学術論文の国際特許分類への自動分類

評価ワークショップNTCIR-7特許マイニングタスク

広島市立大学大学院情報科学研究科講師
難波　英嗣

PROFILE 1996年東京理科大学理工学部電気工学科卒業。2001年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。同年、日本学術振興会特別研究員。2002年東京工業大学精密工学研究所助手。同年、広島市立大学情報科学部講師。2007年4月広島市立大学大学院情報科学研究科講師、現在に至る。博士（情報科学）。言語処理学会、情報処理学会、人工知能学会、ACL、ACM各会員。

nanba@hiroshima-cu.ac.jp　082-830-1584

筑波大学大学院図書館情報メディア研究科准教授
藤井　敦

PROFILE 1998年東京工業大学大学院博士課程修了。博士（工学）。現在、筑波大学大学院准教授。2003年IPAから「天才プログラマー／スーパークリエータ」を受賞。自然言語処理、情報検索、音声言語処理の研究に従事。

株式会社日立製作所　中央研究所
東京工業大学精密工学研究所
岩山　真

PROFILE 1992年（株）日立製作所入社。以来、文書検索、文書分類、自然言語処理等の研究に従事。また、NTCIRにおいて特許検索用テストコレクションの作成に携わる。

東京工業大学統合研究院特任准教授
橋本　泰一

PROFILE 1997年東京工業大学工学部情報工学科卒業。2002年東京工業大学大学院情報理工学研究科博士後期課程修了。2002年4月東京工業大学大学院情報理工学研究科助手。2006年4月東京工業大学統合研究院特任准教授、現在に至る。博士（工学）。言語処理学会、情報処理学会、人工知能学会、科学技術社会論学会各会員。

1 はじめに

本稿では、国立情報学研究所（NII）が主催する評価ワークショップ「NTCIR」において、筆者らが行っている特許と論文を対象とした情報処理のためのテストコレクション（評価用ベンチマーク）の構築研究について述べる。評価ワークショップとは、複数の研究グループが協調と競争を通して、問題設定やテストコレクション、評価方法について共同開発していく枠組みである。

筆者らは本ワークショップにおいて「特許マイニングタスク」を企画し、国内外から参加グループを募り、2007年から研究を開始している。近年、大学研究者自身が関連論文だけでなく関連特許について情報を検索したり、特許を出願したりする機会が増えており、2008年6月に政府の知的財産戦略本部が発表した「知的財産権推進計画2008」においても、推進計画2006、2007に引き続き、大学研究における特許情報の重要性が謳われている。この計画で、大学研究者の利用を想定した特許・論文情報統合検索システムの整備が含まれていることからも、このような傾向が今後さらに強まっていくと思われる。

特許と論文を検索するのは、大学研究者に限った話ではない。例えば、特許庁の審査官は、出願された技術が特許権の取得に該当するかどうか判断するために、過去に同様の特許が出願されたり論文が発表されたりしてい

ないか調査する。これは、一般に無効資料調査と呼ばれている。同様の調査は、サーチャーと呼ばれる専門の担当者が審査官による審査を経た出願技術を再調査し、競合する他者の権利を無効化するために民間企業の社内で行われることもある。

こうした状況を鑑み、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最終目標とし、そのための第一歩として筆者らが位置づけているのが、本評価ワークショップの特許マイニングタスクである。

本タスクでは、日本語または英語論文抄録に、特許分類体系のひとつである「国際特許分類」（International Patent Classification：IPC）を自動的に付与することを目的とする。特許を分類するタスクは、これまでにNTCIR-5と6におけるFターム分類タスクが実施されてきたが、今回のタスクでは、分類対象となる文書が論文に変わるため、特許と論文で使われる用語の違いについて新たに検討する必要がある。

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。また、特許では学術用語よりも多様な表現が用いられることが多い。例えば、「機械翻訳」という学術用語に対する特許用語は「機械翻訳」の他にも「自動翻訳」「言語変換」などがある。このため、単純に表層的な単語の一致度を用いる従来の分類モデルでは、十分な分類精度が得られるとは限らない。本タスクでは、このような論文と特許の用語の使われ方の違いを吸収できる分類や検索のための基礎技術の確立を目指している。

