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

難波英嗣
広島市立大学

竹澤寿幸
広島市立大学

内山清子
国立情報学研究所

相澤彰子
国立情報学研究所

1. はじめに

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。そこで、本研究では、同義語抽出手法を利用し、論文用語を特許用語に自動的に変換する手法を提案する。また、文書分類実験により、その有効性を確認する。

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、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。最後に 5 節で本論文をまとめる。

2. 関連研究

ジャンル横断検索や文書分類に関しては、これまでにいくつかの先行研究がある。NTCIR-3 で実施された技術動向調査タスク[Iwayama 2002]では、与えられた新聞記事と関連する特許を検索する、という課題が設定された。このタスクにおいて、Itoh ら[Itoh 2002]は、“Term Distillation”という手法を提案している。例えば,「社長」という単語は新聞記事中では高頻度で出現するが、特許中では出現頻度が非常に低い。このため、「一般的な用語ほど重要ではない」という考えに基づいて単語の重要度を計算する tf*idf などの手法を用いると、同じ単語でも新聞記事と特許では重要度が大きく異なる。そこで、Itoh らは、単語の新聞記事集合中での出現頻度と特許中での出現頻度の違いを考慮して単語の重み付けを行うことで、ジャンルを横断した文献の対応付けを行っている。しかし、この方法は、たとえば「磁気記録媒体」のように特許中では一般的に使われるが論文中ではまったく使われない用語に対しては適用できない。

この問題に対し、難波ら[難波 2009:a]は、論文用語を特許用語に自動変換する手法を提案している。例えば、論文用語「フロッピーディスク」を特許用語「磁気記録媒体」に自動変換される、しかし、この手法は、論文用語をユーザが 1 語ずつ入力することを前提としており、どの語を変換するかは、ユーザの判断に委ねられている。一般に、論文中には特許用語に変換する必要のない語も含まれている。詳細は後述するが、本タスクのように、学術論文を入力とし、国際特許分類(IPC)に自動分類する場合には、論文中のどの語を変換し、どの語を変換しないかを別途決定する必要が出てくる。また、この手法は、日本語以外の言語への拡張が容易でない、という問題もある。

論文用語の特許用語への自動変換に関連するこの他の研究に難波らのものがある [難波 2009:b]。この研究では、日本語で記載された特許データを訓練用データとして用いて英語論文を IPC に自動分類することを目的とし、特許用および論文用の 2 種類の機械翻訳システムを用いた分類手法を提案している。一般に、特許と論文では使われる用語が違うことから、入力された論文を翻訳する際、特許用の翻訳システムは、論文用のものと同等の翻訳精度が期待できない。しかし、特許用システムによる翻訳結果に特許用語が数多く含まれていれば、文書分類の段階での精度向上が期待できるため、総合的に見れば特許用翻訳システムを用いるメリットがあると考えられる。そこで、難波らは、入力された英語論文を、2 つの翻訳システムを用いて和訳し、その結果を統合して索引を作成することにより、1 つの翻訳システムを用いる場合と比べ、学術論文の分類精度が向上することを実証している。この研究は、理想的な翻訳を介した論文用語の特許用語への変換に関するひとつのアプローチととらえることができるが、本研究では、このような理想的な翻訳を利用しない点で、難波らの研究と比べ、より難しい課題と考えられる。

特許と論文を対象にした情報アクセスに関するこの他の研究として、NTCIR-7 および NTCIR-8 特許マイニングタスク[Nanba 2008, Nanba 2010]が挙げられる。これは、特許と論文を対象にした検索や動向分析など、様々な目的に利用可能な言語処理技術の開発を最終目標とした研究プロジェクトであり、NTCIR-7 および NTCIR-8 では、学術論文を IPC に自動分類するタスクを設定している。上述の難波らの研究[難波 2009:b]は、NTCIR-7 のデータセットを用いたものであり、本研究でも、このデータを実験に用い、提案手法の有効性を検証する。

特許と論文を対象としたこの他の研究として、TREC Chemistry Track[1]が挙げられる。これは、評価ワークショップ TREC において 2009 年より新しく始まったタスクのひとつであり、化学分野の論文と特許に特化したジャンル横断検索を目的としている。

3. 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

本節では、同義語抽出手法を利用した論文用語の特許用語への変換手法を 2 種類提案する。

3.1 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し、情報検索の際の query expansion に用いたり[海野 2008]、機械翻訳[Kauchak 2006]や自動要約の評価[Zhou 2006, 平原 2009]に利用したり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。

3.2 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[相澤 2008]。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(係り受け関係にある語およびその頻度のリスト)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。例えば、以下の例のように、論文データベース中の「フロッピーディスク」の共起語ベクトルと特許データベース中の「磁気記録媒体」の共起ベクトルの一致度が高い場合、論文用語「フロッピーディスク」は特許用語「磁気記録媒体」に変換される。

