広島市立大学審査博士学位論文

技術文書を統合した
動向分析システムの自動構築

2016 年 9 月

福田 悟志

要旨

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．本論文では，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自動的に抽出する手法を提案する．また，要素技術とその効果を用いた応用例を示す．

論文や特許からの技術動向の抽出にあたって，本論文では，特定分野において使用された基礎的な技術(要素技術)とそれを用いて得られた知見(効果)に着目する．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，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ことを試みた．

一般に，技術文書において，「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．また，「が可能になる」や「ができる」の直前には，効果を表す表現が出現する可能性が高い．さらに，効果表現は，「精度(属性)が向上(属性値)」など，属性や属性値になりやすい用語で構成されている場合が多い．本論文では，このような手掛かり語を用いた機械学習による自動解析手法を提案した．さらに，機械学習において，単独のドメインコーパスから得られる情報量には限界があることに着目し，ドメイン適応による学習量の増加を図り，さらなる解析精度の向上を試みた．そして，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，論文解析では再現率 0.254，精度 0.469 および特許解析では再現率 0.441，精度 0.537 が得られた．これらの結果は，特許マイニングタスクの formal run で提示されたシステムの結果よりも優れていることが分かった．

要素技術と効果に関する情報を利用したシステムとして，特定の分類体系に基づく研究領域全般を横断した学術論文の自動分類システムを構築した．すべての研究領域を網羅した自動分類は，網羅的かつ効率的な文書検索を実現する．このようなタスクは文書分類として扱われており，機械学習に基づいた手法が多く提案されている．本論文では，機械学習に用いる手掛かり語として，要素技術と効果に関する表現を用いた．これらの情報は，特定分野の特徴を表す重要な手掛かりになる．なお，使用する分類体系として，すべての研究分野を網羅して構築された科学研究助成事業データベースの分類体系を用いた．そして，この分類体系における「分野・分科・細目表」を論文の分類対象としたとき，実験からそれぞれ平均 0.853，0.712，0.615 の分類精度が得られた．これらの値は，要素技術とその効果に関する表現を用いない場合より高いことから，文書分類における手掛かり語として要素技術とその効果を用いることの有効性が示された．

論文や特許中に含まれる要素技術とその効果を利用したもう一つのシステムとして，技術動向分析システムを構築した．ある技術分野において，「どのような要素技術がいつ頃から使われており，どのような効果が得られているのか」という情報を網羅的に収集し整理することは，その分野の技術動向を概観するのに必要不可欠である．しかし，このような動向調査には多大な時間と労力を要する．そこで，技術文書から技術動向に関する情報を自動的に抽出し，可視化するための方法を提案した．まず，特定のキーワードにより検索された論文と特許から要素技術と効果を抽出する．次に，左から順番に，要素技術，論文または特許の著作年，効果に関する情報を列挙する．これにより，特定のキーワードを中心とした要素技術の変遷を提示することができる．そして，膨大な関連論文や特許を人手で分析することなく，技術の将来性や研究の方向性を効率的に判断できることを示した．

本論文では，日本語論文と特許を対象としたが，今後は，英語や中国語など様々な言語の論文および特許も対象とすることで，海外との技術動向を比較したシステムの構築などを検討している．また，論文や特許のような技術文書だけでなく，ニュース記事や SNS といったメディア情報や評価報告書，決算短信などの文書も分析対象とし，技術的側面と経済的・社会的側面の両方から技術を評価する動向分析システムの構築も検討している．

キーワード:　情報抽出，ドメイン適応，文書分類，技術動向分析

Automatic Construction of a Technical Trend Analysis System that Integrates Technical Documents

Satoshi Fukuda

Abstract

Retrieving and analyzing existing research papers and patents has become an important aspect of assessing the scope of fields with high industrial relevance. We propose a method that automatically extracts information about technical trends from both research papers and patents and demonstrate its applications in the fields.

For the extraction of the technical trends from both research papers and patents, we focus on the elemental or 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 that field. Therefore, we constructed a method that recognizes phrases that represent elemental technologies and their effects in any research field.

In general, phrases beginning with the words, "を用いた (using)," or, "を具備する (equipped)," tend to represent elemental technology. Similarly, phrases beginning with words such as, "が可能になる (be able to)," or, "できる (possible)," tend to represent effect. Moreover, phrases representing effect, such as, "精度(属性)が向上(属性値) (increase (attribute) precision (value))," tend to be constructed using particular words that represent an attribute or value. Based on this idea, we propose a cross domain machine learning method using multiple corpuses to train the system for these cue phrases. 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 and recall and precision scores of 0.441 and 0.537, respectively, for the analysis of patents. These results are an improvement over the system submitted in Patent Mining Task.

As an example of the application of our proposed method, we constructed a system for the automatic classification of research papers across all research areas based on a particular classification index. Automatic classification that covers all research areas promises more comprehensive and efficient document retrieval. Such a challenge is treated as document classification, and a machine learning method is proposed as the

typical approach. We used elemental technologies and their effects as cue phrases for the machine learning method. These cue phrases are useful for characterizing research fields. To investigate the effectiveness of our method, we conducted an experiment that covers all research fields using the KAKEN classification index. From the results, we obtained average recall scores of 0.6220, 0.7205, and 0.8530, respectively, for the Research Field, Discipline, and Area levels in KAKEN classification index. These scores are with a significant improvement over the baseline method, which does not use elemental technologies and their effects as features. As a result, we confirmed the effectiveness of our document classification method using elemental technologies and their effects.

