Japio

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

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 TEL 082-830-1584

広島市立大学大学院情報科学研究科教授
竹澤　寿幸

PROFILE
1984年早稲田大学理工学部電気工学科卒業。1989年早稲田大学大学院博士後期課程修了。同年（株）国際電気通信基礎技術研究所入社。2007年広島市立大学大学院情報科学研究科教授、現在に至る。工学博士。音声対話翻訳の研究開発に従事。平成18年度電子情報通信学会ISS論文賞受賞。電子情報通信学会、人工知能学会、日本音響学会、言語処理学会各会員。
TEL

1 はじめに

学術論文の特許分類体系への分類は、特許と論文を対象とした網羅的かつ効率的な先行技術調査、無効資料調査、技術動向分析などを可能にする。しかし、特許の場合と同様に論文発表時に著者本人に特許分類コードを付与してもらうことや、すでに発表済みのすべての論文に人手で分類コードを付与することは、コスト面から考えて現実的ではない。そこで、本研究では、学術論文を特許分類体系に自動的に分類する手法を提案する。

これまでに、特許を自動的に分類する研究は、国立情報学研究所が主催の評価ワークショップNTCIR-5[Iwayama 2005]と6[Iwayama 2007]において、Fターム分類タスクとして実施されてきたが、学術論文を特許分類体系に分類する場合には、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」という論文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。さらに、より網羅的な調査や分析を可能にするためには、複数の言語で記述された論文を分類対象にする必要がある。

これらの問題を解決するため、本研究では、特許用および論文用の2種類の翻訳システムを用いた分類手法を提案する。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。本研究では、第7回NTCIRワークショップ（NTCIR-7）特許マイニングタスク[Nanba 2008:a]において実施された言語横断サブタスク（E2J）のデータを用い、提案手法の有効性を検証する。

本論文の構成は以下のとおりである。次章では、NTCIR-7特許マイニングタスクについて述べる．3章では、2種類の翻訳システムを用いた論文の分類手法を提

案する。４章では、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章で本論文をまとめる。

2 NTCIR-7 特許マイニングタスク

NTCIR-7 特許マイニングタスク [Nanba 2008:a] は、特許と論文を対象にした情報アクセスに関する研究プロジェクトである。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プロジェクトであり、その第一歩として、NTCIR-7 では、学術論文を国際特許分類に自動分類するタスクを設定している。特許マイニングタスクでは、以下の４つのサブタスクが実施された。

- ●日本語サブタスク：日本語の論文を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分類する。
- ●英語サブタスク：英語の論文を英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訓練用データとして用いて分類する。
- ●言語横断サブタスク（E2J）：英語の論文を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分類する。

本研究では、特許マイニングタスクの中でも、NTCIR-7 で参加者がいなかった言語横断サブタスク（E2J）のデータを用い、提案手法の有効性を検証する。

3 学術論文の特許分類体系への自動分類

3.1 提案手法

ジャンル G1 に属する言語 L1 で記述された文書 I を、ジャンル G2 に属する言語 L2 で記述されたラベル付き文書集合を訓練用データとして用いて文書分類する手順を、【図 1】を用いて説明する。一般的には、（1）ジャンル G1 用の翻訳システムを用いて入力文書の記述言語を L1 から L2 に翻訳した後 (図中 O)、（2）ジャンルにより異なる用語の使われ方を考慮して適宜用語を変換した上で (図中 O')、(3) 文書分類を実施する、という 3 つのステップが必要となる．ここで、ジャンル G2 用の機械翻訳システムが存在する場合、この翻訳システムを用いて入力文書 I を翻訳すれば、出力結果にはジャンル G2 に適合した語彙が使われているため、上述のステップ（1）と（2）を同時に解決できる可能性がある。しかしながら、ジャンル G2 用の翻訳システムは、言語 L1 で記述されたジャンル G2 の文書が入力されることが前提となっているため、G1 用の翻訳システムを用いた場合と比べ、翻訳結果には、誤訳が含まれる可能性も高くなる．そこで、ジャンル G1 用と G2 用の翻訳システムによる結果を組み合わせる（O+O'）ことにより、ジャンル G2 の語彙を含みつつ、G2 用翻訳器の誤訳の影響を最小限にとどめたジャンル横断、言語横断文書分類の実現を目指す。

