情報処理学会論文誌　データベース　Vol. 2　No. 3　76–86 (Sep. 2009)

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

難波英嗣[†1]　竹澤寿幸[†1]

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．論文を特許分類体系に分類するには，特許と論文で使われる用語の違いについて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．このため，単純に表層的な単語の一致度を用いる従来の文書分類モデルでは，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種類の翻訳モデル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．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，特許用翻訳システムと論文用のものを組み合わせたときに，論文用のシステムを単体で用いた場合と比べ，分類精度が改善できることが分かった．

Classification of Research Papers into a Patent Classification System Using Two Translation Systems

Hidetsugu Nanba[†1] and Toshiyuki Takezawa[†1]

Classification of research papers into patent classification systems enables exhaustive and effective invalidity search, prior art search, and technical trend analysis. However, it is very costly to ask research paper's authors or professionals to assign patent classification codes manually. Therefore, we propose a method that automatically classifies research papers into a patent classification system. To classify research papers into the classification system, we should take account of the differences of terms used in research papers and patents, because the terms used in patents are often more abstract or creative than those used in research papers, to try to widen the scope of the claims. Focusing on the classification of research papers written in various languages is also required for exhaustive searches and analyses. To solve these problems, we propose some classification methods using two machine translation systems. Generally, a performance of a machine translation system for patents is inferior to that for research papers, because the terms used in patents are different from those in research papers. However, we consider that the translation system for patents is useful for our task, because translation results by the translation system for patents tend to contain more patent terms than those for research papers. To confirm the effectiveness of our method, we conducted some examinations using the data provided from the Patent Mining Task in the NTCIR-7 Workshop. From the experimental results, we found that our method using translation systems for both research papers and patents could improve a method using single translation system.

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 [12] と NTCIR-6 [13] 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．たとえ

†1 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

ば，「機械翻訳」という論文用語[⋆1]に対する特許用語[⋆2]は「機械翻訳」のほかにも「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[19]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，関連研究について述べる．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. 関 連 研 究

本章では，「ジャンル横断情報アクセス」と「言語横断情報アクセス」に関する関連研究について述べる．

2.1 ジャンル横断情報アクセス

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[11]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[10]は，“Term Distillation” という手法を提案している．たとえば，「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．この考え方は，新聞記事とブログ記事の対応付けにも有効であることが Ikeda らにより確認されている[8]．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[21]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．たとえば，論文用語「フロッピーディスク」は特許用語「磁気記録媒体」に自動変換される．しかし，この手法では，論文用語をユーザが 1 語ずつ入力することが前提となっている．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が，本タスクのように，学術論文を入力とする場合，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．

特許と論文を対象にした情報アクセスに関するこのほかの研究として，NTCIR-7 特許マイニングタスク[19]があげられる．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，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．

特許マイニングタスクでは，論文に付与する国際特許分類（IPC）コード数が 30,855 件と非常に多く，さらに，訓練用データが 350 万〜450 万件と膨大である．このため，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を用いる参加グループは 2 グループのみで，他の参加グループは，k-Nearest Neighbor（k-NN）法を採用した．また，特許と論文で使われる用語の違いについて取り組んだ参加グループは 1 グループ[16]だけあったが，その有効性を確認するまでには至っていない．本研究では，特許マイニングタスクの中でも，

⋆1 ここで，論文用語とは，論文を検索したり分類したりする際に実際に使われうる用語全般を指す．このため，「フロッピーディスク」や「ワードプロセッサ」のように，元々はある分野でのみ使われていた専門用語が一般用語化してしまったものも，論文検索や分類の際に有用と考えられる用語はすべて「論文用語」に含まれる．

⋆2 特許用語とは，「特許を検索したり分類したりする際に有用な用語」と定義する．特許用語の中には，特許固有の用語だけでなく，論文で使われる用語（入力された論文用語の同義語や上位語など）も含まれる可能性がある．

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[⋆1]，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ほかの研究として，TREC Chemistry Track[⋆2]があげられる．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 1 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．その研究の詳細は，2009 年 11 月に開催される会議で報告される予定である．