As another example of using elemental technologies and their effects, we constructed a technical trend analysis system for research papers and patents. The application of the elemental technologies used in a particular research field and the information regarding what kind of effect is obtained by using the technology are essential for analyzing technical trends in the field, but it is costly and time consuming to read all of the related technical documents. Therefore, we propose a method that automatically visualizes information about technical trends. First, elemental technologies and their effects are extracted automatically from research papers and patents retrieved by the particular keyword. Then, elemental technology, copyright year of the document, and effect are displayed from left to right. As a result, we can efficiently know the history of the elemental technologies and their effects in a particular field. Moreover, we can effectively assess the future of the technology and the direction of research without manually analyzing all related papers and patents.

In this work, we focused on research papers and patents written in Japanese. Our future work will focus on research papers and patents written in English and Chinese, and we will consider the construction of a system comparing the technical trends between countries. In addition to technical documents such as research papers and patents, we will consider other documents such as newspapers, SNS, evaluation reports, and financial statements. We will also construct a technical trend system that evaluates a technological aspect from both technical and social/economic viewpoints.

Keyword: information extraction, domain adaptation, document classification, technical trend analysis

目次

図目次

表目次

第1章　序論

1.1 研究の背景

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．2016 年 5 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 2016」[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ると考えられる．

論文と特許を検索するのは，大学研究者に限った話ではない．例えば，特許庁の審査官は，出願された技術や発明が特許権の取得に該当するかどうかを判断するために，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，競合する他社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，論文と特許が検索および分析の対象となる．

しかしながら，限られた時間で特定の分野で発表された論文や特許を網羅的に収集し分析することは容易ではない．文部科学省が発表した「平成 22 年版 科学技術白書」の「論文成果に見る我が国の状況」[2]という解説記事では，SCOPUS[3]と Web of Science[4]における日本で発表された論文数は，2004 年から 2006 年の 3 年間において平均 157,412 件とされている．また，特許では，JPO (JAPAN PATENT OFFICE)で発表された日本での特許出願件数[5]によると，2013 年において 328,436 件と述べられている．これらに加えて，海外で発表された論文や特許[6]も対象とするならば，検索や分析に膨大な時間が必要となるといえる．こうした状況を鑑み，本論文では，論文と特許を対象にした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，および検索支援システムの構築を目指す．

---

[1] http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku20160509.pdf
[2] http://www.mext.go.jp/b_menu/hakusho/html/hpaa201001/detail/1296363.htm
[3] https://www.scopus.com/
[4] http://ip-science.thomsonreuters.jp/products/web-of-science/
[5] https://www.jpo.go.jp/shiryou/toushin/nenji/nenpou2014/honpen/1-1.pdf
[6] 「平成 22 年版 科学技術白書」で発表された米国での論文件数は平均 555,941 件，JPO で発表された 2013 年の PCT (Patent Cooperation Treaty)国際出願件数は 43,075 件とされている．

システムの構築にあたって，本論文では，特定の分野で使用された基礎的な要素技術とその効果に着目する．論文および特許中には，研究課題に対して提案された新しい技術や既存技術を応用した技術，研究課題を解決するための手段などが記述されている．また，これらの技術や手段などを用いて得られた知見を，研究課題に対する成果として述べている場合が多い．そして，特定の研究課題において有用とされている技術や手段が確認されたとき，それらは同一あるいは近い分野の他の研究課題にも利用されることも少なくない．すなわち，論文および特許中で述べられている要素技術とその効果に関する表現は，その分野の特徴を表す重要な手掛かりとなり，技術動向のあらましを把握する重要な情報といえる．

1.2 研究の目的

本論文では，論文と特許から要素技術とその効果を表す表現を自動的に抽出する手法を提案する．また，要素技術とその効果に関する表現が持つ性質を利用した文書分類手法を提案し，検索支援システムの構築を試みる．さらに，抽出した情報を用いた技術動向分析システムを構築することを目指す．

1.2.1 要素技術とその効果の抽出

これまでにも，構造解析による文の意味的な役割の解析や人手で作成したルールに対応付けた解析により，論文や特許から有用な情報を抽出する研究は多く存在する．しかし，著者による表現の違いや論文と特許間の記述スタイルの違いという面から，文章をルールに対応付けて解析することは難しい．本論文ではこの問題を「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」や「発明の詳細な説明」において，「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．また，「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．さらに，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を表す「属性」，および属性に付随する値を示す「属性値」が含まれていることが多い．例えば，「SVM を用いることで精度の向上が可能になった」という文の場合，「SVM」が要素技術，「精度の向上」が効果となる．また，「精度の向上」において，「精度」と「向上」がそ

れぞれ属性，属性値を示す.「を用いた」や「が可能になる」といった手掛かり語の表記は，著者や研究分野によって大幅に異なることはないと考えられる．そこで，要素技術または効果を示す手掛かり語を収集し，これらの手掛かり語の有無を素性とした機械学習による統計的な解析手法により，要素技術とその効果の抽出を行う．

また，効果表現の抽出において，「精度」や「向上」のように，属性または属性値になりやすい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．しかしこれらの用語は，著者や分野によって同じ意味でも表現が異なる傾向にある．例えば，「精度」という属性表現は，著者や分野によって「性能」，「分類精度」または「分類性能」という表現で記述される場合がある．同様に，属性値表現でも，「向上」に対して「改善」あるいは「大幅改善」と記述されることもある．そのため，属性や属性値になりやすい用語を人手で網羅的に収集するのは容易ではない．しかし，これらの用語を収集してリスト化しておけば，属性または属性値の有無を素性とした機械学習による識別を行うことができる．そこで本論文では，係り受け関係や上位下位関係，さらに分布類似度を用いた網羅的な属性および属性値表現の収集手法を提案する．