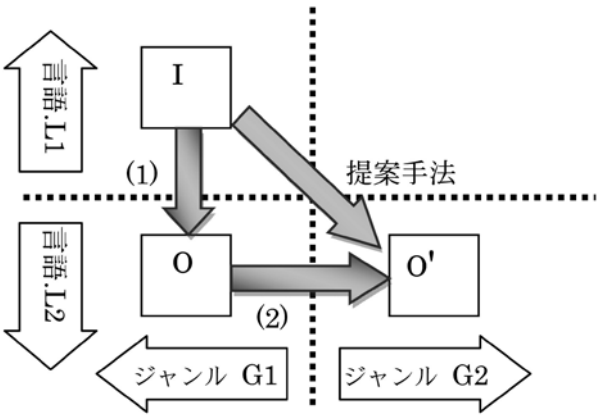

図 1　提案手法

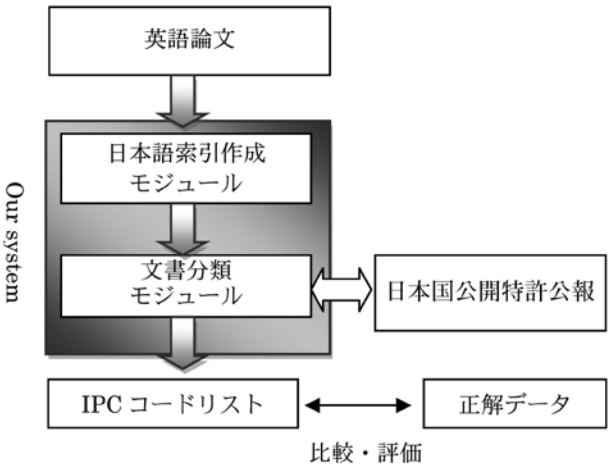

図２　システム概要

3.2　システムの構成

本研究では、「英語論文（言語 L1、ジャンル G1）を、日本語（言語 L2）で記述された特許データ（ジャンル G2）を訓練用データとして用いて特許分類体系に自動分類する」という課題に取り組む。【図 2】にシステムの構成を示す。提案システムは、「日本語索引作成モジュール」と「文書分類モジュール」から構成されている。以下に、各モジュールについて説明する。

日本語索引作成モジュール

日本語索引作成モジュールは、【図 3】に示すような英語表題と概要の対を入力とし、特許用および論文用の 2 種類の翻訳システムを用いて、【図 4】に示す日本語の索引を出力する [1]。日本語索引を作成するには、2 つの方法：（A）入力された英語表題と概要を日本語に翻訳した後、内容語 (名詞 / 名詞句、動詞、形容詞) 抽出して索引を作成する方法と、(B) 入力された英語表題と概要から内容語を抽出して英語索引を作成した後、各索引語を翻訳する方法が考えられる。今回は、（A）（B）2 種類の方法で実験を行った。なお、言語モデルの作成には SRILM を、フレーズテーブル (翻訳モデル) の作成には Giza を、デコーダには Moses を、利用した。言語モデルおよびフレーズテーブル獲得用の日英対訳データとして、NTCIR-7 特許翻訳タスク [Fujii 2008] で配布されている約 1、800、000 日英訳文対を、論文用翻訳システムの獲得には NTCIR-1、２言語横断検索タスクで配布されている論文抄録データから抽出した約 300、000 日英訳文対を、それぞれ全て用いた。また、索引語の翻訳には、Giza と Moses を用いて特許用翻訳システムと論文用翻訳システムを獲得する過程で得られるフレーズテーブルの中で、フレーズの英日翻訳確率が最も高いものを英語索引語の日本語訳として用いた。

