情報処理学会論文誌　データベース　Vol.6　No.2　16–29 (Mar. 2013)

論文と特許からの技術動向情報の抽出と可視化

福田 悟志[1,a)]　難波 英嗣[1]　竹澤 寿幸[1]

受付日 2012年9月20日, 採録日 2012年12月29日

概要：産業と関連性が高い分野の研究者にとって，論文や特許などの技術文書を検索・分析することは，その分野の動向を知るうえで重要である．本研究では，このような作業を支援するため，論文と特許から技術動向に関する情報を抽出し，マップを自動作成して可視化する手法を提案する．技術動向マップの構築には，特定分野において使用された基礎的な要素技術とその効果に着目する．このような要素技術とその効果の変遷を知ることは，その分野における技術動向のあらましを把握する重要な情報となる．そこで本研究では，様々な研究分野における要素技術とその効果に関する表現を自動的に抽出するための手法を提案する．本研究の有効性を確かめるために，NTCIR-8 特許マイニングタスクで提供されたデータを用いて実験を行った結果，本研究で提案した手法において，論文解析では再現率 0.254，精度 0.496 が，特許解析では再現率 0.441，精度 0.537 が得られた．

キーワード：情報抽出，SVM，ドメイン適応，分布類似度

Extraction and Visualization of Technical Trend Information from Research Papers and Patents

Satoshi Fukuda[1,a)]　Hidetsugu Nanba[1]　Toshiyuki Takezawa[1]

Received: September 20, 2012, Accepted: December 29, 2012

Abstract: For a researcher in a field with high industrial relevance, retrieving and analyzing research papers and patents have become an important aspect of assessing the scope of the field. We propose a method for creating a technical trend map automatically from both research papers and patents. For the construction of the technical trend map, we focus on the elemental (underlying) technologies used in a particular field, and their effects. Knowledge of the history and effects of the elemental technologies used in a particular field is important for grasping the outline of technical trends in the field. Therefore, we have constructed a method that can recognize the application of elemental technologies and their effects in any research field. To investigate the effectiveness of our method, we conducted an experiment using the data in the NTCIR-8 Patent Mining Task. From our experimental results, we obtained recall and precision scores of 0.254 and 0.496, respectively, for the analysis of research papers. We also obtained recall and precision scores of 0.441 and 0.537, respectively, for the analysis of patents.

Keywords: information extraction, SVM, domain adaptation, distributional similarity

1. はじめに

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，分析したりする機会が増えている．2012 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 2012」[*1]においても，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論文・特許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，このような傾向は今後さらに強まっていくと考えられる．

論文と特許を検索するのは，大学研究者に限った話では

[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, Hiroshima 731–3194, Japan
[a)] fukuda@ls.info.hiroshima-cu.ac.jp

[*1] http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku2012.pdf

© 2013 Information Processing Society of Japan

ない．たとえば，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．これは一般に先行技術調査と呼ばれている．このほかに，サーチャ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無効資料調査でも，論文と特許が検索や分析の対象となる．

しかしながら，限られた時間で特定分野の論文や特許を網羅的に収集・分析することは容易ではない．こうした状況に鑑み，本研究では，論文と特許を対象に，特定分野の技術動向を把握するのに有用なシステムの開発を目指す．

システムを構築するにあたって，本研究では特定分野において使用された基礎的な要素技術とその効果に着目する．本研究における「要素技術」とは，研究において使用されたアルゴリズムやツール，技術的手法のことを指し，また，その要素技術から得られる知見を「効果」と定義する．また，効果に関する表現の中には，要素技術を用いることで得られた特徴・性質を表す「属性」表現，および属性に付随する値を表す「属性値」表現が含まれているとする．これらの表現を収集することで，ある特定の分野内で使用された要素技術から得られた効果の変遷を知ることができ，その結果，その分野内における技術動向のあらましを効率的に把握することができると考えられる．

これまでにも，人手で作成したルールに対応付けて論文や特許を解析している研究は多く存在する．しかし，論文と特許では，表現や形式など，記述スタイルの面で大きく異なっている．また，同じ論文や特許における同じ意味を持つ文章でも，作成した著者によって表現がそれぞれ異なる．このように，様々な形式・表現で記述されている文章をルールに対応付けて解析することは難しい．本研究ではこの問題を「要素技術とその効果を示すタグを付与」するという系列ラベリング問題として考え，機械学習を用いてタグの自動付与を目指す．

一般に，論文の表題や概要，および特許の「発明の名称（以後，特許の表題）」や「発明の詳細な説明（以後，特許の概要）」では，「を用いた」や「を具備する」といった表現の直前には要素技術を表す用語が出現する．一方で，「が可能になる」や「ができる」の直前には効果を表す用語が出現する可能性が高い．たとえば，「磁気ストライプを用いることで，非常に安価な製造が可能になった」という論文概要の場合，「磁気ストライプ」が要素技術を，「安価な製造」が効果を示している．また，この効果表現の中において，「製造」が属性，「安価」が属性値を表している．そこで，要素技術とその効果を示す手がかり語のリストを作成しておき，各々のリスト中の手がかり語の有無を素性として扱い機械学習に用いることで，解析精度の向上を目指す．このほか，「精度」や「信頼性」のように属性になりやすい用語や，「向上」や「高速化」のように属性値になりやすい用語が存在する．このような用語を収集してリストを作成しておけば，これらの用語の有無を機械学習の素性として用いることができる．しかし，様々な分野の属性と属性値を人手で網羅的に収集するのは容易ではない．そこで本研究では，機械学習に用いる手がかり語の収集方法として，係り受け関係や上位下位関係による人手での収集，さらに分布類似度を用いた語句の自動収集を行うことで，様々な分野における表現を網羅的に収集することを目指す．

また，一般に，論文用および特許用の抽出器を機械学習により獲得するために，論文用または特許用タグ付きコーパスを単独で用いる．しかし，単独のコーパスから得られる学習量には限界がある．そこで本研究では，ドメイン適応手法を用いることで，構造解析する論文または特許に対して，論文用および特許用コーパスの両方を使用し，学習量を増加させることで，解析精度の向上を試みる．

本稿の構成は以下のとおりである．次章では本研究で実際に開発した技術動向分析システムの動作例について説明する．3 章では，関連研究について述べる．4 章では，論文および特許の表題と概要の解析手法を述べ，5 章では，有効性を調べるために行った実験について報告し，結果を考察する．最後に 6 章で本稿をまとめる．

2. 技術動向分析システムの動作例

本章では，論文と特許を対象にした技術動向を分析・可視化するシステムの動作例および仕組みについて説明する．本システムは，国立情報学研究所の論文情報ナビゲータである CiNii [*2] に収録されている論文データ，および NTCIR テストコレクションで配布された公開特許公報全文データを対象に，技術動向に関する情報を自動的に抽出し，マップとしてユーザに提示する．**図 1** は，「論理回路」という用語をシステムに入力したときの技術動向マップの一部を示している．図 1 において，左側に「論理回路」に関する各論文および特許中で使われている要素技術が列挙され，

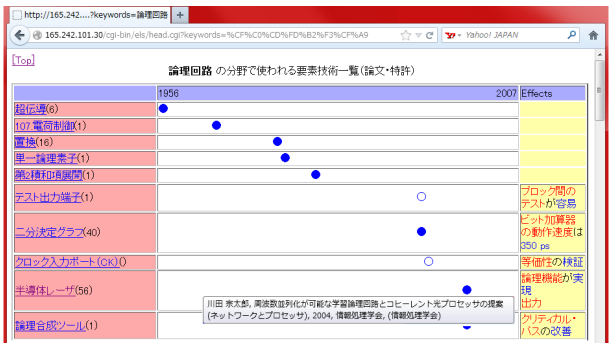


図 1　「論理回路」で使われる要素技術と効果の一覧表示

Fig. 1　A list of elemental technologies and effects used in the “Logic Circuit”.