2.2　言語横断情報アクセス

機械翻訳技術を用いて異なる言語で記述された文書情報にアクセスする，いわゆる言語横断情報アクセスに関する研究は，自然言語処理分野において長い歴史を持つ．ユーザが入力した検索質問と異なる言語で記述された文書を検索する言語横断検索は，特許や論文など技術文書を対象にしたものだけでも，たとえば，NTCIR-1[14] と NTCIR-2[15] では論文を対象に，また，NTCIR-4[3]，NTCIR-5[4]，NTCIR-6[5] では特許対象に，これまでに実施されている．ヨーロッパを中心に開催されている評価ワークショップ CLEF でも，2009 年から特許を対象にした言語横断検索タスク CLEF-IP[⋆3]が始まっている．このほか，前述の NTCIR-7 における特許マイニングタスク[19] の言語横断サブタスクも，言語横断情報アクセス研究の 1 つとしてあげられる．

ここでは，本研究と関連の深い特許マイニングタスクの言語横断サブタスク（J2E）に参加した 2 つの研究グループのシステム[1),2)] について述べる．Bian ら[1] は，Web 上で利用可能な 3 つの翻訳システム（Google，Excite，Yahoo! Babel Fish）を用いて日本語論文を英訳した後，k-NN 法を用いた文書分類器により，言語横断文書分類を実現している．この分類器では，tf*idf 法により単語の重み付けを行うベクトル空間型モデルを採用している．Clinchant ら[2] は，NTCIR-1 の論文データベース[14] から抽出した日英論文表題の対訳約 30 万件から，Giza[⋆4]を用いて日英単語対訳辞書を自動構築している．この辞書を，k-NN 法による言語モデルに基づいた文書分類器と組み合わせることにより，言語横断文書分類を行っている．Bian ら，Clinchant らは，ともに文書分類器のみを用いて英語サブタスクにも参加しており，どちらも言語横断サブタスクとほぼ同等の分類精度を得ている[⋆5]．これらの結果から，Bian ら，Clinchant らは，言語横断という側面からの提案手法の有効性は示したものの，ジャンル横断については特別な処理を行っていない．本研究では，日本語の文書分類器を利用する関係で翻訳の方向は異なるが，言語横断サブタスク（E2J）のデータを用い，言語横断文書分類の実験を行う．本研究でも Clinchant らと同様に，日英対訳文から Giza を用いて辞書を自動的に構築するが，Clinchant らが論文データのみを用いているのに対し，本研究では，論文データおよび特許データから 2 種類の辞書（および翻訳システム）を構築し，それらを組み合わせて利用することにより，「特許と論文で使われる用語が異なる」ジャンル横断文書分類の問題も解決しようとする点が Clinchant らと異なる．

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うえで（図中 O'），(3) 文書分類を実施する，という 3 つのステッ

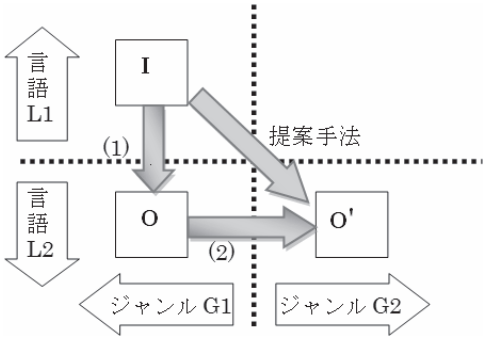

図 1　提案手法
Fig. 1　Our method.

⋆1 このサブタスク（E2J）のデータを用いた理由は，著者らが利用可能な文書分類器が，k-NN 法を用いた日本語サブタスク用のものに限られていたためである．
⋆2 https://wiki.ir-facility.org/index.php/TREC_Chemistry_Track
⋆3 http://www.ir-facility.org/the_irf/current-projects/clef-ip09-track/
⋆4 http://www.fjoch.com/GIZA++.html
⋆5 言語横断サブタスクにおける Bian らの MAP（Mean Average Precision）値：0.0934〜0.1070．Clinchant らの MAP 値：0.4380

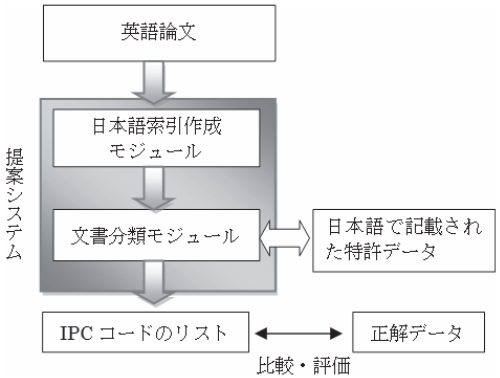

図 2　システム概要
Fig. 2　System configure.