文書分類モジュール

文書分類には、k-NN 法に基づく Nanba の分類器を用いる [Nanba 2008:b]。この分類器は、NTCIR-6 特許検索タスク [Fujii 2007] 用に開発された特許検索

英語表題：A Sandblast-Processed Color-PDP Phosphor Screen
英語概要：Barrier ribs in the color PDP have usually been fabricated by multiple screen printing. However, the precise rib printing of fine patterns for the high resolution display panel is difficult to make well in proportion as the panel size grow larger. On the other hand, luminance and luminous efficiency of reflective phosphor screen will be expected to increase when the phosphor is deposited on the inner wall of display cells. Sandblasting technique has been applied to make barrier ribs for the high resolution PDP and nonffat phosphor screens on the inner wall of display cells.

図３　英語表題と概要

1　なお、各用語の左に記載されている数値は用語の文書内頻度を示している。

18 形成
18 ＰＤＰ
18 型蛍光面
12 障壁形成
12 障壁
12 蛍光
12 カラーＰＤＰ
12 反射型蛍光
12 型蛍光
12 サンドブラ
（以下略）

図４　日本語索引

システム [Nanba 2007] を内部で利用している。

この検索システムは検索モデルとしてベクトル空間型モデルを、索引語には MeCab を用いて抽出された名詞または名詞句 (連続して出現する名詞)、動詞、形容詞を、単語の重みの計算には tf　(term frequency) を、類似度尺度には SMART を、それぞれ採用している。入力された日本語索引に対応する 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（日本語索引）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。

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。また、“Relevance score of each patent”は、検索された各特許と入力クエリとの類似度を示す。

国際特許分類への自動分類

本研究では、学術論文を分類する特許分類体系のひとつとして、国際特許分類（IPC）を用いる。IPC は、国際的に統一されて用いられている分類体系であり、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 IPC コード[2] が存在する。本研究では、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。

4 実験

提案手法の有効性を調べるため、実験を行った。本章では、4.1 節で実験方法について、4.2 節で比較手法について、それぞれ説明する。また、4.3 節で実験結果を報告し、4.4 節で結果を考察する。

4.1　実験方法

NTCIR-7 特許マイニングタスク言語横断サブタスク (E2J) のデータを用い、実験を行った。

正解データ

【図 3】に示すような英語論文抄録 879 件に、人手で IPC コードを付与したデータを用いた。1 抄録あたり平均 2.2 個の正解の IPC コードが付与されている。このデータを、システムが出力した IPC コードのリスト（抄録毎に最大 1000 件）と比較し、MAP（Mean Average Precision）により評価した。

文書データ

実験に用いた文書データを【表 1】にまとめる。

比較手法

以下に示す 3 種類の提案手法、7 種類の比較手法、

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 コードを対象としている。

データ名	年	サイズ	文書数	言語
日本国公開 特許公報	1993-2002 02	100 GB	3.5M	日
NTCIR-1, NTCIR-2 言語横断タスク データ (論文抄録)	1988-19 99	1.4GB	0.26M	日 / 英

表 1　文書データ

および２種類の手法 (システムの上限) で実験を行った。なお、“SMT(X)”と“Index(X)”は、それぞれ「翻訳システム X を使って全文を翻訳した後、日本語索引を作成」と「英語索引を作成した後、翻訳システム X 用のフレーズテーブルを使って日本語索引を作成」を意味する。