さらに本論文では，ドメイン適応の考え方を取り入れた情報抽出性能のさらなる向上を試みる．機械学習における論文/特許用抽出器を構築するとき，論文用または特許用タグ付きコーパスをそれぞれ単独で学習に用いることが一般的であるが，単独のドメインコーパスから得られる学習量には限界がある．しかし上記でも述べたように，要素技術とその効果を示す手掛かり語表現や，属性や属性値になりやすい用語に対する論文と特許間での違いは無いと考えられる．そのため，論文用および特許用コーパスの両方を使用することで，学習量を増加させることができ，抽出性能の向上が期待できる．

1.2.2　学術論文の自動分類

特定の技術について記述された文献を検索し収集する場合，例えば，特許には，国際特許分類(Internatinal Patent Classification: IPC)と呼ばれる，ほぼすべての技術領域を対象とした分類体系が考案されており，IPC コードを利用した検索支援が整備されている．一方で，一部の論文データベースでは，特定の研究領域を対象とした分類体系が考案されているものの，研究領域全般を横断した論文の分類は，すべての研究領域における研究者に対する論文の検索支援を目指す場合において重要な課題といえる．

文書分類は，自然言語処理などのデータ解析の分野における代表的な研究課題の一つであり，機械学習による分類手法が多く考案されている．本論文では，学術論文固有の特徴である要素技術とその効果を用いた機械学習手法を提案する．技術動向を示す要素技術と

それらにより得られた効果に関する情報は，論文を分類する分野を決定するための重要な手掛かりになると考えられる．

論文の分類に用いる分類体系として，本論文では，科学研究費助成事業データベース(KAKEN)[7]の分類体系を用いる．KAKEN とは，国立情報学研究所が文部科学省，日本学術振興会と協力して作成・公開しているデータベースであり，過去に採択された 67 万件以上の研究課題を検索することができる．この分類体系は，理工系，人文社会系，生物系といったほぼすべての研究分野を網羅しており，研究領域によって「系・分野・分科・細目表」という 4 階層から構成されている．このように，KAKEN の分類体系は，すべての研究分野の領域を横断的した学術論文の分類に適している．

1.2.3　技術動向分析システムの構築

近年では，文書内容を解析し，観点別に動向を分析する取り組みが多く行われている．特許庁が公開している「技術分野別特許マップ 活用ガイドマップ」[8]においても，「方法・手段－目的・効果」や「技術開発課題－解決手段」といった，文書の内容を理解しなければ得られないような情報を軸とした可視化技術の開発が求められている．このような観点による可視化技術が実現されれば，課題解決に必要な技術の把握や目標達成への難易度などを見積もることができる．同様に，論文を対象とした観点別による技術動向マップの作成は，特定の分野における技術の将来性や研究の方向性を判断する場合などに有用である．

しかし実際，「技術分野別特許マップ 活用ガイドマップ」によると，技術動向マップの作成は，特許情報解析のエキスパートと各技術の専門家による共同作業により行われており，各技術に対して 2～3 万件の特許文書を対象としていると述べられている．このように，技術動向マップの作成には人手による多大なコストを要している．そのため，技術動向マップの自動作成は，分析作業の効率向上には欠かせない最も重要な課題である．

本論文では，論文および特許から抽出した要素技術とその効果を観点とした技術動向分析システムを構築する．ある技術分野において，「どのような要素技術がいつ頃から使われており，どのような効果が得られているのか」という情報を網羅的に収集し整理することは，その分野の技術動向を概観するのに必要不可欠である．本システムでは，特定のキーワードにより収集した論文と特許から抽出した要素技術を縦軸に，各文書の著作年や出願年を横軸に取り，要素技術を用いて得られた効果もあわせて提示する．これにより，特定のキーワードを中心とした要素技術とその効果の変遷を知ることができると考えられる．

---

[7] http://kaken.nii.ac.jp/

[8] https://www.jpo.go.jp/shiryou/s_sonota/pdf/map_guide/map_guide.pdf

1.3 論文の構成

2 章では，文書からの意味的な表現の抽出に対する諸研究をサーベイし，本手法の方針を述べる．3 章では，論文と特許から要素技術とその効果を抽出する手法を説明する．この章では，機械学習に用いる手掛かり語の利用方法，属性または属性値として利用される用語の自動収集方法，ドメイン適用手法について述べる．そして，提案手法の有効性を調べるための実験を行う．

4 章では，要素技術とその効果に関する表現を利用した文書分類手法，および提案手法を利用した検索支援システムについて述べる．5 章では，既存の動向分析に関するシステムや研究をサーベイし，本論文で構築した要素技術とその効果を観点とした技術動向分析システムについて述べる．最後に 6 章では，結論と今後の課題について述べる．

第2章　関連研究

2.1 論文と特許の構造

本章では，まず論文と特許の構造について述べる．図 2.1 に，一般的な論文の概要例を，図 2.2 に，特許明細書の一部の例を示す．

図 2.1 と図 2.2 を比較すると，特許明細書では，【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】という項目が設けられており，構造化されていることが分かる[9]．これは，特許明細書では，発明に関する背景や従来技術の問題点，発明によってどのように解決するのかといった説明を詳細に記述するため，読解に時間を要する．そのため，あらかじめいくつかの項目を設定し，執筆者に記載させることで，特許を読む人に対して，発明の概観の効率的な理解を促すことができる．

一方で，論文概要には，特許明細書のような項目は設けられていない．しかし，論文概要は主に，「研究背景」「研究目的」「手法」「結果」などで構成されている場合が多い．例えば，図 2.1 の論文概要を上記で述べた型式に当てはめたとき，1 文目が「研究目的」と「手法」，2～4 文目は「研究背景」，5 文目は「手法」，6 文目は「結果」を述べているといえる．このように，一定の意味的な型式に従って論文概要を構造化させたとき，研究の意義や重要性，論文で扱う問題点，問題を解決するための手段，研究で得られた知見を効率的に読み取ることができる．2.2 節で，論文概要を構造化する研究や取り組みについて述べる．