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．

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

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
Fig. 3　An example of an English title and an abstract.

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 サンドブラスト法
9 サンドブラスト
（以下略）

図 4　日本語索引
Fig. 4　An example of a Japanese index.

した後，内容語[⋆2]を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[⋆3]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

⋆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．
⋆2 形態素解析器 MeCab（http://mecab.sourceforge.net/）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，形容詞を内容語とする．
⋆3 英文の品詞タグ付けには TreeTagger（http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/）を用いた．

Giza を，デコーダには Moses[⋆1] 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[6] 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[14]，NTCIR-2[15] 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，それぞれすべて用いた．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．索引語の翻訳には，このほかにも既存の専門用語辞書を使う方法や要素合成法[23] などがあげられるが，専門用語を含む複合名詞の翻訳におけるフレーズテーブルの有効性は，森下ら[17]，Itagaki ら[9] が報告しているため，本研究でもフレーズテーブルを用いて索引語を翻訳する．なお，比較のため，既存の専門用語辞書を用いた場合についても実験を行う．詳細は 4.2 節を参照されたい．

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[20]．この分類器は，NTCIR-6 特許検索タスク[5] 用に開発された特許検索システム[18] 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[22] 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．

$$Score(X) = \sum_{i=1}^{n} Relevance\ score\ of\ each\ patent$$

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．また，単語の重みの計算には tf（term frequency）を，類似度尺度には SMART を，それぞれ用いる．なお，手順 1 における上位 170 件という値を，Nanba は NTCIR-7 特許マイニングタスク訓練用およびドライラン用に配布されたデータを用いて決定している．

3.3　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 1 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．

IPC が付与されたデータは，今回実験に用いる日本国特許データ以外にも，たとえば米国特許データや JAPIO 英文抄録データなど，英語で記述された特許データが存在する．これらのデータを用いれば，処理の過程で機械翻訳を行う必要がないため，「論文を国際特許分類に分類する」という目的を達成するためには，あえて記述言語が異なるデータを利用する必然性はない．しかしながら，特許分類体系には，国際特許分類以外にもファイルインディックス（FI）や F タームといった日本国特許固有の分類体系が存在する．今回は，実験に用いるデータセットの制約上，「国際特許分類への自動分類」を目標としているが，将来は本システムの拡張による FI や F タームなどへの分類も視野に入れている．また，本論文で提案する手法は，翻訳の方向を英日から日英に変えれば米国特許データを用いて日本語論文を国際特許分類に自動分類することも可能になる．この場合，国際特許分類に限らず，米国特許固有の分類体系である US Class への日本語論文の自動分類も将来的には実現可能になると考えられる．

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

⋆1 http://www.statmt.org/moses/

⋆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．

表 1　文書データ
Table 1　Document data.

データ名	年	サイズ	文書数	言語
日本国公開特許公報	1993-2002	100 GB	3.5 M	日
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）	1988-1999	1.4 GB	0.26 M	日/英

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（抄録ごとに最大 1,000 件）と比較し，以下に定義する MAP（Mean Average Precision）により評価した．

$$MAP = \frac{1}{Q}\sum_{j=1}^{Q}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

ここで，Q，m_j，R は，それぞれ分類対象の論文抄録数，適合した IPC コードの数，検索された IPC コード集合を示す．なお，正解データは，2 段階の判定基準により作成されている．879 件の論文抄録は，より信頼度の高い IPC コードが付与されているものを “group A”，group A よりもやや信頼度が劣るものを “group B” として分けられており，前者に属するものが 473 件，後者が 406 件存在する．本研究では，group A のみを用いた場合と，879 件すべてを用いた場合の 2 通りで評価を行う．なお，正解データの作成に関する詳細は，Nanba らの報告書[19] を参照されたい．