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等の手法を用いると、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。

特許と論文を横断的に検索するための研究としてNanbaら［Nanba 2008］の研究が挙げられる。近年、特許中で関連論文を、また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Nanbaらは特許中で関連文献が引用される「従来の技術」という項目を解析して引用論文の書誌情報を抽出し、特許と論文間の引用関係を解析している。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十分とは言えない。

特許と論文を横断的に検索するための別のアプローチとして、釜屋ら［釜屋 2008］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」に自動変換する。釜屋らは、論文用語の特許用語への変換を実現するため、特許と論文間の引用関係に着目している。一般に、引用関係にある特許と論文は、同一トピック（分野）である可能性が高い。そこで、ある用語を表題に含んだ論文を収集し、それらと直接引用関係にある特許から、特許のトピックを示す用語を抽出すれば、入力された論文用語に関連する特許用語の変換が実現できる。

以上は、文書のジャンルを横断した情報アクセス技術の一例であるが、特許マイニングタスクでは、参加者による新しい文書分類技術やジャンルの横断も意識した情報アクセス技術の確立が期待されている。

3 特許マイニングタスクの概要およびテストコレクションの構築

3.1 タスクの概要

前述のとおり、特許マイニングタスクでは日本語また

は英語論文抄録に、特許分類体系のひとつであるIPCのコードを自動的に付与する。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のクラスが存在する。本タスクでは、最下層の「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。【図1】は日本語論文の例である。ここで、<TOPIC-ID>は論文のIDを、<TITLE>と<ABSTRACT>は論文表題と概要を、それぞれ示している。タスクの参加者は、図1のような入力が与えられると、対応するIPCコードを自動的に出力するシステムを構築することが求められる。

特許マイニングタスクでは、以下のサブタスクが実施された。

●日本語サブタスク（Japanese）：日本語の論文を日本語で記載された特許データを用いて分類する。

●英語サブタスク（English）：英語の論文を英語で記載された特許データを用いて分類する。

●言語横断サブタスク（J2E）：日本語の論文を英語で記載された特許データを用いて分類する。

この他、英語の論文を日本語で記載された特許データを用いて分類するサブタスクも参加募集をしたが、参加者がいなかった。以上のサブタスクは【図2】にまとめられる。

<TOPIC>
<TOPIC-ID>312</TOPIC-ID>
<TITLE>二値画像用高速符号化/復号LSI</TITLE>
<ABSTRACT>二値画像データを高速で符号化，復号するLSIを開発した。参照ラインデータ上に「基準色変化点」を探すのと並行して、それを参照するランのイメージデータを生成する方式により、復号性能を向上させた。また、符号化時と復号時共に同じ方向にデータが流れるパイプライン構成とし、さらに主な回路は共通化する構成によって回路を簡略化した。</ABSTRACT>
</TOPIC>

図1　システムの入力例

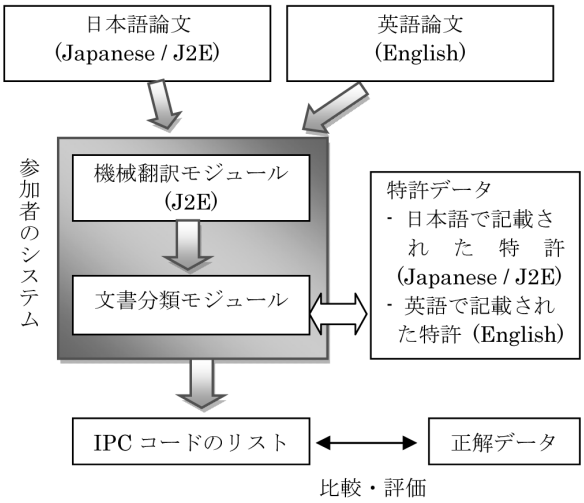

図2　サブタスクのまとめ

3.2　特許データ

各サブタスクで使われたデータを【表1】にまとめる。

3.3　正解データの作成

正解データの作成方法について説明する。論文抄録をIPCに自動分類するタスクを実施するには、正解データとして、論文抄録にIPCコードを付与したものを準備する必要がある。ここで、前述のとおり国際特許分類第6版ではサブグループ数が約50,000あり、論文抄録に人手でサブグループを付与する作業は、知識と経験を持った専門家でも非常に時間がかかる。特許以下の手順で正解データの作成を行う。原則的に、ある発明を論文等で発表すると、その時点で発明の内容は公知になるとみな