論文や特許の構造化は，技術内容の効率的な理解を促進する．一方で，論文や特許から，技術内容を表す特徴的な情報を文よりも短い単位で抽出することで，最小限の情報で効果的にその文書を理解することができる．例えば，図 2.1 は，「テキスト分類問題に対して，SVM を用いることで高い分類精度を実現する」という内容の論文概要であるが，この内容は，概要における「SVM がテキスト分類問題に対して有効」，および「高い分類精度を実現する」という情報を抽出することで読み取ることができる．このような，文章から必要な情報を取得するというタスクは，一般的に情報抽出と呼ばれており，様々な研究が行われている．2.3 節で詳細を述べる．

---

[9] 実際の特許明細書では，【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】の他に，【発明の名称】【技術分野】【背景技術】【発明を実施するための最良の形態】（【実施例】）【産業上の利用可能性】【図面の簡単な説明】【符号の説明】という項目が設定されている．

概要

本稿では Support Vector Machine (SVM)を用いたテキスト分類法を提案する．テキスト分類問題に対して学習手法を適用する場合，出現頻度の小さい単語まで考慮して学習を行なわないと，分類精度が落ちることが知られている．このため高い分類精度を実現するためには，高次元の単語ベクトルを用いなければならないが，過学習により分類精度が落ちてしまう危険性が生じる．SVM は Kernel 関数により非線形学習も可能であり，高次元の入力ベクトルを用いても過学習なしに最適解が得られる．SVM をテキスト分類に適用し，1. 異なる次元の単語ベクトル，2. 異なる Kernel 関数，3. 異なる目的関数，の 3 点について比較実験を行なった．その結果，SVM がテキスト分類問題に対して有効であることが確認された．

図 2.1 論文の概要例

【発明が解決しようとする課題】しかしながら，前述の燃料噴射ノズルでは，噴射圧力，噴射量，噴射期間等が，燃料噴射ノズルへ燃料を送油する噴射ポンプによって決定されてしまう構造であり，噴孔の数が固定されており，噴射に有効な噴孔の総面積を増減できない構造となっている．このため，エンジン低速回転時には噴射圧が低くなってしまったり，エンジン低負荷時には，噴射時間が短くなってしまったりするなど，良好な燃焼状態を継続させることが困難になるという問題点があった．

【課題を解決するための手段】しかして，この発明にかかる燃料噴射ノズルは，加圧燃料を噴射する噴孔がノズルボディの先端部に形成され，前記ノズルボディの周囲に摺動自在に回動するカバー部材を設け，このカバー部材を回動させることにより噴射に有効な前記噴孔の総面積を可変させる噴孔面積可変機構を備えたことにある(請求項１)．この噴孔面積可変機構は，例えば，カバー部材をコントロールユニットからの信号で回転するマイクロモータによって回動させる機構が考えられる．

【発明の効果】以上述べたように，請求項１乃至３にかかる発明によれば，噴射に有効な噴孔の総面積を噴孔面積可変機構によってカバー部材を回動させることにより調節することができるので，エンジンの負荷と回転数とに対応した噴射圧力，噴射期間，噴射量を得ることが可能となり，ＮＯｘの低減や燃費の向上を図ることができる．

図 2.2 特許明細書の記述例

2.2 構造解析技術

技術文書，特に論文概要の構造化に対して，人手で構造タグを付与することでタグ付きコーパスを構築する取り組みや特定の構造に従った概要執筆を推奨する取り組み，さらに論文概要を自動的に解析して構造タグを付与する研究が行われている．以下では，各取り組みや研究，および従来の構造解析技術について述べる．