文書データ

実験に用いた文書データを**表 1** にまとめる．

日本国公開特許公報には，全文データの他に人手で IPC コードを付与したデータも含まれており，本研究ではこのデータを文書分類モジュールで k-NN 法を適用する際に利用している．論文抄録データは，3.2 節で述べたとおり，論文用翻訳モデルの獲得に利用している．

4.2　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，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．また，“Trans(X)” と “Lang(X)” は，それぞれ「X 翻訳用フレーズテーブル」と「X 翻訳用言語モデル」を示す．

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

比較手法

- SMT(Paper)$^{\star 1}$：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)$^{\star 2}$：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
- Trans(Paper)+Lang(Patent)：論文翻訳用フレーズテーブルと特許翻訳用言語モデルを組み合わせて英語論文を和訳した後，日本語索引を作成
- Trans(Paper)+Lang(Patent)+Index(Paper)：Trans(Paper)+Lang(Patent) と Index(Paper) の和集合
- Trans(Paper)+Lang(Patent)+Index(Patent)：Trans(Paper)+Lang(Patent) と Index(Patent) の和集合
- Trans(Patent)+Lang(Paper)：特許翻訳用フレーズテーブルと論文翻訳用言語モデルを組み合わせて英語論文を和訳した後，日本語索引を作成
- Trans(Patent)+Lang(Paper)+Index(Paper)：Trans(Patent)+Lang(Paper) と Index(Paper) の和集合
- Trans(Patent)+Lang(Paper)+Index(Patent)：Trans(Patent)+Lang(Paper) と Index(Patent) の和集合

$\star 1$ = Trans(Paper)+Lang(Paper)
$\star 2$ = Trans(Patent)+Lang(Patent)

表 2　評価結果
Table 2　Evaluation results.

	手法	group A+B	group A
提案手法	Index(Paper)*Index(Patent)	0.1830	0.2230
	Index(Paper)+Index(Patent)	0.2258	0.2596
	SMT(Paper)+Index(Patent)	**0.2633**	**0.2897**
比較手法	SMT(Paper)	0.2518	0.2777
	SMT(Patent)	0.2214	0.2507
	Index(Paper)	0.2169	0.2433
	Index(Patent)	0.2000	0.2373
	Trans(Paper)+Lang(Patent)	0.2108	0.2507
	Trans(Paper)+Lang(Patent)+Index(Paper)	0.2352	0.2758
	Trans(Paper)+Lang(Patent)+Index(Patent)	0.2338	0.2758
	Trans(Patent)+Lang(Paper)	0.2296	0.2621
	Trans(Patent)+Lang(Paper)+Index(Paper)	0.2454	0.2750
	Trans(Patent)+Lang(Paper)+Index(Patent)	0.2347	0.2669
上限	日本語サブタスク (理想的な翻訳)	0.2958	0.3267
	日本語サブタスク+Index(Patent)	0.3001	0.3277

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
- 日本語サブタスク+Index(Patent)：“日本語サブタスク” で作成された日本語索引を “Index(Patent)” により拡張

システムの上限として，論文著者自身が作成した理想的な論文翻訳から日本語索引を作成した “日本語サブタスク”，さらにこの “日本語サブタスク” を “Index(Patent)” で拡張した方法でも実験を行った．

4.3 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”）を用いたとき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 1 つである “SMT(Paper)+Index(Patent)” が最も高い MAP 値を得た．

4.4 考察

実際に作成された Index(Paper) と Index(Patent)

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．

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
Table 3　Translation results of an English index using two phrase tables.

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
high resolution display	高精細	高解像度ディスプレィ
resolution	分解能	解像度
barrier	障壁	バリア

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
Fig. 5　An example of a Japanese research paper.

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．

Index(Paper) と Index(Patent) の比較

“Index(Paper)” と “Index(Patent)” の MAP 値の差は小さく，どちらも高い値が得ら

表 4　翻訳確率の上位 n 件を訳語として用いた場合の Index(Paper) と Index(Patent) の重なり索引語数
Table 4　The number of overlapped index terms between Index(Paper) and Index(Patent).

