様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 6 月 7 日現在

機関番号：25403
研究種目：若手研究（B）
研究期間：2010～2012
課題番号：22700154
研究課題名（和文）
同義語抽出手法を利用した論文用語の特許用語への自動変換および情報検索への応用
研究課題名（英文）
Automatic Translation of Scholarly Terms into Patent Terms Using Synonym Extraction Techniques and Its Applications to Information Retrieval
研究代表者
難波　英嗣（NANBA HIDETSUGU）
広島市立大学・情報科学研究科・准教授
研究者番号：50345378

研究成果の概要（和文）：
本研究では，論文中で使われる用語(特許用語)を特許中で使われる用語(特許用語)に自動変換する手法を提案する．これは，たとえば「ワードプロセッサ」という論文用語を入力すると，「文書編集装置」や「文書作成支援装置」といった特許用語に自動変換する技術のことである．ユーザがある分野の特許と論文を同時に検索する際にこの技術を用いれば，その作業を支援することが可能になる．
研究成果の概要（英文）：
In this research, we propose a method to convert scholarly terms into patent terms (e.g. converting “word processor” into “document editing device” or “document writing support system”). The method is useful when users search both research papers and patents in a particular field.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,000,000	300,000	1,300,000
2012 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：情報検索，言い換え，文書分類，特許，論文，技術動向分析

１．研究開始当初の背景

近年，学術情報量が爆発的に増加し，専門家は自分の専門分野の最新動向を把握するために，絶えず膨大な量の文献を読まなければならない状況に直面している．また，研究分野の専門分化に伴い，ある分野の知識を得るために，さらに複数の別の分野についても知らなければならないということも，もはや

一般的になりつつある．バイオテクノロジー，半導体，情報科学のように研究・開発・製品化のサイクルが非常に短い分野では，論文だけでなく，特許等，他のジャンルの文献にも注意を払う必要がある．しかし，特許では権利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．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．また，検索結果の文献数が膨大な場合，その全てに目を通すことは困難である．

そこで，本研究では，特許と論文間の引用関係などを利用して，学術用語(例：「フロッピーディスク」)を特許用語(例：「磁気記録媒体」)に自動的に変換し，日本語論文と特許を横断的に検索する研究を行っており，これまでにプロトタイプシステムを構築している．しかし，既存の手法は日本語のみを対象としており，英語など他の言語に適用するためには，様々な言語資源を用意し，さらに，それらを組み合わせるためのチューニングが必要であるが，それには多大な時間と労力を要するという問題がある．また，検索結果の文献数が膨大な場合でも，上述のプロトタイプシステムでは，結果を一覧表示する機能しか現状では備えていない．

２．研究の目的

本研究では，以下の 2 つの課題に取り組む．

- [課題 1] 複数言語を対象にした論文用語の特許用語への自動変換
  日本語以外の言語にも容易に適用可能な論文用語の特許用語への自動変換手法を提案し，実験により，その有効性を確認する．
- [課題 2] 分かりやすい検索結果の表示
  情報抽出技術と言い換え技術を統合し，技術内容や技術の効果に応じて検索結果を分類して提示するインタフェイスを構築する．

課題 1 については，例えば，ユーザが“floppy disk”という用語を入力すると“magnetic recording device”（磁気記録媒体）や“removable recording media”（リムーバブル記録媒体)といった英語の特許用語を出力する手法を開発する．この手法を，申請者がこれまでに構築してきた特許，論文横断検索プロトタイプシステム上で利用できるよう，システムを拡張する．

課題 2 については，あらかじめ各文書から「処理速度が速い」や「速度が向上」といった技術の効果に関する表現，「HMM」や「CRF」といった要素技術(手段)に関する表現を抽出し，課題 1 の手法を用いて検索した結果を，要素技術や技術の効果ごとに分類して検索結果を提示する．