表1　文書セット

データ名	年	サイズ	文書数	言語
(1) 日本国公開特許公報	1993-2002	100GB	3.5M	日
(2) 米国特許	1993-2000	33GB	0.99M	英
(3) Patent Abstracts of Japan（特許抄録の英訳）	1993-2002	4.2GB	3.5M	英
(4) NTCIR-1、NTCIR-2言語横断タスクテストコレクション（論文抄録データ）[Kando 1999] [Kando 2001]	1988-1999	1.4GB	0.26M	日/英

され、特許を受ける権利が失われるが、日本では、特許庁の承認を受けた学協会、団体での発表、出版であれば、発表後6ヶ月の間に限り、特許出願を行うことができることが、特許法第30条で規定されている。特許法30条が適用される特許には、特記事項として、論文を発表した論文誌名、発表した日付などが特許中に記載されるため、これを手がかりに効率的に正解データを作成することができる。例えば、以下のように記載される。

> 【新規性喪失の例外の表示】特許法第30条第1項適用申請有り2000年3月14日　社団法人情報処理学会発行の「第60回（平成12年前期）全国大会講演論文集（4）」に発表

各特許に付与されているIPCコードを、「新規性喪失の例外の表示」の欄に記載されている論文のIPCコードと見なすことで、正解データを効率的に作成する。

特許マイニングタスクで用いる公開公報（1993年～2002年）約350万件には、「新規性喪失の例外の表示」を含んだものが約9,000件存在する。これらから、まず、発表日、論文誌名を抽出する。この欄には、論文表題や発表者名が記載されることはほとんどないが、多くの場合、特許の発明者と論文の発表者が同一（連名の場合、共通する人名がある）であるため、特許の発明者を論文の著者として抽出しておく。次に、これらの情報を抄録データベース※1の書誌情報と照合し、対応付けを行う。ここで、ひとつの「新規性喪失の例外の表示」に記載の論文に対し、平均6件程度まで対応付けの候補となる抄録を絞り込むことができる。最終的に、これらの候補を人間が確認し該当する抄録を特定する。抄録にIPCを付与するのは専門的な知識が必要であるが、「新規性喪失の例外の表示」記載の論文と、候補として挙げられた論文が同一であるかどうかは、専門知識を持たないものでも判断できるため、学生アルバイトにこの判定作業を依頼し、正解データ作成の作業を行った。その結果、976対の論文と特許の対応データが得られ、1論文あたり平均2.3個のIPCが正解として付与されたデータが作成された。

※1　本タスクでは，NTCIR-1，2の言語横断検索タスクで用いた約46万件の抄録データベースを用いる
※2　http://ntcir.nii.ac.jp

4 おわりに

本稿執筆当時はNTCIR-7が進行中であり、本タスク参加グループの技術の詳細は、2008年12月に国立情報学研究所で開催される最終報告会で発表される予定である。また、報告書もNTCIRのウェブページ※2で公開されるので、詳細はそちらを参照されたい。

参考文献

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

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

[Kando 1999] Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., and Hidaka, S. “Overview of IR Tasks at the First NTCIR Workshop”, In Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.11-44（1999）

[Kando 2001] Kando, N., Kuriyama, K., and Yoshioka, M. “Overview of Japanese and English Information Retrieval Tasks（JEIR）at the Second NTCIR Workshop”, In Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, pp.4-37 - 4-60（2001）

[Nanba 2008] Nanba, H., Anzen, N., and Okumura, M. “Automatic Extraction of Citation Information in Japanese Patent Applications”, Special Issue of International Journal on Digital Libraries on Very Large Digital Libraries（to appear）

[釜屋 2008] 釜屋英昭，難波英嗣，竹澤寿幸，奥村学．“特許用語の論文用語への自動変換”，言語処理学会第14回年次大会，pp.801-804（2008）