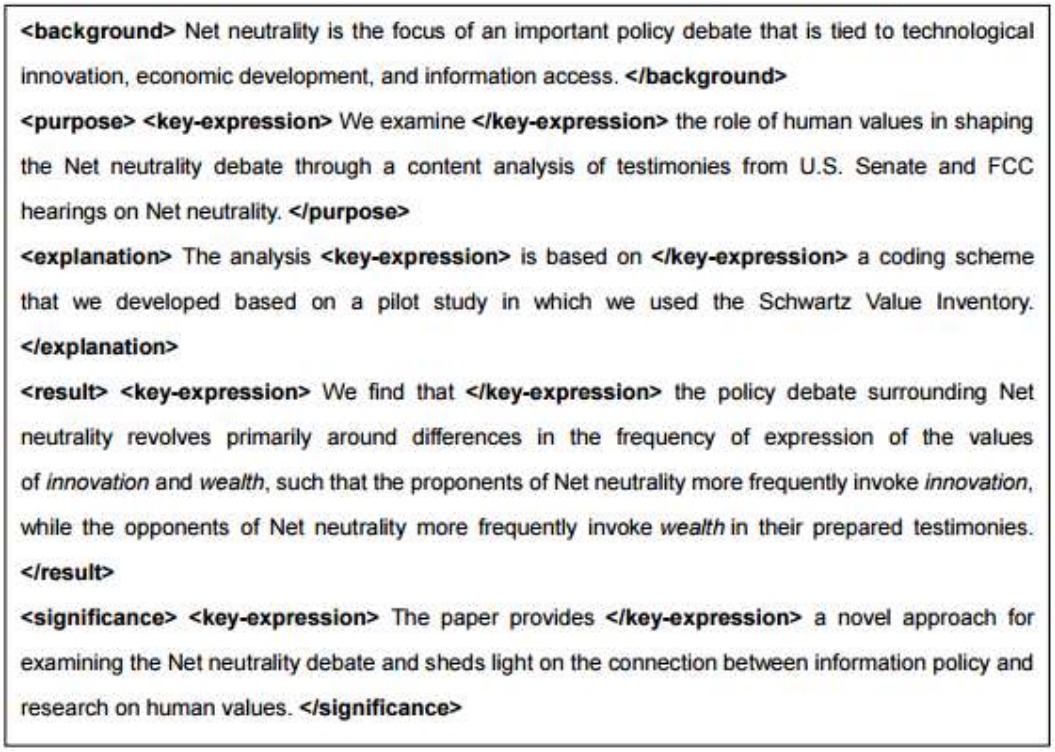
2.2.1 タグ付きコーパスの構築に関する研究・取り組み

人手で構造タグを付与し，タグ付きコーパスを構築するという取り組みは，論文概要に対する構造解析技術への研究や，構造を解析した上での論文の組織化に関する研究などに対する学習/評価データの提供や支援を目的として行われている[1, 2]．また，米国医学会雑誌(The Journal of the American Medical Association; JAMA)や日本の診療ガイドラインでは，論文に記載された臨床試験の内容や問題点などを効率的に読み取ることを目的として,「BACKGROUND」,「OBJECTIVE」,「METHODS」,「RESULTS」,「CONCLUSION」といった概要に記載する項目を設け，記述させるという取り組みが行われている．このような構造化された概要は，タグ付きコーパスとして構築することができる．表 2.1 に，上記で述べたタグ付きコーパスの整備に関する取り組みについてまとめる．また，図 2.3 に，富浦らが構築したタグ付きコーパスの例を示す．図 2.3 において，文の役割を推定した根拠となるキー表現(付与した役割の文に特徴的に出現する表現)を key-expression タグで示している．図 2.4 に，医学・生物学分野の学術文検索サービス PubMed[10]に収録されている構造化概要の例を示す．PubMed は，米国国立医学図書館[11] (National Library of Medicine)における国立生物科学情報センター[12](National Center for Biotechnology Information)が作成しているデータベースであり，世界の主要医学系雑誌に掲載された文献が 300 万件以上収録されている．このうち，約 30%程度の概要が構造化されている．

[10] http://www.ncbi.nlm.nih.gov/pubmed
[11] https://www.nlm.nih.gov/
[12] http://www.ncbi.nlm.nih.gov/