*2　http://ci.nii.ac.jp/

© 2013 Information Processing Society of Japan

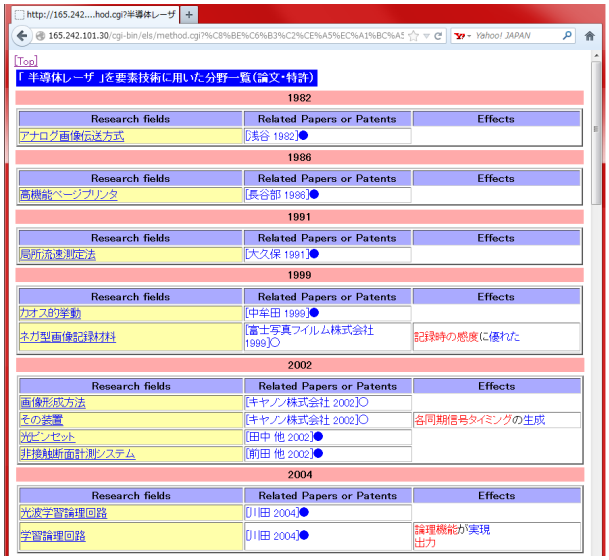


図 2　「半導体レーザ」を要素技術として用いている分野と効果の一覧表示

Fig. 2　A list of research fields and effects that use “Semiconductor Laser” as an elemental technology.

その右側に各技術が使われた年が表示されている．たとえば図 1 において「半導体レーザ」が論理回路の要素技術として 2004 年に使われていることを示している．図中の「●」（「○」）は，「半導体レーザ」を要素技術として用いている論文（特許）を意味しており，ユーザが「●」にカーソルを重ねることで，その文献の書誌情報がポップアップウィンドウ内に表示される．さらに，「●」をクリックすれば，文献の詳細な情報にアクセスできる．「●」は論文を，「○」は特許を表している．

図 1 において，要素技術として提示されている用語をユーザがクリックすることで，その要素技術がどのような分野で利用されているのかを，年代順に一覧表示することができる．**図 2** は，図 1 中の「半導体レーザ」をクリックした結果を示している．学術分野では 2002 年までにおいて，主に画像系の分野で使われていた技術が，2004 年に入ると論理回路の分野でも利用されていることが，一覧表示の結果より分かる．真ん中には，「半導体レーザ」を要素技術に用いた分野における関連文書情報を列挙しており，「●」は論文を，「○」特許を表している．

さらに，各要素技術の効果に関する情報が，各図の右端に表示される．図 1 では，「論理回路」の分野で「論理合成ツール」という技術から「クリティカル・パスの改善」という効果が得られることが分かる．また，図 2 では，様々な分野においてある要素技術にどのような効果があるのか一覧できる．

3.　関連研究

これまでに，論文や特許を対象にした技術動向分析に関する研究やシステムの構築が多く行われている．本節では，これらの研究や関連システムについて述べる．

3.1　表題の構造解析およびその応用

Matsumura ら [1] は，処理の負荷が小さく精度が高い形態素解析を使って，単語間の係り受け関係を情報検索に利用することを提案している．Matsumura らの検索手法では，名詞，形容詞のような，ある概念を表すキーワードである「概念語」と，助詞，動詞のような，概念語どうしの関係を表す「関係語」からなる「構造化インデックス」を作成している．「概念語」とは，ある概念を表すキーワードである．たとえば，名詞，形容詞，副詞である．「関係語」とは，概念語どうしの関係を表すものである．たとえば，助詞，助動詞，動詞である．Matsumura らは，このインデックスを情報検索エンジンのインデックスに利用している．自然文に対して係り受け関係を解析し，構造化インデックスを作成することは，複雑な自然言語処理を必要とするため，文書表題のような擬似的な自然文を対象としている．Matsumura らは，人手で作成したパターンに基づいて概念語と関係語の係り受け関係を判定していたが，本研究では，表題の構造解析する際に機械学習を取り入れてパターン作成の自動化を図る．

今井 [2] は，日本語論文表題の構造を解析し，その結果を用いて論文を自動分類する手法を提案している．この手法は，「標準化」と「コード割当て」という 2 つの処理から構成される．「標準化」処理では，文字列処理による不要部分の削除・分割を行い，文字列処理による木構造の変形を行う．その後，単語列処理による不要部分の削除・分割を行う．この処理を繰り返して適用することによって，論文表題をいくつかの部分要素に分割する．「コード割当て」処理では，それぞれの部分要素中の専門用語を抽出し，その用語を岩波情報科学辞典のコードと対応付けることで論文の分類を実現している．論文表題を構造解析し主題を抽出するという点では今井の研究と共通するが，本研究では，表題解析に機械学習を取り入れている点と，要素技術に着目して処理を行う点が異なる．

3.2　技術動向分析

研究動向の調査に関して，村田ら [3] の研究がある．村田らは，言語処理学会年次大会および論文誌の第 1 回から第 10 回までの 10 年間において，どういった研究がなされてきたかを調べている．調査方法は，電子書誌情報から形態素解析器を用いて名詞を抽出し，その頻度を並べることで，様々な側面から自然言語処理分野の研究動向分析を行っている．この分析では，論文表題中の名詞はすべて等価に扱われているが，本研究では，論文表題の構造を解析することで，要素技術と効果を示す用語を識別する．

難波ら [4] の研究では，「を用いた」，「における」のような論文題目中に非常に多用される表現を手がかり語とし

© 2013 Information Processing Society of Japan

て題目中の名詞句（専門用語）間の関係を明らかにしている．そして，論文表題から要素技術用語を以下の解析手順で抽出している．まず，「を用いた」，「における」，「のための」のような手がかり語と「METHOD」，「RESTRICT」，「GOAL」のような構造タグの対応リストを用意しておく．次に，論文表題と作成した手がかり語を比較し，表題中で一致する文字列を構造タグに置き換える．この解析手法を，たとえば「ニュース番組における字幕生成のための文短縮」という論文表題に用いた場合，「ニュース番組」と「字幕生成」に「RESTRICT」タグと「GOAL」タグがそれぞれ付与される．この解析結果から，「文短縮」の目的が「字幕生成」であることが分かり，「文短縮」が「字幕生成」の要素技術であることが分かる．難波らは，作成した対応リストのうち，「METHOD」と「GOAL」に着目し，要素技術用語を抽出している．本研究でも名詞句間の関係づけのため，手がかり語を提案手法の素性として用いる．

西山ら [5] は，技術文書から特定の技術エリアで生み出される新製品・新技術に関する記述をすばやく把握したいというニーズに応える技術文書マイニング手法を提案している．技術文書の中には，新技術，新製品が持つ好ましい性質や新機能などの効果に関することを述べているものがある．たとえば，「通話音質が向上する」である．このような表現を特長表現と呼んでいる．西山らは，複数の手がかり語を用いて，その手がかり語から特定量の単語分戻る形で特長表現を抽出している．しかしながら，抽出に用いる手がかり語が限定的であるため，たとえば，「精度が 0.935」など，数値で表現される効果には対応できないという問題がある．

また近年では，文書の潜在的な構造を抽出するための手法として，潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）[6] に基づくトピックモデルを用いた研究動向の調査が多くなされている．トピックモデルとは，Bag-of-words で表現された文書の生成過程を潜在的意味（トピック）に基づいて確率的に表現するモデルであり，科学技術の分野における研究アイディアの発展過程を調べる際に有効であることが報告されている [7], [8]．また，トピックモデルの特徴として，一般的な生成モデルと比べて拡張が容易であり，引用情報 [9] や著者情報 [10] など，多様な情報を統合することができることがあげられる．これにより，Bag-of-words のみでは考慮できない，学術論文における重要な固有表現を考慮することができる．トピックモデルでは，あるトピックに対して特徴的な語句を抽出するが，本研究では，係り受け関係や分布類似度などの統計的手法を利用して半自動的に収集した手がかり語を用いることにより，多様な効果表現の抽出を目指す．

論文と特許を対象にした技術動向分析に関するこのほかの研究プロジェクトとして，国立情報学研究所主催の第 8 回 NTCIR ワークショップ [11]（NTCIR-8）で実施されている特許マイニングタスクがある．このタスクでは，ある分野の論文と特許から，「要素技術」と「効果」という観点から分類した技術動向マップを自動的に作成することを目的としている．このような技術動向マップを自動的に作成するツールは，先行技術調査や，無効資料調査の支援ツールとして利用することができる．このようなマップを自動的に作成するために，以下の 2 つの手順が必要であると定めている．