翻訳確率 上位 n 件	Index(Paper)	Index(Patent)
1	0.394 (17066/ 43347)	0.435 (17066/ 39235)
3	0.291 (36995/127334)	0.324 (36995/114243)
5	0.245 (51120/208747)	0.276 (51120/185493)
7	0.217 (62655/288251)	0.246 (62655/254455)
9	0.202 (73993/366362)	0.230 (73993/321942)

表 5　翻訳確率の上位 n 件を訳語として用いた場合の Index(Paper) と Index(Patent) の MAP 値
Table 5　MAP scores of Index(Paper) and Index(Patent) using top n translations.

翻訳確率 上位 n 件	Index(Paper)		Index(Patent)	
	group A+B	group A	group A+B	group A
1	0.217	0.243	0.200	0.237
3	0.220	0.250	0.215	0.255
5	0.220	0.248	0.211	0.254
7	0.221	0.252	0.212	0.257
9	0.225	0.258	0.210	0.253

れている[⋆1]．Index(Paper) と Index(Patent) の MAP 値の差について調べるため，“Index(Paper)” と “Index(Patent)” により作成された日本語索引の比較を行った．英語索引中の総用語数 69,100 語について，論文用フレーズテーブルで翻訳できたものは 47,055 語（47,055/69,100=0.681）であるのに対し，特許用では 40,427 語（40,427/69,100=0.585）であり，特許用が論文用と比べ 10%近く翻訳できていないことが分かった．この差は，論文と特許で使われる用語に差異があることを示唆しており，これが “Index(Patent)” と “Index(Paper)” の分類精度の差となって現れていると考えられる．

“Index(Paper)” と “Index(Patent)” は，どちらもフレーズテーブル中で最も翻訳確率の高いものだけを用いて索引語を和訳している．ここで，翻訳確率が最も高いものだけでなく，翻訳確率が高い順に上位 n 語を索引語として採用した場合に，“Index(Paper)” と “Index(Patent)” 間で索引語の重なりがどの程度あるのか，また，n の値を 1，3，5，7，9[⋆2]と増やした場合に，分類精度がどのように変化するのか，調査を行った．前者については**表 4** に，後者は**表 5** に，それぞれ結果を示す．

表 4 から，n の値を増やすと重なり索引語数は増加しているものの，索引語総数の中で占める割合は低下している．一方，表 5 において，n の値が増えるにつれ MAP 値が若干増加するものの，提案手法を上回るほど大幅には変動していない．以上のことから，(1) n の値を増やしたときに増加する重なり索引語の大半は，MAP 値上昇に貢献しないものであり[⋆3]，(2) 分類精度に影響する内容語のほとんどは翻訳確率上位 1 件に含まれており，(3) n の値を変えても “Index(Paper)” と “Index(Patent)” の MAP 値に差があることから，論文と特許で使われる用語に差異があることを示唆していると考えられる．

Index(Paper) と Index(Patent) の組合せ

論文用と特許用のフレーズテーブルを使って，ある英語の索引語の和訳したとき，両者が一致すれば，一致しない場合と比べ正しい翻訳である可能性が高いと考えられる．そこで，提案手法の 1 つとして，訳語が一致したもののみを日本語索引として用いる “Index(Paper)*Index(Patent)” で MAP 値を調べた．その結果，group A+B を用いた場合に 0.1830，group A のみを用いた場合に 0.2230 と，比較的良い値が得られた．これは，論文と特許では語彙空間の重なりが比較的大きく，この重なり部分の語彙のみを使って分類しても，ある程度の分類精度が得られることを示唆している．しかしながら，“Index(Paper)” および “Index(Patent)” を単体で用いた場合と比べ，MAP 値が約 2%低下していることから，重なり部分の語彙だけでは十分ではないと考えられる．

逆に，“Index(Paper)” と “Index(Patent)” の共通項だけでなく差分もまとめて 1 つの日本語索引を作成した場合（“Index(Paper)+Index(Patent)”）についても MAP 値を調べた．その結果，group A+B を用いた場合に MAP 値 0.2258 が，group A のみを用いた場合に 0.2596 が，それぞれ得られた．この値は，“Index(Paper)” および “Index(Patent)” を単体で用いた場合と比べ，MAP 値が 1〜2%高い．この結果から，2 つの翻訳システムを用いることにより，ジャンルの違いにより使われる用語の違いを考慮した文書分類が実現できたといえる．