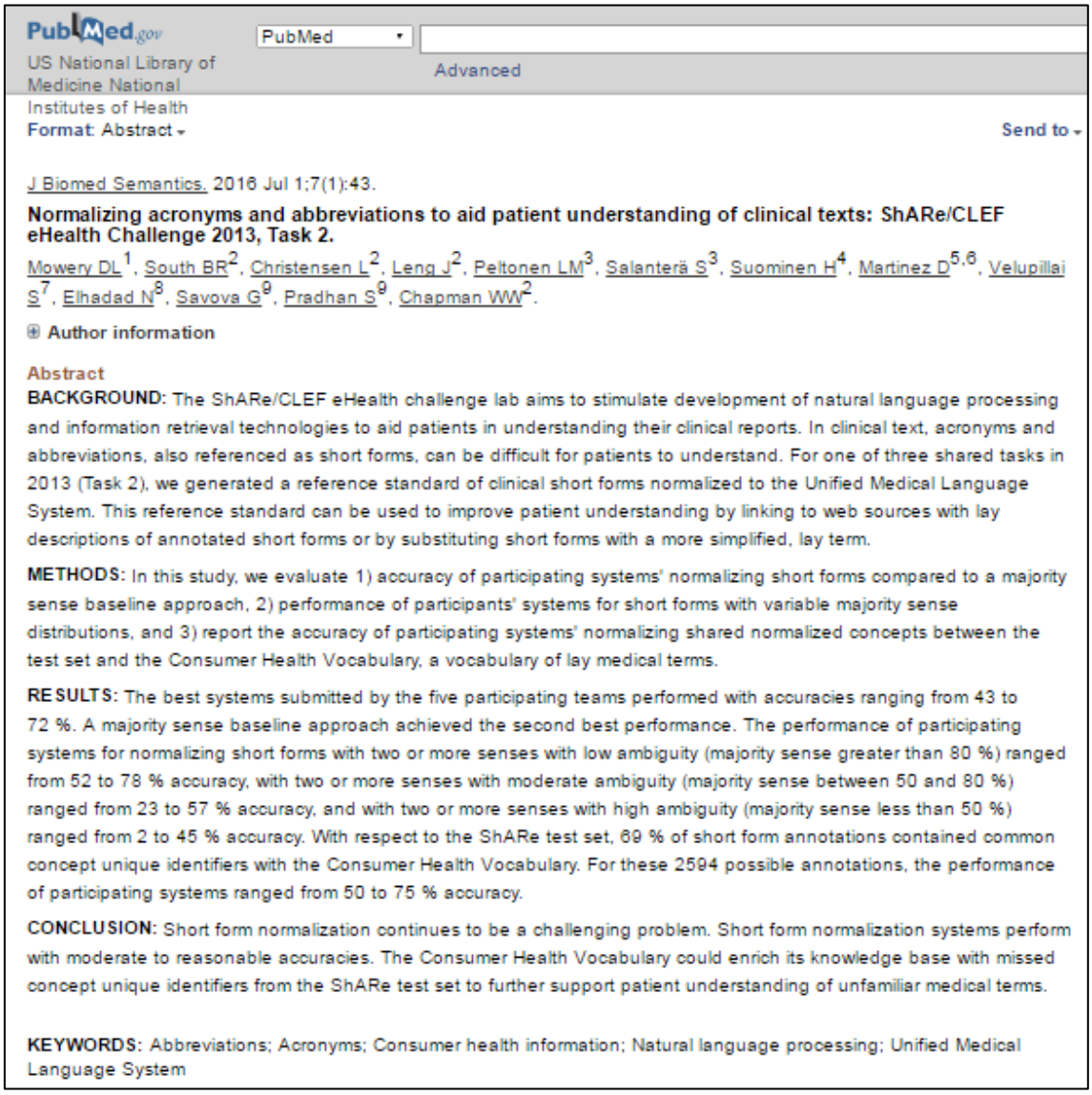
表 2.1 論文概要の構造化に関する取り組みの概観

研究者・雑誌名など	構造タグ	詳細内容
冨浦ら[1]	BACKGROUND (背景), PURPOSE (目的), EXPLANATION (手法等の説明), RESULT (結果), SIGNIFICANCE (意義)	英語論文の各文に人手で構造タグを付与し，タグ付きコーパスを作成
Yamamoto ら[2]	BACKGROUND (背景), PURPOSE (目的), METHOD (手法), RESULT (結果)	日本語論文の各文に人手で構造タグを付与し，タグ付きコーパスを作成
米国医学会雑誌, 診療ガイドライン など	BACKGROUND (背景), OBJECTIVE (目標，目的), METHODS (手法), RESULTS (結果), CONCLUSION (まとめ) など	一部の医学誌や日本での診療ガイドラインでは，構造化された各項目に記載するといった概要の作成を推奨している



(富浦, 2012)[1]より抜粋

図 2.3 論文概要への構造タグの付与例


PubMed.gov
US National Library of Medicine National Institutes of Health
PubMed
Advanced
Format: Abstract
Send to

J Biomed Semantics. 2016 Jul 1;7(1):43.

Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2.

Mowery DL[1], South BR[2], Christensen L[2], Leng J[2], Peltonen LM[3], Salanterä S[3], Suominen H[4], Martinez D[5,6], Velupillai S[7], Elhadad N[8], Savova G[9], Pradhan S[9], Chapman WW[2].

Author information

Abstract

BACKGROUND: The ShARe/CLEF eHealth challenge lab aims to stimulate development of natural language processing and information retrieval technologies to aid patients in understanding their clinical reports. In clinical text, acronyms and abbreviations, also referenced as short forms, can be difficult for patients to understand. For one of three shared tasks in 2013 (Task 2), we generated a reference standard of clinical short forms normalized to the Unified Medical Language System. This reference standard can be used to improve patient understanding by linking to web sources with lay descriptions of annotated short forms or by substituting short forms with a more simplified, lay term.

METHODS: In this study, we evaluate 1) accuracy of participating systems' normalizing short forms compared to a majority sense baseline approach, 2) performance of participants' systems for short forms with variable majority sense distributions, and 3) report the accuracy of participating systems' normalizing shared normalized concepts between the test set and the Consumer Health Vocabulary, a vocabulary of lay medical terms.

RESULTS: The best systems submitted by the five participating teams performed with accuracies ranging from 43 to 72 %. A majority sense baseline approach achieved the second best performance. The performance of participating systems for normalizing short forms with two or more senses with low ambiguity (majority sense greater than 80 %) ranged from 52 to 78 % accuracy, with two or more senses with moderate ambiguity (majority sense between 50 and 80 %) ranged from 23 to 57 % accuracy, and with two or more senses with high ambiguity (majority sense less than 50 %) ranged from 2 to 45 % accuracy. With respect to the ShARe test set, 69 % of short form annotations contained common concept unique identifiers with the Consumer Health Vocabulary. For these 2594 possible annotations, the performance of participating systems ranged from 50 to 75 % accuracy.

CONCLUSION: Short form normalization continues to be a challenging problem. Short form normalization systems perform with moderate to reasonable accuracies. The Consumer Health Vocabulary could enrich its knowledge base with missed concept unique identifiers from the ShARe test set to further support patient understanding of unfamiliar medical terms.

KEYWORDS: Abbreviations; Acronyms; Consumer health information; Natural language processing; Unified Medical Language System

図 2.4 PubMed に収録されている構造化概要の例

2.2.2　構造解析技術に関する従来研究

一方で，論文概要を自動的に構造解析する研究は多く存在する．表 2.2 に，自動構造解析技術に関する従来研究をまとめる．

表 2.2 論文概要に対する構造解析技術の概観

研究者	構造タグ	研究内容
Salanger-Meyer[4]; Swales[5]; Orasan[6]	problem, solution, evaluation, conclusion	修辞構造理論に基づいて論文概要を分析
Marcu[7]; Teufel ら[8]; 三池ら[9]	話題，背景，従来の問題，特徴，結果，結論，課題など	構造タグの特徴を表す手掛かり語を用いた構造解析
Biber ら[12]	Introduction, Methods, Discussion, Results	各項目間で使用される言語的な手掛かり語の特徴の違いを調査
Kando[13, 14]	階層的な構成要素カテゴリ (図 2.5)	構成要素カテゴリの特徴を表す手掛かり語を用いた構造解析
Willot ら[18]	BACKGROUND, METHODS, RESULTS, CONCLUSION	ニューラルネットワークを用いた構造解析
Teufel ら[19]; Ruch ら[20]	AIM, BACKGROUND, OWN, BASIS, CONTRAST など	ナイーブベイズモデルを用いた構造解析
Lin ら[21]; Wu ら[22]	INTRODUCTION, METHODS, RESULTS, CONCLUSION など	隠れマルコフモデルを用いた構造解析
Dai ら[23]	BACKGROUND, PURPOSE, METHODS, RESULTS, CONCLUSION	Markov Logic Network を用いた構造解析
McKnight ら[24]; Shimbo ら[25];Ito ら[25]; Yamamoto ら[27]	INTRODUCTION, METHODS, RESULTS, CONCLUSION など	Support Vector Machine を用いた構造解析
Hirohata ら[29]; Lin ら[30]	OBJECTIVE, METHODS, RESULTS, CONCLUSION など	Conditional Random Field を用いた構造解析