（手順 1）　ある分野の論文と特許を網羅的に収集する．
（手順 2）　手順 1 で収集された論文と特許から要素技術と効果の対を抽出し，技術動向マップとしてまとめる．

特許マイニングタスクでは，これら 2 つの手順について，以下のサブタスクを設定している．

- 学術論文分類サブタスク：論文抄録に，特許分類体系の 1 つである国際特許分類（International Patent Classification: IPC）コードを自動的に付与するシステムを構築する．
- 技術動向マップ作成サブタスク：要素技術とその効果を示す表現を，論文や特許から自動的に抽出する．

本研究では，NTCIR-8 特許マイニングタスクにおいて実施された技術動向マップ作成サブタスクのデータセットを用いて本研究の提案手法の評価を行う．

なお，技術動向マップ作成サブタスクでは，効果における属性値表現が数値となるものも対象としている．もし，たとえば「形態素解析」や「機械翻訳」などの特定分野の論文や特許から，「93.5%」などのような精度値を抽出できれば，精度値を縦軸に，論文の著者年や特許の出願年を横軸にとった，対象とする分野における精度値の時間的な推移を示すグラフが描画できる．このグラフを用いることで，対象の分野への新規参入を検討している企業に対して，参入する余地があるかどうかの判断材料として利用することができると考えられる．本研究では，特定分野における，抽出された要素技術を縦軸に，各文献の著者年や出願年，およびその要素技術を用いて得られた効果を横軸にとって可視化することにより，その分野における要素技術の変遷や技術動向を効果的に提示することを目指す．

このワークショップにおいて，Nishiyama ら [12] は，日本語論文および日本語特許を対象として FEDA [13] を用いることにより，解析精度が向上することを示している．FEDA とは，元ドメインのデータを併用して，目標ドメインの性能を改善するドメイン適応である．一般に，ドメイン適応では，元ドメインの訓練データによって得られたパラメータを目標ドメインでの学習の指標として用いることで，目標ドメインに適応するようなパラメータ調整を行う．FEDA は，元ドメインの特徴ベクトルと目標ドメインの特徴ベクトルをそれぞれ長さが 3 倍の高次元の特徴ベクトルに変換を行う．そして，変換後の特徴ベクトルを用い

© 2013 Information Processing Society of Japan

て通常の方法で学習を行う．この手法により，従来のドメイン適応手法とほぼ同程度の精度結果を得られることが明らかにされている．本研究でも同様に，ドメイン適応手法である FEDA を用いて有効性を確認する．また，本研究では FEDA とは異なる新たなドメイン適応手法を提案し，FEDA との比較を行う．

4. 論文と特許の表題および概要の構造解析

本章では，論文と特許の表題および概要の構造解析手段について述べる．4.1 節では，表題および構造解析について述べ，4.2 節では，ドメイン適応手法の 1 つである FEDA，および本研究で新たに提案するドメイン適応手法について述べる．

4.1 表題および概要の構造解析

4.1.1 表題および概要構造におけるタグの定義

本研究では，表題および概要の構造解析において，機械学習を用いて構造化を行う．以下に，本研究で使用する構造タグとそのタグを付与する際に使用する手がかり語を示す．

- **TECHNOLOGY**：要素技術を示す（例：“SVM”，“HMM”）．
- **EFFECT**：効果（新しい機能の追加，新しく得られた物質，精度などの数値または増加・減少，問題点の抑制や解決したこと，明らかになったこと）を示す．EFFECT タグは，以下に示す ATTRIBUTE タグと VALUE タグを含む．
- **ATTRIBUTE, VALUE**：たとえば，「処理速度（ATTRIBUTE）が向上（VALUE）」のように「属性（ATTRIBUTE）」と「属性値（VALUE）」の対で表現する．

表題において，本研究では TECHNOLOGY タグのみを用いて解析を行う．これは，論文や特許の表題には，要素技術から得られる効果に関する記述はほとんどされていないからである．以下に，「隠れマルコフモデルを用いた形態素解析に関する研究」という論文表題に上記のタグを付与した例を示す．

<TECHNOLOGY> 隠れマルコフモデル </TECHNOLOGY> を用いた形態素解析に関する研究

概要の解析においては，TECHNOLOGY，EFFECT，ATTRIBUTE，および VALUE タグを用いる．以下に，「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適用するため，電気損失を最小化できる．」という概要に上記のタグを付与した例を示す．

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> を適用するため，<EFFECT><ATTRIBUTE> 電気損失 </ATTRIBUTE> を <VALUE> 最小化 </VALUE> </EFFECT> できる．

4.1.2 構造解析の手段

論文および特許の表題や概要中の「を用いた」や「を具備する」といった表現の直前には要素技術（TECHNOLOGY）を表す用語が出現する．一方で，「が可能になる」や「ができる」の直前には効果を表す用語が出現する可能性が高い．また，「信頼性」や「精度」のように属性（ATTRIBUTE）になりやすい用語や，「向上」や「改善」のように属性値（VALUE）になりやすい用語も存在する．これらの用語をあらかじめリストとしてまとめておき，概要中の各単語がリストに含まれるか否かを機械学習の素性として用いる．ここで，要素技術に関する手がかり語表現には定型的なものが非常に多く，「を用いた」などの表現は，様々な分野の論文や特許に出現する．そのため，要素技術の手がかり語表現は分野依存性が低く，人手での手がかり語の収集も比較的容易であると考えられる．一方で，属性や属性値になりやすい用語を様々な分野の論文や特許を対象に人手で網羅的に収集するのは容易ではない．そこで本研究では，係り受け関係や分野類似度などの統計的な手法を用いて半自動的に手がかり語リストを作成する．

4.1.3 手がかり語リストの作成

以下に，手がかり語リストの作成手順について述べる．

（Step 1）上位下位関係による収集

まず，NTCIR-1 と NTCIR-2 で使用された論文文書集合と，1993 年から 2002 年の 10 年において出版された特許文書集合（合計 255,960 件）から，「A などの効果」や「A 等の特徴」などの表現を含む文を収集し，その後，A に該当する個所から，“改善” や “最適化” など，属性値に関する表現を抽出する．以下に，上位概念が “効果” となる下位概念の表現の一部を示す．

頻度	上位概念が “効果” となる下位概念
4192	向上
2075	防止
1424	低減
1201	提供
651	抑制

その後，属性値になりえない表現を人手で削除し，最終的に，300 の手がかり語からなる属性値リストを作成した．

（Step 2）係り受け関係による収集

（Step 1）で得られた属性値に関する用語と依存関係にある名詞/名詞句は，属性になりやすい．そこで，（Step 1）で用いた文書集合から，“向上する” などの属性値になりうる特定の動詞に対して，“精度（が）” や “効率（を）” など，ガ格やヲ格で係る名詞/名詞句を，属性に関する表現として収