３．研究の方法

［課題 1］論文用語の特許用語への自動変換

ここでは，同義語抽出手法を用いた論文用語の特許用語への自動変換手法を 2 種類提案する．

(1) 統計翻訳技術を用いた用語の変換

「磁気記録媒体」の英訳が“magnetic recording medium”，「磁気記憶媒体」の英訳も“magnetic recording medium”であるとき，英訳が共通である「磁気記録媒体」と「磁気記憶媒体」は同義語であると考えられる．この考え方に基づき，統計的機械翻訳技術を用いて自動的に獲得された翻訳モデルから，同義語を抽出する研究が近年行われるようになってきている．ここで，論文用翻訳モデ

ルにおいて“high resolution”と「高分解能」が，特許用翻訳モデルにおいて“high resolution”と“高解像度”とがそれぞれ対応付けられている場合，論文用語「高分解能」を特許用語「高解像度」に変換できるはずである，というのが，基本的な考え方である．

(2) 分布類似度を用いた用語の変換

文書中から自動的に同義語を抽出するこの他の手法として，分布類似度を用いた手法がある．この手法は，文書集合中で，「ある用語がどの語と何回係り受け関係にあるか」(以後，共起語ベクトル)により，その用語の意味を表現し，係り受け関係にある語の一致度に応じて，用語と用語の意味的な類似度を数値化する手法である．この考え方に基づき，あらかじめ，共起語ベクトルを，論文データベース，特許データベースから，それぞれ作成しておくことにより，論文用語を特許用語に変換することができる．

日本語と英語を対象に，上記の 2 手法を実装し，国立情報学研究所が主催する第 7 回および第 8 回 NTCIR ワークショップ特許マイニングタスクのデータを用い，評価を行う．

[課題 2] 分かりやすい検索結果の表示

(3) 特許と論文からの要素技術と効果の抽出

要素技術とその効果を示す表現を，特許や論文から自動的に抽出することを目的とする．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 1 に示すように，要素技術と効果を示す箇所に，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 種類のタグが付与される．技術の効果に関する表現は多様であり，そのすべてを処理対象とするのは，現在の言語処理技術では非常に困難である．このため，例えば，「処理速度(Attribute)が向上(Value)」や「ノイズ(Attribute)が減少(Value)」のように，技術の効果が「属性(Attribute)」と「属性値(Value)」の対で表現できるもののみを対象とする．近年の自然言語処理分野では，テキスト中に出現する属性と属性値の対の抽出が活発に研究されており，技術の蓄積が急速に進みつつある．特許や論文中の属性と属性値の対で表現可能な技術の効果に関する表現の抽出も，このような既存の技術の利用が期待できる．

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 1　特許と論文からの要素技術と効果の抽出例

(4) 特許と論文を対象にした検索システムの構築

(1)～(3)の技術を用い，特許と論文を対象にした検索システムを構築する．

４．研究成果

[課題 1] 論文用語の特許用語への自動変換

2 種類の同義語抽出手法を利用した論文用語の特許用語への自動変換手法：(1)統計翻訳技術を用いた手法と(2)分布類似度を用いた手法を提案した．提案手法の有効性を確認するため，NTCIR-7 特許マイニングタスクのデータを用い，学術論文を国際特許分類(IPC)に自動分類する実験を行った．実験の結果，統計翻訳技術を用いた変換手法は IPC のサブグループレベルでベースライン手法の MAP 値を 0.0020，分布類似度を用いた手法は IPC をサブクラスレベルでベースライン手法を 0.0024 向上できることが確認された．

[課題 2] 分かりやすい検索結果の表示

・学術論文分類システムの構築

あるキーワードに関する検索結果に複数のトピックの文献が混在する場合，それらはあらかじめトピックごとに自動分類した方が，結果の視認性が高い．これには文書分類技術が必要となるが，より高い精度で文書分類を行うには，同義語の認識が必要不可欠である．このため，日英上位・下位シソーラスを構築し，さらにそこから日本語および英語の同義語を抽出する手法を提案した．