SMT(Paper)+Index(Patent) の有効性

論文用翻訳システムを用いた手法は，“Index(Paper)” だけでなく “SMT(Paper)” も単体で高い MAP 値を得ている．そこで，論文用翻訳システムは先に全文を和訳した後日本語索引を作成，特許用翻訳システムは先に英語索引を作成した後，フレーズテーブルを用いて索

⋆1 なお，このほかにも，既存の日英専門用語辞書として科学技術 45 万語対訳辞典（日外アソシエーツ株式会社，2001）および英辞郎を用いた実験も行ったが，どちらの場合も Index(Paper) および Index(Patent) よりも MAP 値が低い結果となった．
⋆2 フレーズテーブル獲得の際に，ある用語に対する訳語候補を最大 10 件までしか獲得しない設定にしているため，n の値を 1〜9 の範囲で調査を行った．
⋆3 おそらく翻訳確率の低い語の多くは，不適切な訳語であると推測される．

引語を作成した後，両者を組み合わせた日本語索引を用いて実験を行った．実験の結果，この手法では group A を用いた場合と group A+B を用いた場合の両方で，最も高い MAP 値：0.2633 と 0.2807 が得られた．この値は，理想的な翻訳を用いたシステムの上限 “日本語サブタスク” にせまる値である．

フレーズテーブルと言語モデルの組合せ

ベースライン手法 “Trans(Paper)+Lang(Patent)” と “Trans(Patent)+Lang(Paper)” は，いずれも “SMT(Paper)” の MAP 値を下回る結果となった．特に，“Trans(Paper)+Lang(Patent)” の MAP 値が “Trans(Patent)+Lang(Paper)” の MAP 値よりも低いことから，ジャンルの違いによる言語モデルの差が，翻訳の質だけでなく分類精度にも影響していると考えられる．

SMT(Paper) と Index(Paper) の比較

“SMT(Paper)” と “Index(Paper)” を比較すると，“SMT(Paper)” の方が “Index(Paper)” よりも MAP 値が 0.0349 高い．文を生成すると MAP 値が向上するということから，本タスクでは，論文翻訳用言語モデルの影響により，より適切な訳語が選択できていると考えられる．実際に，“SMT(Paper)” と “Index(Paper)” の索引語の重なりを調査したところ，“SMT(Paper)” で作成された全索引語 56567 語のうち，“Index(Paper)” と重なっているものはわずか 24.5%（13889/56567）であることが分かった．

Index(Patent) を用いた日本語論文抄録の自動分類

“Index(Patent)” は，“SMT(Paper)” と “Index(Paper)” の MAP 値が改善できることが確認できたが，理想的な翻訳を用いた “日本語サブタスク” も改善できるか，調査を行った（“日本語サブタスク+Index(Patent)”）．日本語サブタスク用のトピックデータから日本語インディックスを作成した後，“Index(Patent)” 中で日本語インディックスに含まれないもののみを追加し，実験を行った．実験の結果，group A+B を用いた場合と group A を用いた場合の両方において，若干ではあるが MAP 値が改善されることが確認できた．

提案手法の実用性

最後に，提案手法の 1 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ものかを調べるため，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．

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

結果を**表 6** に示す．表では，group A のトピックのみを用いた場合と，すべてのトピックを用いた場合の結果を示している．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 6 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．

表 6　上位 n 件の再現率（SMT(Paper)+Index(Patent)）
Table 6　Recall for top n results (SMT(Paper)+Index(Patent)).

順位	group A	group A+B
1	0.117 (131/1115)	0.110 (226/2051)
2	0.186 (207/1115)	0.169 (347/2051)
3	0.239 (267/1115)	0.215 (440/2051)
4	0.278 (310/1115)	0.250 (512/2051)
5	0.311 (347/1115)	0.277 (567/2051)
10	0.420 (468/1115)	0.377 (774/2051)
20	0.524 (584/1115)	0.467 (958/2051)
50	0.659 (735/1115)	0.597 (1224/2051)
100	0.733 (817/1115)	0.673 (1381/2051)
500	0.775 (864/1115)	0.728 (1494/2051)
1000	0.775 (864/1115)	0.728 (1494/2051)