© 2013 Information Processing Society of Japan

集する．係り受け解析器として，本研究では CaboCha [*3] を使用した．以下に，“向上する” に係る名詞/名詞句の例を示す．

頻度	係り受け関係にある名詞/名詞句
12066	信頼性_を
9792	大幅_に
6155	作業性_を
5218	生産性_を
4364	操作性_を

その後，属性になりえない表現を人手で削除し，最終的に，700 の手がかり語からなる属性リストを作成した．

（Step 3）分布類似度による収集

テキストから語の関係を自動抽出する方法として共起語に着目し，テキストの指定した範囲内で共起する語のベクトルで各語を特徴づけ，これらの共起語ベクトルどうしの類似度によって語の類似度を数値化する方法がある [14], [15]．相澤 [16] はこれについて，大規模コーパスを用いて語の類似度計算する際における問題点を調べ，同義語について自動獲得と考察を行った．その結果，広範囲の語と共起する語が類似度計算におけるノイズとなるという前提のもと，提案手法の有効性を確認している．本研究では，（Step 1）と（Step 2）で得られた属性および属性値リストを基に，この大規模コーパスを用いた分布類似度の使用を 1 つの方法として，新たな属性および属性値に関する表現を収集することを目指す．これらの表現を収集する際，あらかじめ，10 年分の特許公開広報約 5 億文に対して CaboCha を用いて構文解析を行い，名詞ごとに共起語ベクトル（各名詞と係り受け関係にある動詞を頻度順にまとめた検索語リスト）を作成する．次に，汎用連想計算エンジン GETA [*4] を用いて，属性または属性値リスト中の用語と類似する語を新たな手がかり語として収集する．それぞれの用語間の類似度計算には SMART [17] を用いた．特許文書集合から属性になりうる新たな語を収集するまでの概念図を**図 3** に示す．この図では，特許文書集合中に記述されている “駆動周波数” と共起する語のベクトルと，属性リスト内にある “信頼性” や “作業性” などの語句と共起する語（属性値）のベクトルとの類似度から，“駆動周波数” を属性になりうる手がかり語として収集している．その後，収集した手がかり語集合に対して閾値を設定し，高頻度で出現した語句のみを新たな手がかり語として追加する．この結果，属性表現に対して 510 語，属性値表現に対して 108 語を新たに収集することができた．なお，この手法で後述のすべてのリストを拡張することも可能であるが，予備実験の結果から，属性・属性値の用語リストの拡張のみにおいて，精度が向上することが確認されている．

*3 http://code.google.com/p/cabocha/
*4 http://geta.ex.nii.ac.jp/geta.html

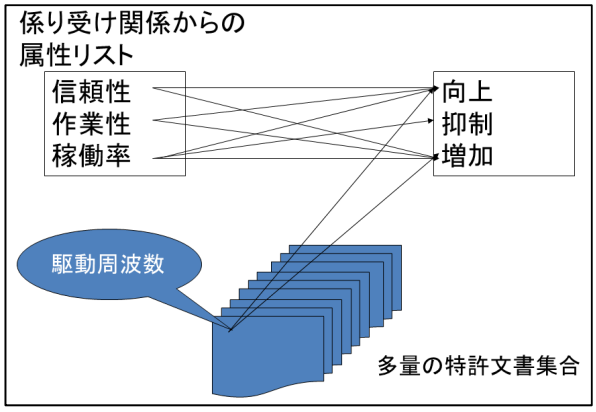


図 3　分布類似度により手がかり語として “駆動周波数” を収集する例

Fig. 3　An example of collecting “driving frequency” as a cue phrase using distributional similarity.

（Step 1）および（Step 3）を用いて属性値に関する語を収集した際，“kg” や “cm” など，単位を表す語はほとんど収集されなかった．そこで，単位を表す語を半自動的に収集し，属性値における手がかり語として用いることを行う．収集方法として，論文文書集合から直前に数値が記述されている単語を収集し，その後，明らかに単位ではない語や出現頻度の少ない語を人手で削除する．この結果，単位になりうる語として 178 語を収集することができた．なお，本研究では日本語論文だけでなく，日本語特許の解析も行う．日本語論文では英字や記号を半角で記述するが，日本語特許では全角で記述する傾向がある．そこで本研究では，日本語特許の解析にも対応させるために，作成した単位語リスト内の全角英字や半角記号に対する，それぞれの全角英字，全角記号を単位語リストに加える．最終的に，274 の手がかり語からなる単位語リストを作成した．

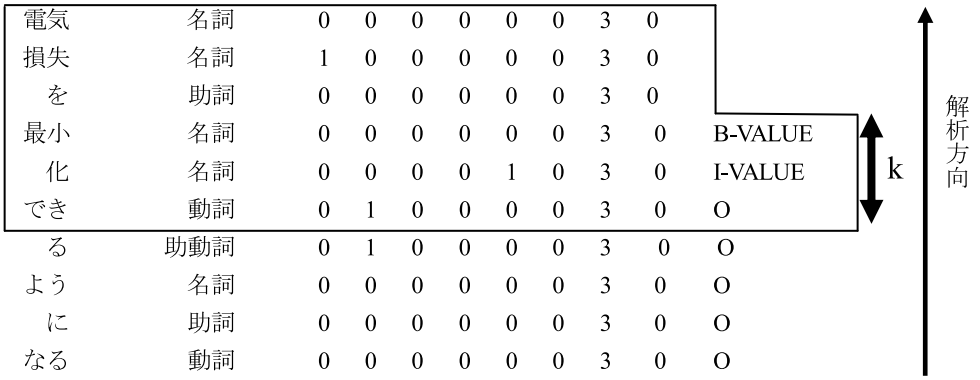
4.1.2 項でも述べたように，多くの論文および特許の表題や概要中において，「を用いた」や「を具備する」などの表現の直前には，要素技術を表す用語が出現する．たとえば，4.1.1 項で示した「隠れマルコフモデルを用いた形態素解析に関する研究」という文の場合，TECHNOLOGY タグが付与されている “隠れマルコフモデル” の直後に，“を用いた” という表現が記述されている．このような，要素技術を示すような手がかり語を機械学習の素性として用いることで，様々な分野における要素技術表現を網羅的に解析できると考えられる．本研究では，要素技術になりうる専門用語（TECHNOLOGY-internal）を人手で収集するとともに，要素技術を示すような手がかり語（TECHNOLOGY-external）を収集する．

また，論文や特許の概要には，主題が記述されている個所がある．このような個所に TECHNOLOGY タグが誤って付与されないように，手がかり語を用いて判定する．たとえば，“提案する” の直前の語句は主題となる場合が多いが，TECHNOLOGY タグが付与されることはない．そこ

© 2013 Information Processing Society of Japan

表 1　概要における機械学習に用いる入力データ
Table 1　Features and tags given to the machine learning for abstract analysis.

各単語	品詞	F1	F2	F3	F4	F5	F6	F7	F8	タグ
電気	名詞	0	0	0	0	0	0	3	0	
損失	名詞	1	0	0	0	0	0	3	0	
を	助詞	0	0	0	0	0	0	3	0	
最小	名詞	0	0	0	0	0	0	3	0	B-VALUE
化	名詞	0	0	0	0	1	0	3	0	I-VALUE
でき	動詞	0	1	0	0	0	0	3	0	O
る	助動詞	0	1	0	0	0	0	3	0	O
よう	名詞	0	0	0	0	0	0	3	0	O
に	助詞	0	0	0	0	0	0	3	0	O
なる	動詞	0	0	0	0	0	0	3	0	O



で，論文や特許の主題となるような手がかり語の有無を素性の 1 つとして用いる．

さらに本研究では，論文と特許の概要における特徴的な記述様式に着目する．論文概要は，前半部に研究目的や提案技術・手法，中間部に要素技術に関する説明，後半部にまとめや効果部に関する説明で構成されている．また，特許においても【発明が解決しようとする課題】【課題を解決するための手段】【発明の効果】という 3 つの項目で構成されている．そこで抽出したある語句が，これらの 3 つの構成部分のうち，どこに属するのかを素性として用いる．

4.1.4　機械学習に用いる素性

表題の構造解析を行う際，機械学習に以下の 1 個の素性を用いる．括弧内の数値は各リストの個数である．

1）TECHNOLOGY-external（45）：要素技術の手がかり語の有無（例：“を用いた”，“に基づいた”）．

また，概要の構造解析を行う際，機械学習に以下の 10 個の素性を用いる．括弧内の数値は各リストの個数である．

1）概要中の各単語
2）品詞情報
3）ATTRIBUTE-internal（1210）：属性の手がかり語の有無（例：“処理量”，“精度”）．
4）EFFECT-external（21）：効果部の手がかり語の有無（例：“できる”，“実現する”）．
5）TECHNOLOGY-external（45）：要素技術の手がかり語の有無（例：“を用いた”，“に基づいた”）．
6）TECHNOLOGY-internal（17）：要素技術専門用語の有無（例：“HMM”，“SVM”）．
7）VALUE-internal（408）：属性値の手がかり語の有無（例：“増加”，“抑止”）．
8）HEAD-exclusion（12）：主題となる不要語または主題の手がかり語の有無（例：“を提案”，“開発”）．
9）Location：概要構造に関する素性．前半部を “1”，中間部を “2”，後半部を “3” で表す．
10）UNIT-internal（274）：数値付き単位の有無（例：“kg”，“cm”）．