提案手法

- Index(Paper)*Index(Patent): 論文翻訳用および特許翻訳用システム構築の際に作成されたフレーズテーブルを使って英語索引をそれぞれ和訳し、その積集合を利用
- Index(Paper)+Index(Patent): 論文翻訳用および特許翻訳用のフレーズテーブルを使って英語索引をそれぞれ和訳し、その和集合を利用
- SMT(Paper)+Index(Patent): 論文用翻訳システムを使って英語論文を和訳した後、日本語索引を作成したものと Index(Patent) の和集合
- SMT(Paper)+SMT(Patent): 論文用翻訳システムと特許用翻訳システムを使って英語論文を和訳した後、それらの和集合を利用

比較手法

- SMT(Paper): 論文用翻訳システムを使って英語論文を和訳した後、日本語索引を作成
- SMT(Patent): 特許用翻訳システムを使って英語論文を和訳した後、日本語索引を作成
- Index(Paper): 論文翻訳用フレーズテーブルを使って英語索引を和訳
- Index(Patent): 特許翻訳用フレーズテーブルを使って英語索引を和訳

システムの上限

- 日本語サブタスク : 入力された英語論文の対訳データを理想的な翻訳と考え、日本語索引を作成

4.2　実験結果

【表 2】に実験結果を示す。論文用翻訳モジュールを利用した“SMT(Paper)”は理想的な翻訳 (“日本語サブタスク”) を用いた時の結果に非常に近い結果となった。提案手法および比較手法の中では、提案手法のひとつである“SMT(Paper)+Index(Patent)”が最も

	手法	MAP 値
提案手法	Index(Paper)*Index(Patent)	0.2230
	Index(Paper)+Index(Patent)	0.2596
	SMT(Paper)+Index(Patent)	0.2897
	SMT(Paper)+SMT(Patent)	0.2826
比較手法	SMT(Paper)	0.2777
	SMT(Patent)	0.2507
	Index(Paper)	0.2433
	Index(Patent)	0.2373
上限	日本語サブタスク (理想的な翻訳)	0.3267

表 2　実験結果

高い MAP 値を得た。

4.3 考察

【表 3】に、フレーズテーブルを用いた英語索引の翻訳結果を示す。表は、NTCIR-7 特許マイニングタスク言語横断（E2J）サブタスクのトピック番号 351（【図 3】）から生成された英語索引から、論文用および特許用翻訳システムのフレーズテーブルを用いて日本語に翻訳した結果の一部である。また、参考までに、【図 5】に、論文の著者自身が作成したトピック番号 351 の対訳データを示す。

この表において、たとえば“screen”という用語は論文用のフレーズテーブルを用いた場合には「スクリーン」、特許用のフレーズテーブルを用いた場合には「画面」と翻訳されている。一般的には「スクリーン」と「画面」は同義語であるが、統計的に見れば、論文では「スクリーン」、特許では「画面」と表記する可能性が高いと考えられる。実際、【図 5】における著者自身が作成した日本語論文内でも「スクリーン」という用語が使用されている。

提案手法の実用性

最後に、提案手法のひとつであり、上限以外の手法で最も高い MAP 値を得た“SMT(Paper)+Index(Patent)”が、どの程度実用に耐えうる物かを調べるため、上位 n 件の再現率を調べた。ここで、再現率 (Recall) は以下の式により定義される。

英語索引	論文用翻訳システムのフレーズテーブル	特許用翻訳システムのフレーズテーブル
screen	スクリーン	画面
display	ディスプレイ	表示
high resolution	高分解能	高解像度
inner wall	ひだの集中	内壁
inner	内部	インナー
resolution	分解能	解像度
barrier	障壁	バリア

表 3　2 種類のフレーズテーブルを用いた英語索引の翻訳結果

日本語表題：サンドブラスト法によるカラー PDP 蛍光面の試作
日本語概要：PDP の大型化、高精細化においてスクリーン印刷による障壁形成は、種々の問題点を有する。我々は、この障壁形成プロセスを検討した結果、サンドブラスト法による新規形成プロセスを見いだし、大型 PDP、高精細 PDP の障壁形成に有効であることを確認した。また、障壁壁面に反射型蛍光面を形成することにより輝度効率の向上が期待される。我々は、サンドブラスト法の検討を進めることにより、この手法が上記の反射型蛍光面の形成にも有効であることを確認し、8 インチ DC 型カラー PDP を試作し、従来の透過型蛍光面を有するカラー PDP と比較検討を行った。