フロッピーディスク (論文用語)	磁気記録媒体 (特許用語)
3 に_書き込む ←→	**4 に_書き込む**
2 に_収める	2 を_作る
2 に_取り込む	**2 は_読み取る**
1 は_読み取る	1 を_傷つける

4. 実験

提案手法の有効性を調べるため、実験を行った。

4.1 実験方法

タスク

本研究では、NTCIR-7[Nanba 2008]特許マイニングタスクにおける学術論文分類サブタスクのデータを用いて実験を行った。このタスクでは、入力された学術論文(論文表題と概要)を IPC に自動的に分類することを目的としている。IPC は、特許文献の技術内容によって上から順に「セクション」、「クラス」、「サブクラス」、「メイングループ」、「サブグループ」の 5 階層から構成・分類されており、国際特許分類第 6 版ではサブグループのレベルで約 50,000[2]の IPC コードが存在する。本課題では、「サブクラス」、「メイングループ」、「サブグループ」レベルの IPC コードを論文抄録に付与することを目的とする。

評価用データおよび評価尺度

評価用データには、879 論文に人手で「サブクラス」、「メイングループ」、「サブグループ」レベルで IPC コードを付与したものを、評価尺度は、MAP (Mean Average Precision)を、それぞれ用いた。なお、評価用データには、1 論文あたり平均 2.2 個の IPC コードが付与されている。

学術論文分類システム

IPC のコード数はサブグループレベルで 30,855

[1] https://wiki.ir-facility.org/index.php/TREC_Chemistry_Track

[2] NTCIR-7 特許マイニングタスクでは、これらのうち、学術分野とは関連性の低い分野を除外した 30,885 の IPC コード(サブグループレベル)を対象としている。

個と非常に多く、さらに、訓練用データ(1993～2002 年の公開公報。各公報にはひとつ以上の IPC が人手で付与されており、これを分類の際の訓練用データとして用いる)も約 350 万件と膨大である。このため、本研究では、自然言語処理分野における分類問題では一般的な機械学習は莫大な計算コストがかかると判断し、代わりに k-Nearest Neighbor（k-NN）法を採用した。

以下、k-NN 法に基づいた学術論文の分類システムについて述べる。まず、学術論文から内容語(名詞句, 動詞, 形容詞)を抽出して索引を作成する。次に、この索引に対応する IPC コードを、以下の手順で自動的に付与する。

1. 入力クエリ（索引）に対して特許検索システムを用いて検索し、上位 170 件の結果を得る。
2. 手順 1 で得られた各特許に付与された IPC コードを獲得する。
3. 以下の式に基づいて IPC コードをランク付けし、出力する。

$$\text{Score}(X) = \text{検索された各特許の、入力クエリとの類似度}$$

ここで、X は IPC コード、n は検索結果上位 170 件の中で X が付与されている特許数を示す。なお、170 という値は、NTCIR-7 特許マイニングタスクドライランデータを用いて決定した。

提案手法による索引の変換

上記手順 1 の入力クエリ(索引)を、3.1 節および 3.2 節で述べた手法で特許用語に変換する。ここで、「研究」や「手法」など、論文中で一般的に使われる(分類に寄与しない)用語を変換の対象にすると、かえって分類精度を低下させる可能性がある。そこで、各索引語の IDF 値が閾値以上のもののみを変換の対象とした場合(以後、IDF)と、すべての索引語を変換の対象とした場合の 2 種類で実験を行う。

翻訳モデル

特許用翻訳モデル構築用のデータとして NTCIR-7 特許翻訳タスク[Fujii 2008]で配布されている約 1,800,000 対の日英対訳文を、論文用翻訳モデル構築用には NII-ELS 論文データベースから抽出した 1,763,217 対の日英論文表題(以後、TITLE モデル)を、それぞれ用いた。この他、NII-ELS の概要から抽出した 600,000 対の日英対訳文からも論文用翻訳モデル(以後、ABST モデル)を構築した。また、翻訳モデル構築のためのツールとして、GIZA を用いた。

分布類似度

特許用共起語ベクトルの構築には、1993～2002 年の公開公報約 6 億文、論文用には、NII-ELS の日本語論文概要約 60 万文を用いた。また、構文解析には CaboCha を用いた。

比較手法

以下に示す 6 種類の提案手法および 1 種類のベースライン手法を用いて、比較実験を行った。

提案手法

- **SMT_ABST**: PAPER(ベースライン)手法を ABST モデルで拡張
- **SMT_ABST+IDF**: SMT_ABST 手法において、IDF 値の低い論文用語は特許用語に変換しない
- **SMT_TITLE**: PAPER 手法を TITLE モデルで拡張
- **SMT_TITLE+IDF**: SMT_TITLE 手法において、IDF 値の低い論文用語は特許用語に変換しない
- **DS**:分布類似度を用いた用語変換
- **DS+IDF**: DS 手法において、IDF 値の低い論文用語は特許用語に変換しない