4.1.5　機械学習に用いるツールとデータ

論文および特許の構造解析には SVM ベースのチャンキングツールである yamcha[*5]を，また形態素解析には MeCab[*6]を用いる．機械学習で用いる入力データの例を**表 1** に示す．表において，1 列目は概要中の単語を，2 列目は各単語の品詞を示す．3 列目以降から，ATTRIBUTE-internal リスト（F1），EFFECT-external リスト（F2），TECHNOLOGY-external リスト（F3），TECHNOLOGY-internal リスト（F4），VALUE-internal リスト（F5），HEAD-exclusion（F6）リストの語の有無を示している．また，9 列目は Location 素性（F7）を示しており，10 列目は UNIT-internal リストの語（F8）の有無を示している．右端の列は教師用データを示しており，要素技術（TECHNOLOGY）とその効果に関する属性（ATTRIBUTE），および属性値（VALUE）に関する語句は，IOB2 表現 [18] でエンコードする．yamcha は，表 1 の枠で囲まれた個所にタグを付与する場合，窓幅を k とすると，前後 k 行の素性と現在の行の素性，前 k 個のタグを素性として用いる．本研究では，人手で k の値を変更していき，論文，特許それぞれにおいて最も精度が高かった窓幅を採用する．この予備実験の結果から，本研究では，論文および特許表題の構造解析には窓幅 5 を用いる．また，論文の概要解析には窓幅 3 を，特許の概要解析には窓幅 4 を用いる．これは，一般に，特許概要は論文概要と比べ，1 文が長く記述される．ゆえに，特許に付与される構造タグも論文と比べて，長く付与される．その結果，特許の概要解析における，素性として用いる窓幅の範囲を論文の概要解析より広く設定する必要があるため，このように違いが生じたと考えられる．

4.2　ドメイン適応を用いた情報抽出

本研究では，3.2 節で述べたように，Nishiyama ら [12]

*5　http://chasen.org/~taku/software/yamcha/
*6　http://mecab.sourceforge.net/

© 2013 Information Processing Society of Japan

が用いたドメイン適応手法である FEDA を用いて有効性を確認する．使用するコーパスとして，論文用および特許用コーパスを用いる．また，本研究では FEDA に加えて，新たなドメイン適応手法を提案する．

論文ドメインと特許ドメインの両方の素性を用いて情報抽出を行う場合，FEDA ではそれぞれを混ぜあわせて学習に用いたが，このほかに，あるドメインの素性を用いていったん学習させ，タグの付与を行った後，もう一方のドメインの素性を用いて学習させ，タグの付与を行う方法が考えられる．本研究では，以下のような手法を提案する[*7]．

［提案手法 1：SEQ］

(1) 論文ドメインを訓練用データとして用いてモデル A を獲得する．
(2) 特許ドメインを訓練用データとして用いてモデル B を獲得する．
(3) モデル A を用いて対象の論文にタグ付けを行った後，モデル B を用いて先ほどタグ付けされた論文にタグ付けを行う．

また，本研究では次の点を考慮する．4.1.5 項でも述べたように，特許における要素技術に該当する語句は，論文に比べて長く記述される．その結果，特許における TECHNOLOGY タグが付与される長さは論文より長くなる．このように，論文と特許で性質が異なる要素技術を考慮せずに用いた場合，精度が低下する可能性がある．本研究では上記の SEQ 手法に加え，この問題を考慮した手法を提案する[*8]．

［提案手法 2：SEQ（T）］

(1) 論文ドメインを訓練用データとして用いてモデル A を獲得する．
(2) 特許ドメインを訓練用データとして用いてモデル B を獲得する．このとき，訓練用データ内に付与されている TECHNOLOGY タグは除去する．また，機械学習に用いる入力データにおける F3，F4 も使用しない．
(3) モデル A を用いて対象の論文にタグ付けを行った後，モデル B を用いて先ほどタグ付けされた論文にタグ付けを行う．

5. 実験

提案手法の有効性を調べるため実験を行った．5.1 節では実験条件について述べ，5.2 節で実験結果を報告し，5.3 節で考察を行う．

[*7,*8] 対象とする文書が論文である場合を提案手法 1，2 では述べたが，対象とする文書が特許である場合，モデル A が獲得する訓練用データは特許ドメイン，モデル B が獲得する訓練用データは論文ドメインとなる．

表 2　評価用データにおける人手で付与されたタグの数
Table 2　The number of manually assigned tags in test data.

	表題 (TECHNOLOGY)	概要 (TECHNOLOGY)	概要 (ATTRIBUTE)	概要 (VALUE)	概要 (EFFECT)
論文	93	362	296	294	293
特許	9	740	506	474	489

5.1 実験条件

5.1.1 実験データ

本研究では NTCIR-8 特許マイニングタスク [11] のデータを用いて実験を行った．このデータは，1993～2002 年の日本国公開特許公報から任意に選択された 500 件に含まれる 3 つの項目【発明が解決しようとする課題】【課題を解決するための手段】【発明の効果】に TECHNOLOGY，EFFECT，ATTRIBUTE，VALUE タグが人手で付与されている．また，同一のタグが論文 500 件に付与されている．このうち，300 件を訓練用データ，200 件を評価用データとして用いる．また，評価用データにおいて，論文と特許の表題および概要に正解として付与されているタグの数を**表 2** に示す．

5.1.2 評価尺度

評価尺度には，以下に示す再現率と精度および F 値を用いる．

$$再現率 = \frac{提案手法により正しく付与されたタグの数}{正解として付与したタグの総数}$$

$$精度 = \frac{提案手法により正しく付与されたタグの数}{提案手法により付与されたタグの総数}$$

$$F\ 値 = \frac{2 \cdot 再現率 \cdot 精度}{再現率 + 精度}$$

また，表題と概要の TECHNOLOGY タグ，および概要の ATTRIBUTE タグと VALUE タグにおける再現率，精度，および F 値の項目における平均値を AVERAGE とする [11]．

5.1.3 比較手法

本研究では，論文と特許の表題および概要に対して以下の 5 種類の提案手法と，NTCIR-8 特許マイニングタスクで設定されている，技術動向マップ作成サブタスクの formal run に参加した 4 つのシステムの結果をベースラインとし，それぞれの手法と比較を行う．

提案手法

- J-ML（HAND）：4.1.4 項で述べたすべての素性を用いて機械学習（SVM）を行う．ただし ATTRIBUTE-internal，VALUE-internal では，人手で収集した語句（4.1.3 項における Step 1，Step 2）のみを用いる．
- J-ML：4.1.4 項で述べたすべての素性を用いて機械学習（SVM）を行う．
- J-ML+FEDA：4.1.4 項で述べたすべての素性を用いて機械学習（SVM）を行う．概要解析には SVM に加えて，ドメイン適応手法 FEDA を用いる．

© 2013 Information Processing Society of Japan

表 3　論文の表題および概要の解析結果
Table 3　Experimental results for analyzing titles and abstracts of research papers.

提案手法	評価	表題（TECHNOLOGY）	概要（TECHNOLOGY）	概要（ATTRIBUTE）	概要（VALUE）	概要（EFFECT）	AVERAGE
J-ML (HAND)	再現率	0.688	0.169	0.105	0.126	0.072	0.185
	精度	0.800	0.598	0.413	0.425	0.339	0.561
	F 値	0.740	0.263	0.167	0.194	0.118	0.278
J-ML	再現率	0.688	0.169	0.115	0.139	0.075	0.191
	精度	0.800	0.678	0.557	0.603	0.449	0.669
	F 値	0.740	0.270	0.190	0.227	0.129	0.298
J-ML+ FEDA	再現率	0.688	0.152	0.166	0.180	0.109	0.211
	精度	0.800	0.495	0.450	0.510	0.376	0.547
	F 値	0.740	0.233	0.242	0.266	0.169	0.305
J-ML+ SEQ	再現率	0.688	0.138	0.226	0.259	0.130	0.246
	精度	0.800	0.284	0.347	0.432	0.295	0.411
	F 値	0.740	0.186	0.274	0.323	0.180	0.308
J-ML+ SEQ (T)	再現率	0.688	0.169	0.213	0.262	0.130	0.254
	精度	0.800	0.678	0.339	0.433	0.295	0.496
	F 値	0.740	0.270	0.261	0.326	0.180	0.336

表 4　特許の表題および概要の解析結果
Table 4　Experimental results for analyzing titles and abstracts of patents.