図 5　日本語論文データの例

$$Recall = \frac{The\ number\ of\ correctly\ identified\ IPC\ codes}{The\ number\ of\ relevant\ IPC\ codes}$$

結果を【表 4】に示す。この結果より、上位 10 件で約 40%、上位 100 件で約 70% の IPC コードが正しく付与できていることが分かる。特許と論文を対象にした技術動向分析の支援を行うためには、上位 1 位における再現率のさらなる向上が必要であるものの、今回の結果は、特許の検索初心者にとってはある程度有効であると考えられる。一般に、特許を効率的に検索するためには、検索キーワードに加え IPC などの特許分類コードも併用される。しかし、検索初心者にとって、適切な特許分類コードの選択そのものが困難であり、これには

ある程度の技術と経験が必要とされる。このような場合、ユーザが本システムに調べたい分野の論文を入力すれば、その論文と関連する IPC コードが列挙される。【表 4】から、ユーザが結果の上位 20 件まで見れば、50% 以上の確率で該当する IPC コードが得られることから、特許検索初心者に対する IPC コードを用いた特許検索の敷居をある程度下げる効果があり、検索支援につながると思われる。

(SMT(Paper)+Index(Patent))

順位	Recall
1	0.110 (226/2051)
2	0.169 (347/2051)
3	0.215 (440/2051)
4	0.250 (512/2051)
5	0.277 (567/2051)
10	0.377 (774/2051)
20	0.467 (958/2051)
50	0.597 (1224/2051)
100	0.673 (1381/2051)
500	0.728 (1494/2051)
1000	0.728 (1494/2051)

表 4　上位 n 件の再現率

5 おわりに

本研究では、2 種類の翻訳システムを用いた学術論文の国際特許分類への自動分類手法を提案した。提案手法の有効性を検証するため、第 7 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て実験を行った。実験の結果、入力された英語論文から英語索引を作成した後、特許用フレーズテーブルを用いて索引語を作成する手法と、論文用翻訳システムを用いて英語論文を和訳した後、日本語索引を作成する手法を組み合わせた場合（“SMT(Paper)+Index(Patent)”）に、最も高い MAP 値：0。2897 が得られた。この値は、論文用翻訳システムを単体で用いた場合（“SMT(Paper)”）よりも高く、今回提案した 2 種類の翻訳システムを用いた分類手法が、ジャンルの異なる分類体系への文書分類に有効であることが実証された。

6 謝辞

本研究では、NTCIR-1、 2 の言語横断検索タスクおよび NTCIR-7 の特許マイニングタスクのデータを利用させていただいた。

参考文献

[Fujii 2007] Fujii, A., Iwayama, M., and Kando, N.: Overview of the Patent Retrieval Task at NTCIR-6 Workshop, Proc. the 6th NTCIR Workshop Meeting, pp. 359-365 (2007).

[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).

[Iwayama 2005] Iwayama, M., Fujii, A., and Kando, N.: Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task, Proc. the 5th NTCIR Workshop Meeting (2005).

[Iwayama 2007] Iwayama, M., Fujii, A., and Kando, N.: Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task, Proc. the 6th NTCIR Workshop Meeting (2007).

[Nanba 2007] Nanba, H.: Query Expansion using an Automatically Constructed Thesaurus, Proc. the 6th NTCIR Workshop Meeting, pp. 414-419 (2007).

[Nanba 2008:a] 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 on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 325-332 (2008).

[Nanba 2008:b] Nanba, H.: Hiroshima City University at NTCIR-7 Patent Mining Task. Proc. the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 369-372 (2008).