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．

謝辞　本研究では，NTCIR-1，NTCIR-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

1) Bian, G.-W. and Teng, S.-Y.: Integrating Query Translation and Text Classification in a Cross-Language Patent Access System, *Proc. 7th NTCIR Workshop Meeting*, pp.341–346 (2008).
2) Clinchant, S. and Renders, J.-M.: XRCE's Participation to Patent Mining Task at NTCIR-7, *Proc. 7th NTCIR Workshop Meeting*, pp.351–353 (2008).
3) Fujii, A., Iwayama, M. and Kando, N.: Overview of Patent Retrieval Task at NTCIR-4, *Working Notes of the 4th NTCIR Workshop*, pp.225–232 (2004).
4) Fujii, A., Iwayama, M. and Kando, N.: Overview of Patent Retrieval Task at NTCIR-5, *Proc. 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp.269–277 (2005).
5) Fujii, A., Iwayama, M. and Kando, N.: Overview of the Patent Retrieval Task at NTCIR-6 Workshop, *Proc. 6th NTCIR Workshop Meeting*, pp.359–365 (2007).
6) Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp.389–400 (2008).
7) Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proc. 14th International Conference on Computational Linguistics*, pp.539–545 (1992).
8) Ikeda, D., Fujiki, T. and Okumura, M.: Automatically Linking News Articles to Blog Entries, *Proc. AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs*, pp.78–82 (2006).
9) Itagaki, M., Aikawa, T. and He, X.: Automatic Validation of Terminology Translation Consistency with Statistical Method, *Proc. MT summit XI*, pp.269–274 (2007).
10) Itoh, H., Mano, H. and Ogawa, Y.: Term Distillation for Cross-db Retrieval, *Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task*, pp.11–14 (2002).
11) Iwayama, M., Fujii, A., Kando, N. and Takano, A.: Overview of Patent Retrieval Task at NTCIR-3, *Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task*, pp.1–10 (2002).
12) Iwayama, M., Fujii, A. and Kando, N.: Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task, *Proc. 5th NTCIR Workshop Meeting* (2005).
13) Iwayama, M., Fujii, A. and Kando, N.: Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task, *Proc. 6th NTCIR Workshop Meeting* (2007).
14) Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H. and Hidaka, S.: Overview of IR Tasks at the First NTCIR Workshop, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.11–44 (1999).
15) Kando, N., Kuriyama, K. and Yoshioka, M.: Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop, *Proc. 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pp.4-37–4-60 (2001).
16) Mase, H. and Iwayama, M.: NTCIR-7 Patent Mining Experiments at Hitachi, *Proc. 7th NTCIR Workshop Meeting*, pp.365–368 (2008).
17) 森下洋平，宇津呂武仁，山本幹雄：対訳特許文書からの専門用語対訳辞書半自動獲得におけるフレーズテーブルと既存対訳辞書の併用，情報処理学会自然言語処理研究会，NL-187, pp.91–98 (2008).
18) Nanba, H.: Query Expansion using an Automatically Constructed Thesaurus, *Proc. 6th NTCIR Workshop Meeting*, pp.414–419 (2007).
19) Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp.325–332 (2008).
20) Nanba, H.: Hiroshima City University at NTCIR-7 Patent Mining Task, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp.369–372 (2008).
21) 難波英嗣，釜屋英昭，竹澤寿幸，奥村　学，新森昭宏，谷川英和：論文用語の特許用語への自動変換，情報処理学会論文誌：データベース，pp.81–92 (2009).
22) Salton, G.: *The SMART Retrieval System — Experiments in Automatic Document Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ (1971).
23) 外池昌嗣，宇津呂武仁，佐藤理史：ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定，自然言語処理，Vol.14, No.2, pp.33–68 (2007).

(平成 21 年 3 月 20 日受付)
(平成 21 年 7 月 6 日採録)

(担当編集委員　江口 浩二)

難波　英嗣（正会員）

1996 年東京理科大学理工学部電気工学科卒業．1998 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．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）．テキストマイニング，情報検索，自動要約，特許情報処理に関する研究に従事．言語処理学会，人工知能学会，ACL，ACM 各会員．

竹澤　寿幸（正会員）

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．