提案手法	評価	表題（TECHNOLOGY）	概要（TECHNOLOGY）	概要（ATTRIBUTE）	概要（VALUE）	概要（EFFECT）	AVERAGE
J-ML (HAND)	再現率	0.556	0.435	0.377	0.432	0.280	0.418
	精度	0.455	0.492	0.563	0.677	0.432	0.553
	F 値	0.500	0.462	0.452	0.528	0.340	0.476
J-ML	再現率	0.556	0.427	0.393	0.513	0.276	0.441
	精度	0.455	0.478	0.535	0.643	0.419	0.537
	F 値	0.500	0.451	0.453	0.570	0.333	0.484
J-ML+ FEDA	再現率	0.556	0.418	0.372	0.506	0.262	0.429
	精度	0.455	0.466	0.545	0.676	0.422	0.540
	F 値	0.500	0.440	0.442	0.579	0.323	0.478
J-ML+ SEQ	再現率	0.556	0.416	0.421	0.546	0.227	0.454
	精度	0.455	0.445	0.482	0.581	0.314	0.493
	F 値	0.500	0.430	0.449	0.563	0.264	0.473
J-ML+ SEQ (T)	再現率	0.556	0.422	0.419	0.544	0.227	0.455
	精度	0.455	0.472	0.483	0.586	0.315	0.507
	F 値	0.500	0.445	0.449	0.565	0.264	0.480

- J-ML+SEQ：4.1.4 項で述べたすべての素性を用いて機械学習（SVM）を行う．概要解析には SVM に加えて，4.2 節で述べた提案手法 1 を用いる．
- J-ML+SEQ（T）：4.1.4 項で述べたすべての素性を用いて機械学習（SVM）を行う．概要解析には SVM に加えて，4.2 節で述べた提案手法 2 を用いる．

ベースライン

- TRL7_1&TRL6_2 [12]：表題および概要の解析を，機械学習（CRF [19]）を用いて行う．概要解析には CRF に加え，ドメイン適応手法 FEDA を用いる．素性には，各単語，品詞情報，字種タイプ，単語接頭詞タイプ，単語接尾辞タイプ，特許内のセクション，論文内の相対的位置，各概要に人手で付与された IPC コード，評価的な語句，依存木内における語句間の距離を用いる．
- ONT [20]：表題および概要解析を機械学習（SVM）で行う．機械学習を行う前に，あらかじめ訓練用データを複数のクラスタに分類する．その後，各クラスタに対して SVM を適用する．素性には，各単語，品詞情報，単語の原型，言語解析結果からの意味的ラベルを用いる．
- smlab [21]：表題および概要解析を機械学習（SVM）で行う．SVM に使用する素性には，エントロピーベースのスコアを用いて収集した手がかり語（例：“用い”，“備え”）を用いる．
- HTC_1&HTC1_1 [22]：表題および概要解析を，機械学習（SVM）を用いて行い，3 タプル表現に基づいた語句の抽出を行う．素性には，各単語，品詞情報，特許内の項目の 1 つである【発明の効果】から人手で作成した手がかり語リスト，日本語依存文法の構文解析を用いた修飾関係を用いる．

5.2　実験結果

提案手法における論文と特許解析の評価結果を**表 3** と**表 4** にそれぞれ示す．

まず，論文の解析結果について見ていく．表 3 において，人手で収集した語句のみを素性として用いた J-ML（HAND）手法と，分布類似度を利用して収集した語句を

© 2013 Information Processing Society of Japan

表 5　各ベースラインと比較した場合の論文の実験結果（AVERAGE）
Table 5　Experimental results for research papers when compared with each baseline.

	手法	再現率	精度	F 値
提案手法	J-ML(HAND)	0.185	0.561	0.278
	J-ML	0.191	0.669	0.298
	J-ML+FEDA	0.211	0.547	0.305
	J-ML+SEQ	0.246	0.411	0.308
	J-ML+SEQ(T)	0.254	0.496	0.336
ベースライン	TRL7-1	0.181	0.573	0.275
	ONT	0.114	0.246	0.156
	smlab	0.081	0.354	0.132
	HTC-1	0.100	0.188	0.131

表 6　各ベースラインと比較した場合の特許の実験結果（AVERAGE）
Table 6　Experimental results for patents when compared with each baseline.

	手法	再現率	精度	F 値
提案手法	J-ML(HAND)	0.418	0.553	0.476
	J-ML	0.441	0.537	0.484
	J-ML+FEDA	0.429	0.540	0.478
	J-ML+SEQ	0.454	0.493	0.473
	J-ML+SEQ(T)	0.455	0.507	0.480
ベースライン	TRL6-2	0.437	0.506	0.469
	ONT	0.178	0.271	0.215
	smlab	0.272	0.547	0.363
	HTC-1-1	0.233	0.346	0.278

追加して用いた J-ML 手法と比較すると，AVERAGE において精度，再現率，F 値すべてにおいて J-ML 手法が上回っていることが分かる．この結果から，分布類似度を用いることが有効に機能していることが分かる．次に，J-ML 手法と，ドメイン適応手法を組み込んだ各手法とそれぞれ比較すると，AVERAGE において再現率，F 値が向上していることが分かる．この結果から，ドメイン適応手法を用いることが有効に機能していることが分かる．このうち，本研究で提案した J-ML+SEQ（T）手法が最も有効に機能しており，再現率，F 値それぞれにおいて J-ML 手法から 0.063，0.038 改善されている．一方，特許の解析結果について見ていくと，表 4 において，分布類似度を用いた J-ML 手法が最も有効に機能していることが分かる．しかし，J-ML 手法とドメイン適応手法を組み込んだ各手法の AVERAGE をそれぞれ比較すると，すべての手法において F 値が低下していることが分かる．ただし，これは J-ML 手法の性能がすでに高く，論文を対象とした場合と比べて，ドメイン適応手法を用いても向上の余地があまりなかったためと考えられる．

また，ベースラインと設定した 4 つのシステムとの比較結果を**表 5** と**表 6** にそれぞれ示す．表 5 には論文の解析結果を，表 6 には特許の解析結果を示す．表 3，表 4 と同様に，表題と概要それぞれのタグにおける再現率，精度，

表 7　各ベースラインと比較した場合の論文の実験結果（表題 TECHNOLOGY）
Table 7　Experimental results for research papers when compared with each baseline.

	手法	再現率	精度	F 値
提案手法	J-ML+SEQ(T)	0.688	0.800	0.740
ベースライン	TRL7-1	0.323	0.811	0.462
	ONT	0.280	0.634	0.388
	smlab	0.000	0.000	0.000
	HTC-1-1	0.000	0.000	0.000

表 8　各ベースラインと比較した場合の特許の実験結果（表題 TECHNOLOGY）
Table 8　Experimental results for patents when compared with each baseline.