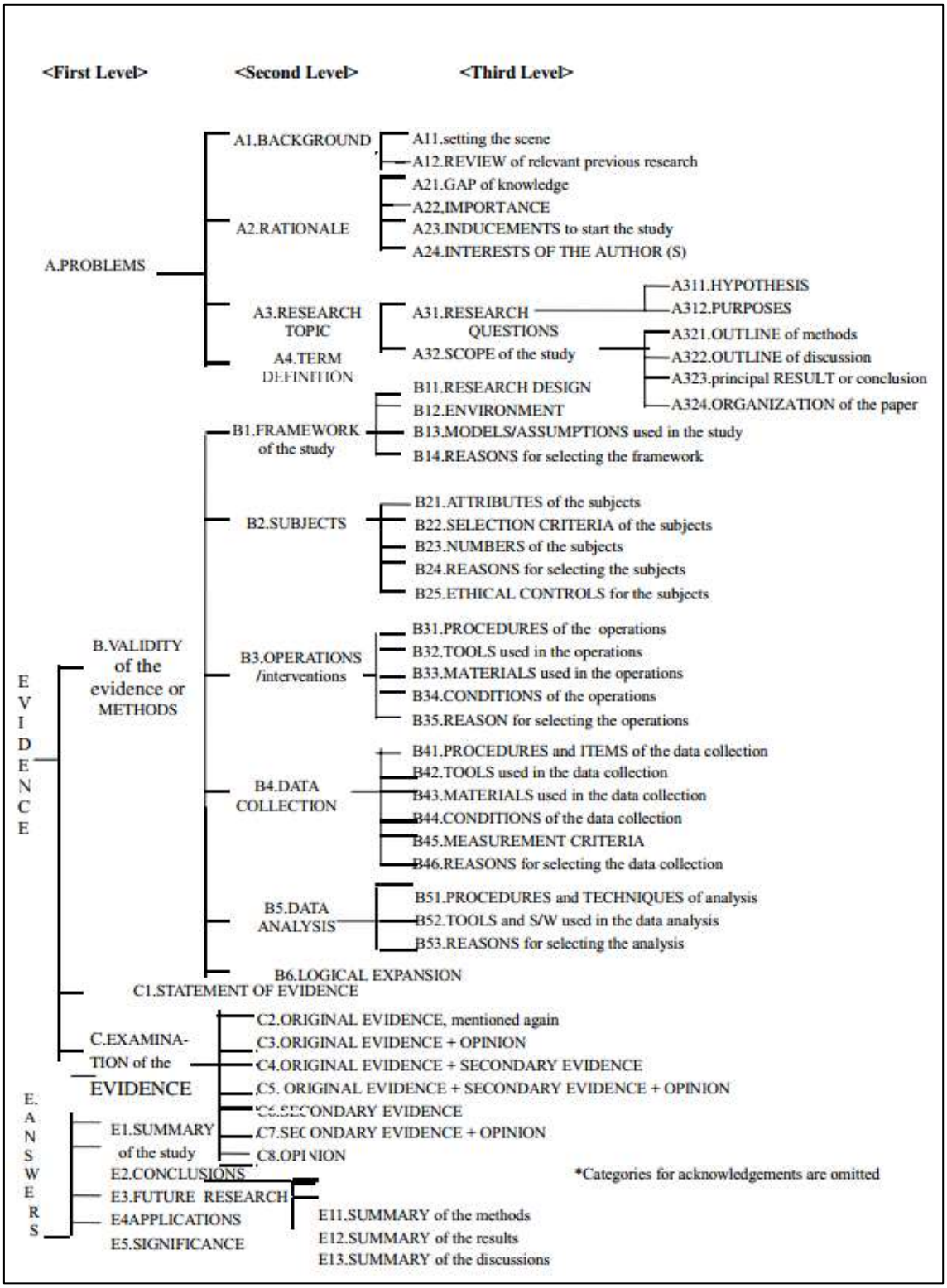
古くは，修辞構造理論に基づいて構成されていると仮定して分析を試みた研究も多く存在する．修辞構造理論とは，談話解析において，修辞関係として文間の関係を意味的に表す理論である[3]．修辞構造理論により学術論文を解析したとき，論文概要は主に，problem,

solution, evaluation, conclusion で構成されていると報告されている[4, 5, 6]．このような修辞構造理論に基づいた構造解析では，手掛かり語を用いることが有効であると報告されており[7, 8]，その後，手掛かり語に基づく解析手法がいくつか提案された[9, 10, 11]．また，Biber ら[12]は，医学分野の論文における Introduction, Methods, Discussion, Results の 4 種類の項目で使用される言語の特徴を調査し，各項目間で言語的な特徴の違いを明らかにした．さらに，Kando [13, 14]は，階層的な構成要素カテゴリ(図 2.5)に基づく構造解析を試みている．このカテゴリは 3 階層から構成されており，「A.PROBLEM - A1.BACKGROUND - A11. Stating background WITHOUT REFERENCE」というように，論文の構造をより詳細に分類できることを示している．構造の解析には，手掛かり語を用いている．このような修辞構造や手掛かり語に基づいた構造解析は，論文のみならず，特許請求項[15]や法令文[16, 17]など, 様々な文書の解析においても汎用的に用いられている．

