0.172	0.246	0.264
		precision	0.072	0.061	0.055	0.045
理想値		ε	1	1	1	1
		recall	1	1	1	1
		precision	0.587	0.294	0.196	0.147

考察

事前実験である Mase 比重実験, 引用・シソーラス手法合成比率実験の考察を行う. 引用手法の特徴として, トップの用語の重みに対し, その他の用語の重

みが比較的緩やかに減少している. それに対し, シソーラス手法では, トップの用語の重みに対し, 2 位以下の用語の重みが急激に減少するという特徴が見られた. そのため, Mase 比重実験において, 引用手法のトップの重みに対し, Mase 比重を 0.8 と高めに設定しても, 引用手法自体の重みを考慮したまま, 改良できたと思われる. また, シソーラス手法では, トップと下位の重みの差が非常に大きいため, Mase 比重を 0.2 と低めに設定することで, シソーラス手法自体の重みを考慮したまま, 改良できたと考えられる.

引用手法とシソーラス手法の合成比率の実験においては, 上述したような, 各手法のトップとそれ以下の用語の重みの割合の差があるために, 引用手法とシソーラス手法を 3 対 7 の割合で合成した場合に各手法の順位, 重みの両面において, つり合いが取れたのではないかと思われる.

最終的な比較実験を行った表 1, 2 より, 提案手法がどれもベースラインを上回った. さらに, 引用手法, シソーラス手法を合成したものが一番良い結果となり, 各手法を合成することで, 高精度かつ, 網羅的に特許用語が収集可能になったと考えられる.

5．おわりに

本研究では, 特許, 論文間の引用関係及び用語間の上位下位関係に着目し, 論文用語を特許用語に自動的に変換するシステムの構築を行った. 提案手法の有効性を確認するために行った実験の結果, Mase 手法による改良, 引用手法とシソーラス手法の合成が有効であるということが分かった.

6．今後の課題

実験結果から, 提案手法のある程度の有効性は確認できた. なお, 今回は引用手法とシソーラス手法を合成する際に, トップの重みを基準として正規化し, 合成を行ったが, 今後は他の合成方法についても検討していく必要がある. また, 出力された用語そのものの評価だけでなく, 出力された用語リストを使って検索や分類等のタスクを実行し, その検索または分類精度によって用語リストの質を測る必要がある. 例えば, NTCIR-7 特許マイニングタスク[藤井 2008]では, 特許と論文を対象にした検索や技術動向分析など, 様々な目的に利用可能な言語処理技術の開発を目指しており, このような技術に本研究の結果を用いた際に, どのぐらいの精度が得られるのかといった評価方法についても検討していく必要がある.

参考文献

[安善 2005] 安善奈津美, 難波英嗣, 相沢輝昭, 奥村学“特許、論文データベースを統合した検索環境の構築”情報処理学会 研究報告, NL-168, pp.21-26, 2005.

[安善 2006] 安善奈津美, 難波英嗣, 相沢輝昭, 奥村学“特許、論文データベースを統合した検索環境の構築”言語処理学会 第 12 回年次大会, pp.743-746, 2006.

[Fujii 2007] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. “Introduction to the Special Issue on Patent Processing.” Information Processing & Management, Vol.43, No.5, pp.1149-1153, Sep. 2007.

[Iwayama 2006] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. “Evaluating Patent Retrieval in the Third NTCIR Workshop.” Information Processing & Management, Vol.42, No.1, pp.207-221, Jan. 2006.

[釜屋 2006] 釜屋英昭, 難波英嗣, 相沢輝昭, 奥村学“特許, 論文間の引用関係を用いた論文用語の特許用語への変換”言語処理学会 第 12 回年次大会, pp.723-726, 2006.

[釜屋 2007] 釜屋英昭, 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山“特許, 論文間の引用関係を用いた論文用語の特許用語への変換”情報処理学会 自然言語処理研究会, NL-178, 97-102, 2007

[清田 2004] 清田陽司, 黒橋禎夫, 木戸冬子“自動抽出した換喩表現を用いた係り受け関係のずれの解消” 自然言語処理, Vol.11, No.4, pp.127-145, 2004.

[Mase 2005] Mase H, Matsubayashi T, Ogawa Y, Yayoi T, Sato Y. and Iwayama M. “NTCIR-5 Patent Retrieval Experiments at Hitachi,” Proc. of NTCIR-5 Workshop Meeting, pp.318-323, 2005.

[難波 2005] 難波英嗣“論文間の引用情報を利用した関連用語の自動収集”言語処理学会 第 11 回年次大会, 2005.

[難波 2007] 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山“特許データベースからのシソーラスの自動構築”言語処理学会 第 13 回年次大会, pp.1113-1116, 2007.

[新森 2004] 新森昭宏, 奥村学, 丸川雄三, 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.