	手法	再現率	精度	F 値
提案手法	J-ML+SEQ(T)	0.556	0.455	0.500
ベースライン	smlab	0.444	0.190	0.267
	ONT	0.222	0.222	0.222
	TRL6-2	0.000	0.000	0.000
	HTC-1-1	0.000	0.000	0.000

および F 値の平均値（AVERAGE）を示す．表 5，6 それぞれの結果から，本研究で提案したすべての手法は，F 値においてすべてのベースライン手法を上回っていることが分かる．

5.3　考察

5.3.1　分布類似度とドメイン適応手法の有効性

表 3，表 4 において，人手で収集した語句のみを素性として用いた J-ML（HAND）手法と，分布類似度を利用して収集した語句を追加して用いた J-ML 手法，およびドメイン適応手法組み込んだ各手法を比べると，論文と特許両方において AVERAGE の再現率が全体的に向上していることが分かる．

ここで，論文や特許の表題または概要のほとんどには，研究で使用した要素技術と要素技術を用いて得られた効果が記述されており，これらの情報がその研究において主張したい重要な部分となる．ゆえに本研究では，要素技術とその効果に関する情報を漏れなく網羅的に解析する必要がある．以上の結果から，本研究において分布類似度およびドメイン適応手法を用いることは，論文や特許に対して網羅的な解析を行うための重要なアプローチであったといえる．しかし 4.1.3 項でも述べたように，分布類似度を利用した語句の収集は，属性および属性値に対してのみ効果があったため，概要における ATTRIBUTE と VALUE の再現率が向上したものの，概要における TECHNOLOGY の再現率は向上していない．ここで，表題における TECHNOLOGY の再現率，精度，F 値を，各ベースライン手法と比較した

© 2013 Information Processing Society of Japan

場合の結果を**表 7** と**表 8** に示す．なお，表 7 と表 8 において，再現率，精度，F 値が 0.000 と記載されている場合，そのシステムは，表題に対して TECHNOLOGY タグを付与することができなかったことを示している．それぞれの表において，F 値の結果を比較すると，論文，特許ともに各ベースライン手法より高い値を示していることが分かる．また，精度の値を見てみると，特許では各ベースライン手法より高い値を示しており，論文においても比較的高い値を示していることが分かる．これらの結果から，本研究の手法は，論文および特許の表題に記述されている要素技術表現を正しく解析できていることが分かる．実際の論文や特許において，表題中に記述されている手法やアルゴリズムは，その研究において特に主張したい重要な要素技術である．そのため，本研究の表題解析手法を利用して大規模コーパスの表題中に記述されている要素技術表現を抽出するという，分布類似度とは異なる手法を用いて要素技術リストを拡張することで，概要における TECHNOLOGY の再現率が向上すると考えられる．

5.3.2　論文解析におけるドメイン適応手法の比較

表 5 の結果において，J-ML 手法と J-ML+FEDA 手法の再現率を比べると，0.191 から 0.211 と，値が 0.020 向上していることが分かる．一方で，J-ML 手法と J-ML+SEQ 手法の再現率を比較すると，0.191 から 0.246 と，値が 0.055 向上していることが分かる．さらに，J-ML 手法と J-ML+SEQ（T）手法の再現率を比較すると，0.191 から 0.254 と，値が 0.063 向上していることが分かる．これらの結果から，本研究で提案した SEQ と SEQ（T）は，FEDA と比べ，より有効なドメイン適応手法であると考えられる．

次に，J-ML+SEQ 手法と J-ML+SEQ（T）手法の解析結果のうち，どの部分が向上したのか，具体的に調査していく．表 3 における論文の各解析結果を見ていくと，ATTRIBUTE タグと VALUE タグにおける再現率が向上していることが分かる．J-ML+SEQ 手法の場合，ATTRIBUTE タグでは 0.115 から 0.226，VALUE では 0.139 から 0.259 と，約 2 倍の改善が見られた．また，J-ML+SEQ（T）手法においても，ATTRIBUTE タグでは 0.115 から 0.213，VALUE タグでは 0.139 から 0.262 と，約 2 倍の改善が見られ，FEDA を用いた場合と比べ，大幅に改善されたことが分かる．5.3.1 項でも述べたように，本研究では要素技術とその効果に関する情報を漏れなく収集することを目的としているため，概要解析における再現率の向上は，本研究の重要なタスクであるといえる．また，いくつの事例において再現率が向上しているか具体的に調査した．ATTRIBUTE タグ，VALUE タグにおける，各手法を用いたときに正解と判定された件数，改善された件数，および改悪された件数を**表 9** に示す．この結果から，SEQ 手法，SEQ（T）手法を用いたことによって改悪された件数はそれぞれ 2 件以下程度にとどまっていることに対して，改善された件数は

表 9　論文概要解析における各手法を用いた場合の正解件数の比較（ATTRIBUTE，VALUE）

Table 9　Comparison of the number of correct answers when used with our methods for research papers.

	手法	正解件数	J-ML 手法と共通の正解件数	改善数	改悪数
ATTRIBUTE	J-ML	34			
	J-ML+SEQ	67	32	35	2
	J-ML+SEQ(T)	63	32	31	2
VALUE	J-ML	41			
	J-ML+SEQ	76	40	36	1
	J-ML+SEQ(T)	77	41	36	0

J-ML 手法を用いて得られた正解件数の約 2 倍であることが分かる．これらの結果から，本研究で提案した，「モデル A を用いていったんタグ付けを行った後，モデル B を用いてさらにタグ付けをする」というドメイン適応手法は，論文中に記載されている属性・属性値の解析に対してより効果的であり，また，特許コーパスにおける ATTRIBUTE，VALUE に関する素性は論文解析において十分な改善をもたらしたといえる．

さらに，概要中の TECHNOLOGY タグにおける解析結果について考察する．J-ML 手法と J-ML+SEQ（T）手法を比べると，再現率，精度，F 値すべてにおいて変化がなかった．一方，J-ML 手法と J-ML+SEQ 手法と比べたとき，再現率，精度，F 値すべてが低下しており，特に精度が大幅に低下している．精度が低下した原因について調査したところ，モデル B を用いてタグ付けを行ったとき，モデル A を用いてすでにタグ付けが行われた個所に対して，TECHNOLOGY タグが付与されたことが主な原因であることが分かった．実際，論文概要中の「線形分離不可能な 4 ビットパリティチェック問題を用いた動作試験により ...」という文に対して，論文ドメインから獲得したモデル A を用いてタグ付けを行ったとき，以下のように TECHNOLOGY タグが付与された．

> 線形分離不可能な <TECHNOLOGY>4 ビットパリティチェック問題 </TECHNOLOGY> を用いた動作試験により ...

しかし続けて，特許ドメインから獲得したモデル B を用いてタグ付けを行ったとき，上記のタグ付けされた文に対して，以下のようにタグが付与された．

> <TECHNOLOGY> 線形分離不可能な <TECHNOLOGY>4 ビットパリティチェック問題 </TECHNOLOGY></TECHNOLOGY> を用いた動作試験により ...

© 2013 Information Processing Society of Japan

この結果,「線形分離不可能な <TECHNOLOGY>4 ビットパリティチェック問題」が要素技術であると判断され，精度や再現率を低下させる要因となった．一方，J-ML+SEQ（T）手法では，特許ドメイン内の要素技術関連の素性を除去して獲得したモデル B を用いているため,「線形分離不可能な 4 ビットパリティチェック問題」に TECHNOLOGY タグが付与されることなく,「4 ビットパリティチェック問題」が要素技術であると判断されている[*9]．以上の結果より，4.2 節で述べた「SEQ 手法において，論文と特許で性質が異なる要素技術を考慮せずに用いた場合，精度が低下する」という仮定は正しいと判断でき，これを考慮した本研究の J-ML+SEQ（T）手法は妥当であったと考えられる．しかし，特許で記述される要素技術には，一般的な表現で長く記述されているものだけでなく，端的に表現しているものも存在する．そのため，今回の SEQ（T）手法のような，特許中の要素技術関連すべての素性を除去するのではなく，端的に表している要素技術表現の素性を用いて論文の解析を行うことで，さらなる再現率や精度の向上が見込まれる．

5.3.3　特許解析におけるドメイン適応手法の比較

表 4 の結果において，J-ML 手法とそれにドメイン適応手法を組み込んだ各手法の AVERAGE をそれぞれ比較すると，F 値はすべてのドメイン適応手法において低下していることが分かる．しかし，J-ML+SEQ 手法および J-ML+SEQ (T) 手法において，概要における ATTRIBUTE，VALUE タグの再現率は J-ML 手法より向上している．これは，J-ML 手法を用いた場合に発生した解析誤りの要因の 1 つである，ATTRIBUTE と VALUE の出現順による解析誤りを，論文ドメインを用いることで考慮できたからと考えられる．具体的に述べると，J-ML 手法の解析結果において,「高い認識率」という例では,「高い」の個所に VALUE タグが付与され,「認識率」の個所に ATTRIBUTE タグが付与されるべきであるが，いずれのタグも付与されていなかった．これは，本研究で用いた特許用の訓練用データに「精度が高い」のような ATTRIBUTE，VALUE となる表現の並びが多かったため，VALUE，ATTRIBUTE の順番で単語が出現した場合,「高い」より前の語に ATTRIBUTE が存在しないか，もしくは「認識率」の後ろの語に VALUE がないかと判断し，タグの付与ができなかったからと考えられる．一方で，論文における効果表現には,「少ない計算量」や「10 倍の速度性能」などのように，VALUE，ATTRIBUTE の並び順で記述される場合が少なくない．実際，論文用の訓練用データを見ると，VALUE，ATTRIBUTE の順でタグが付与されていたものが多く存在した．このことから，論文を訓練用データとして用いて，すでにタグ付けされた特許文書をさらに解析することによって，特許モデルを用いただけでは十分に解析できなかった VALUE，ATTRIBUTE の並び順を補って解析することができたと考えられる．