さらに，国立情報学研究所が提供する学術論文データベース CiNii (http://ci.nii.ac.jp)を対象に，あるキーワードに関する検索結果の文献をトピックごとに自動分類し，さらに各文献の概要から要素技術と効果を自動抽出・ハイライト表示するシステム CiNii Mining (http://www.ls.info.hiroshima-cu.ac.jp/cgi-bin/cinii/index.cgi)を構築した．

・特許と学術論文を対象にした検索システムの構築

技術動向マップを構築するため，各特許と論文から抽出された要素技術と効果を，言い換え技術を用いて統合した．要素技術の言い換えでは，例えば「サポートベクトルマシン」という用語に対し，表記の揺れ(「サポートベクターマシン」)，訳語対(「Support Vector Machine」)，略語(「SVM」)の 3 種類を扱った．この他，特許から自動抽出した用語の上位・下位関係辞書を用い，例えば，「HMM 」と「SVM」を「機械学習 」としてまとめる手法についても検討した．他方，効果の言い換えでは，例えば，「ノイズの低減」と「ノイズの減少」といった表現の自動同定を行う手法を開発した．以上述べた技術と，(1)論文用語の 特許用語への自動変換技術，(2)分かりやすい検索結果の表示技術，(3)上述の言い換え技術を用いた要素技術と効果に関する表現の統 合技術を用い，すでに構築済のプロトタイプシステムを拡張することにより，特許と論文を対象にした検索システムを構築した．シ ステムの構築には，国立情報学研究所から提供を受けた論文データベース CiNii(約 370 万エントリ)および特許データベース(日本国公開公報および米国特許 1993 年-2007 年)を利用した．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

1. 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．

〔学会発表〕（計 7 件）

1. Nanba, H. (2013) “Towards Automatic Generation of Survey Articles ” Workshop on Content-based Approach to Scientific Information Systems.
2. Nanba, H., Takezawa, T., Uchiyama, K., and Aizawa, A. (2012) “Automatic Translation of Scholarly Terms into Patent Terms Using Synonym Extraction Techniques”. In Proceedings of the 8th International Conference on Language Resources and Evaluation
3. Fukuda, S., Nanba, H., and Takezawa, T. (2012) “Extraction and Visualization of Technical Trend Information from Research Papers and Patents”. In Proceedings of the 1st International Workshop on Mining Scientific Publications, collocated with JCDL 2012.
4. 福田悟志，難波英嗣，竹澤寿幸，武田英明，相澤彰子，大向一輝，宮尾祐介，内山清子．(2012) “CiNiiデータベースを用いた研究動向分析システムの構築” 言語処理学会 第 18 回年次大会，539-542.
5. Nanba, H., Mayumi, S., and Takezawa, T. (2011) “Automatic Construction of a Bilingual Thesaurus using Citation Analysis”.

In Proceedings of the 4th International CIKM Workshop on Patent Information Retrieval (PaIR'11), 25-30.
6. 間弓沙織，難波英嗣，竹澤寿幸．(2011) “日英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 17 回年次大会，892-895.
7. Nanba, H., Kondo, T., and Takezawa, T. (2010) “Automatic Creation of a Technical Trend Map from Research Papers and Patents”. In Proceedings of the 3rd International CIKM Workshop on Patent Information Retrieval (PaIR'10), 11-15.

〔図書〕（計 1 件）
1. 奥村学監修, 藤井敦, 谷川英和, 岩山真, 難波英嗣，山本幹雄，内山将夫著『特許情報処理:言語処理的アプローチ(自然言語処理シリーズ)』，コロナ社，2012.

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
CiNii Mining
http://www.ls.info.hiroshima-cu.ac.jp/cgi-bin/cinii/index.cgi

６．研究組織
(1)研究代表者
難波　英嗣　（NANBA HIDETSUGU）
広島市立大学・大学院情報科学研究科・准教授
研究者番号：50345378

(2)研究分担者
(3)連携研究者