ベースライン手法

- PAPER (ベースライン): 論文表題+概要中の用語をそのまま利用

4.2 結果

実験結果を表 1 に示す。表から分かるとおり、統計翻訳技術を用いた用語変換手法は、サブグループおよびメイングループレベルでの分類で、分布類似度を用いた手法は、サブクラスレベルでの分類で、ベースライン手法を上回っている。

表 1 「サブクラス」、「メイングループ」、「サブグループ」 レベルにおける提案手法およびベースライン手法の MAP 値

手法	サブグループ	メイングループ	サブクラス
SMT_ABST	0.3786	**0.5186**	0.6691
SMT_ABST+IDF	**0.3812**	**0.5197**	0.6709
SMT_TITLE	**0.3797**	**0.5208**	0.6688
SMT_TITLE+IDF	**0.3799**	**0.5204**	0.6710
DS	**0.3793**	0.5182	0.6717
DS+IDF	**0.3794**	0.5175	**0.6744**
PAPER (baseline)	0.3792	0.5185	0.6720

4.3 考察

IDFの有効性について

6 種類の提案手法のうち、IDF を用いたものと

用いないものを比較すると、メイングループレベルで一部若干の逆転があるものの、概ね、IDF を用いた手法が良好な結果が得られている。このことから、一般性の高い論文用語は特許用語に変換しない方が良いと言える。

2種類の変換手法について

索引の特許用語への変換結果を確認したところ、統計翻訳技術を用いた変換手法では、同義語に変換されるのに対し、分布類似度を用いた手法では、索引語と共通の性質を持つ関連語や類義語などに変換される傾向にあることが分かった。前者は、より狭い範囲の関連特許の収集に、後者は前者よりも広い範囲の特許の収集に向いていると言える。実際、4.2 節で述べたとおり、統計翻訳技術を用いた用語変換手法は、サブグループおよびメイングループレベルでの分類で、分布類似度を用いた手法は、サブクラスレベルでの分類で、ベースライン手法の分類精度を向上させるのに有用であることは、すでに 4.2 節で述べたとおりである。

5. おわりに

本研究では、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を提案した。また、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 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(SMT_ABST+IDF)は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(DS+IDF)は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。

参考文献

[相澤 2008] 相澤彰子: 大規模テキストコーパスを用いた語の類似度計算に関する考察, 情報処理学会論文誌, Vol.49, No.3, pp.1426-1436 (2008).

[Fujii 2008] Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, Proc. the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 389-400 (2008).

[平原 2009] 平原一帆, 難波英嗣, 竹澤寿幸, 奥村学: 言い換えを用いたテキストの自動評価, 情報処理学会 自然言語処理・音声言語情報処理合同研究会, NL-191 / SLP-76, (2009).

[Itoh 2002] Itoh, H., Mano, H., and Ogawa, Y.: Term Distillation for Cross-DB Retrieval, Proc. Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task (2002).

[Iwayama 2002] Iwayama, M., Fujii, A., Kando, and N., Takano, A.: Overview of Patent Retrieval Task at NTCIR-3, Proc. Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task (2002).

[Kauchak 2006] Kauchak, D. and Barzilay, R: Paraphrasing for Automatic Evaluation, Proc. 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pp.455-462 (2006).

[Nanba 2008] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-7 Workshop, Proc. the 7th NTCIR Workshop Meeting, pp.325-332 (2008).

[難波 2009:a] 難波英嗣, 釜屋英昭, 竹澤寿幸, 奥村学, 谷川英和, 新森昭宏: 特許用語の論文用語への自動変換, 情報処理学会論文誌データベース, Vol.2, No.1, pp.81-92 (2009).

[難波 2009:b] 難波英嗣, 竹澤寿幸: 2 種類の翻訳システムを用いた学術論文の特許分類体系への自動分類, 情報処理学会論文誌データベース, Vol.2, No.3, pp.76-86 (2009).

[Nanba 2010] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-8 Workshop". Proc. the 8th NTCIR Workshop Meeting (2010) (to appear)

[海野 2008] 海野裕也, 宮尾祐介, 辻井潤一: 自動獲得された言い換え表現を使った情報検索, 言語処理学会第 14 回年次大会, pp.123-126 (2008).

[Zhou 2006] Zhou, L., Lin, C.-Y., Munteanu,D.S., and Hovy, E: ParaEval: Using Paraphrases to Evaluate Summaries Automatically. Proc. 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pp.447-454 (2006).