しかしながら，J-ML 手法と J-ML+SEQ 手法および J-ML+SEQ（T）手法の EFFECT タグにおける解析結果において，再現率と精度は低下している．これは上記で述べた，すでにタグ付けされた個所に対してさらにタグ付けされたことが主な原因であることが考えられる．たとえば,「磁気抵抗効果を有しかつバイアス磁界が付与されると共に該磁気抵抗効果を用いて情報を再生するための再生素子」という文に対して，まず特許ドメインから獲得したモデル A を用いて解析したとき，以下のように TECHNOLOGY タグが付与された．

> <TECHNOLOGY> 磁気抵抗効果を有しかつバイアス磁界が付与されると共に該磁気抵抗効果を用いて情報を再生するための再生素子 </TECHNOLOGY>

しかし続けて，論文ドメインを用いて獲得したモデル B を用いてさらに解析をしたとき，以下のように ATTIBUTE タグと VALUE タグ，および EFFECT タグが付与された．

> <EFFECT><ATTRIBUTE><TECHNOLOGY> 磁気 </ATTRIBUTE> 抵抗 <VALUE> 効果 </VALUE> </EFFECT> を有しかつバイアス磁界が付与されると共に該磁気抵抗効果を用いて情報を再生するための再生素子 </TECHNOLOGY>

この問題は，すでにタグが付与された個所の周り対して，新たなタグの付与は行わないようにするといった処理を加えることで解決すると考えられる．

6.　おわりに

本研究では，特定分野の論文と特許から，要素技術とその効果を示す表現を，機械学習を用いて自動的に抽出し，論文と特許を「要素技術」と「効果」という 2 つの観点で分類する手法を提案した．機械学習に用いる素性として，本研究では単語や品詞に加えて，要素技術，属性，属性値の手がかり語表現の有無を使用した．そして，様々な分野における手がかり語表現を網羅的に収集するために，係り受け関係や上位下位関係による人手での収集，さらに分布類似度を用いて自動的に収集した．さらに本研究では，論文または特許の解析を行う際に，ドメイン適応手法を用いることでさらなる解析精度の向上を試みた．その結果，論文の解析において，本研究で提案した「あるドメインの素性を用いてモデルを獲得し解析を行った後，要素技術関連の素性を除いたもう一方のドメインの素性を用いてモデルを獲得し，さらに解析を行う」というドメイン適応手法が最も有効に機能し，再現率，精度，F 値による評価でそれ

[*9] 実際，正解データでは「4 ビットパリティチェック問題」の個所に TECHNOLOGY タグが付与されることは正しいとされている．

© 2013 Information Processing Society of Japan

ぞれ，0.254，0.496，0.336 の値が得られた．一方，特許の解析では，機械学習のみを用いた手法が最も有効であり，再現率，精度，F 値による評価でそれぞれ，0.441，0.537，0.484 の値が得られた．これらの結果は，NTCIR-8 特許マイニングタスクにおける技術動向マップ作成サブタスクの formal run において提示されたシステムの結果よりも優れており，提案手法の有効性が確認された．

参考文献

[1] Matsumura, A., Takasu, A. and Adachi, J.: Structured Index System at NTCIR1: Information Retrieval using Dependency Relationship between Words, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.117–122 (1999).

[2] 今井　俊：表題解析による科学技術論文の自動分類，北陸先端科学技術大学院大学修士論文 (1999).

[3] 村田真樹，一井康二，馬　青，白土　保，金丸敏幸，井佐原均：過去 10 年間の言語処理学会論文誌・年次大会における研究動向調査，言語処理学会 11 回年次大会 (2005).

[4] 難波英嗣，谷口裕子：学術論文データベースからの研究動向情報の抽出と可視化，言語処理学会第 12 回年次大会併設ワークショップ「言語処理と情報可視化の接点」，pp.35–38 (2006).

[5] 西山莉紗，竹内広宣，渡辺日出雄，那須川哲哉：新技術が持つ特長に注目した技術調査支援ツール，人工知能学会論文誌，Vol.24, No.6, pp.541–548 (2009).

[6] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[7] Blei, D.M. and Lafferty, J.D.: A Correlated Topic Model of Science, *The Annals of Applied Statistics*, Vol.1, No.1, pp.17–35 (2007).

[8] Hall, D., Jurafsky, D. and Manning, C.D.: Studying the History of Ideas Using Topic Models, *Proc. EMNLP 2008*, pp.363–371 (2008).

[9] He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P. and Giles, C.L.: Detecting Topic Evolution in Scientific Literature: How Can Citations Help?, *Proc. 18th ACM Conference on Information and Knowledge Management* (*CIKM'09*), pp.957–966 (2009).

[10] Bolelli, L., Ertekin, S., Zhou, D. and Giles, C.L.: Finding Topics Trends in Digital Libraries, *Proc. Joint Conference on Digital Libraries* (*JCDL'09*), pp.69–72 (2009).

[11] Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-8 Workshop, *Proc. 8th NTCIR Workshop Meeting*, pp.293–302 (2010).

[12] Nishiyama, R., Tsuboi, Y., Unno, Y. and Takeuchi, H.: Feature-Rich Information Extraction for the Technical Trend-Map Creation, *Proc. 8th NTCIR Workshop Meeting*, pp.318–324 (2010).

[13] Daumé III, H.: Frustratingly Easy Domain Adaptation, *Proc. 45th Annual Meeting of the Association for Computational Linguistics*, pp.256–263 (2007).

[14] Lee, L.: Measures of Distributional Similarity, *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp.25–32 (1999).

[15] Lin, D.: Automatic Retrieval and Clustering of Similar Words, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp.768–774 (1998).

[16] 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).

[17] Salton, G.: *The SMART Retrieval System – Experiments in Automatic Document Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ (1971).

[18] Tjong Kim Sang, E.J. and Veenstra, J.: Representing Text Chunks, *Proc. 9th Conference on European Chapter of the Association for Computational Linguistics*, pp.173–179 (1999).

[19] Peng, F. and McCallum, A.: Accurate Information Extraction from Research Papers using Conditional Random Fields, *Proc. HLT-NAACL*, pp.329–336 (2004).

[20] Mizuguchi, H. and Kusui, D.: An Information Extraction Method for Multiple Data Sources, *Proc. 8th NTCIR Workshop Meeting*, pp.348–353 (2010).

[21] Suzuki, Y., Nonaka, H., Sakaji, H., Kobayashi, A., Sakai, H. and Masuyama, S.: NTCIR-8 Patent Mining Task at Toyobashi University of Technology, *Proc. 8th NTCIR Workshop Meeting*, pp.364–369 (2010).

[22] Sato, Y. and Iwayama, M.: Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi, *Proc. 8th NTCIR Workshop Meeting*, pp.359–363 (2010).

福田 悟志

2011 年広島市立大学情報科学部知能情報科卒業．現在，同大学大学院情報科学研究科博士前期課程在学中．

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 年同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科助手．同年広島市立大学情報科学部講師．2010 年広島市立大学大学院情報科学研究科准教授．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．

© 2013 Information Processing Society of Japan

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 年同大学大学院理工学研究科博士後期課程修了．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．知能工学専攻言語音声メディア工学研究室に所属．現在に至る．工学博士．音声対話翻訳の研究開発に従事．2006 年電子情報通信学会 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．

（担当編集委員　宇田川 佳久）

© 2013 Information Processing Society of Japan