近年では，論文の構造解析をテキスト分類問題と捉え，各文にラベル付けするための手法が多く提案されており，ニューラルネットワーク[18]やナイーブベイズモデル[19, 20]や隠れマルコフモデル[21, 22]，Markov Logic Network (MLN) [23]といった確率モデルを用いた手法や，Support Vector Machine (SVM) [24, 25, 26, 27], Conditional Random Field (CRF)[28, 29, 30]といった機械学習を用いた解析手法が用いられている．特に，SVM や CRF といった機械学習手法は，教師あり学習と呼ばれており，事前に与えられた訓練データに基づき学習を行い，未知のデータを統計的に予測・分類する．一般に，訓練データには，表 2.1 で示したタグ付きコーパスなどが用いられることが多く，各カテゴリ(構造タグ)の特徴を機械学習により分析する．

2.3 意味的な情報の抽出技術

文章の意味的な役割推定に関する研究が多く行われている一方で，文章からの意味的な情報の抽出に関する研究も多く存在する．これは，文よりも短い単位で文章から重要な情報を抽出し，それが持つ意味的な役割を推定するタスクである．このような抽出技術は，文書検索，動向分析，知識体系の構築など，様々な言語処理技術の支援につながる[31, 32, 33, 34, 35, 36]．また，論文や特許といった技術文書だけでなく，新聞記事やレビュー文，電子カルテなどの文書を対象とした抽出技術が提案されており，様々な産業で利用できる実用的なツールやシステムの開発の基礎技術となっている．このタスクでも，構造解析技術と同様に，タグ付きコーパスの整備や意味的な情報の抽出手法が多く提案されている．以下で詳細を述べる．



(Kando, 1999)[14]より抜粋

図 2.5 階層的な構成要素カテゴリの例

2.3.1　タグ付きコーパスの構築に関する研究・取り組み

2.2節で述べた構造解析技術と同様に，情報抽出技術の客観的な評価や機械学習に用いる学習データとしての利用のために，タグ付きコーパスの整備が活発に行われている．表2.3に，その概観をまとめる．建石ら[37, 38]は，論文中に出現するモノとモノの意味関係を同定するためのタグ付きコーパスを構築した．彼女らは，論文概要中の各文に対して，エンティティを示す語句には「TEAM」「OBJECT」「MEASURE」のいずれかのタグを付与し，エンティティ間には「PERFORM (動作主体)」や「CONDITION (実験条件)」など16種類の有向関係タグを付与した．エンティティタグ，有向関係タグの詳細をそれぞれ表2.4, 表2.5にまとめ，実際にタグを付与した例を図2.6に示す．大田ら[39, 40]は，生命科学分野の論文概要を用いてGENIA専門用語コーパスを作成し, 公開している[13]. このコーパスでは，生命科学の専門家により，物質とその所在とともに，タンパク質名や細胞名などの意味クラスが人手で付与されている．意味クラスは，大田らが構築したGENIAオントロジー(図2.9)と呼ばれる小規模なオントロジーに基づいており，Substance(物質名)，Source(所在)，およびOther(その他)というサブコンポーネントからなる概念の階層関係で構成されている．また，専門用語にはこの葉ノードのいずれかを意味クラスと付与されている．新里ら[41]は，レストランについて書かれた文書から，レストランの雰囲気やジャンル，提供される料理名といった属性情報を付与したタグ付きコーパスを構築した. 彼らは, Web上の掲示板サイトなどに投稿されたレストランに関する質問文を収集し，その中で頻繁に尋ねられている事柄を分析し，64種類の属性情報を定義した(図2.8)．そして，118件のレストランについて書かれた745件の文書(6,080文)に対して人手でタグ付けを行い，レストランコーパスとして構築した．実際に付与されたタグの例を図2.9に示す．このコーパスの特徴として，図2.9における文1の <料理>，<食材>タグのように，タグがネストする場合や，文2の<予約>タグのように，複数の文にまたがってタグが付与され，述語を含む文表現もタグづけの対象となっていることが挙げられる．橋本ら[42]は，特定領域研究「日本語コーパス」[43]で構築されているBCCWJの白書，書籍，Yahoo!知恵袋各コアデータに対して，「関根の拡張固有表現階層」[14] [44, 45, 46]で定義されている200種類のタグを人手で付与した.「関根の拡張固有表現階層」は，MUC (Message Understanding Conference)プロジェクト[47]，IREXプロジェクト[48]，ACE (Automatic Content Extraction)[15]で策定された固有表現の定義に基づいて関根が拡張した固有表現の定義であり，4階層から構成されている(図2.10).

[13] http://www.geniaproject.org/
[14] http://nlp.cs.nyu.edu/ene/
[15] http://www.itl.nist.gov/iad/mig//tests/ace/

表 2.3 人手による情報抽出用タグ付きコーパス構築の概観

研究者・ワークショップ名	付与するタグ	対象文書
建石ら[37, 38]	エンティティタグ（表2.4） 有向関係タグ（表2.5）	論文概要
大田ら[39, 40]	GENIAオントロジー（図2.7）	論文概要
新里ら[41]	属性情報（図2.8）	レストランについて書かれた文書
橋本ら[42]	関根の拡張固有表現階層（図2.10）	BCCWJの白書，書籍，Yahoo!知恵袋
NTCIR-8 特許マイニングタスク 技術動向マップ作成サブタスク[49]	要素技術， 効果（属性，属性値） (例: 図2.11)	論文，特許
MedNLP 匿名化タスク [32, 34, 35]	年齢（aタグ），日時（tタグ），病院名（hタグ），場所（lタグ），個人名（pタグ），性別（xタグ）(例: 図2.12)	医療文書
MedNLP 症状と診断タスク [32, 34, 35]	病状・診断診断名（cタグ） (例: 図2.12)	医療文書

表 2.4 (建石, 2013) [37]で使用されたエンティティタグの説明

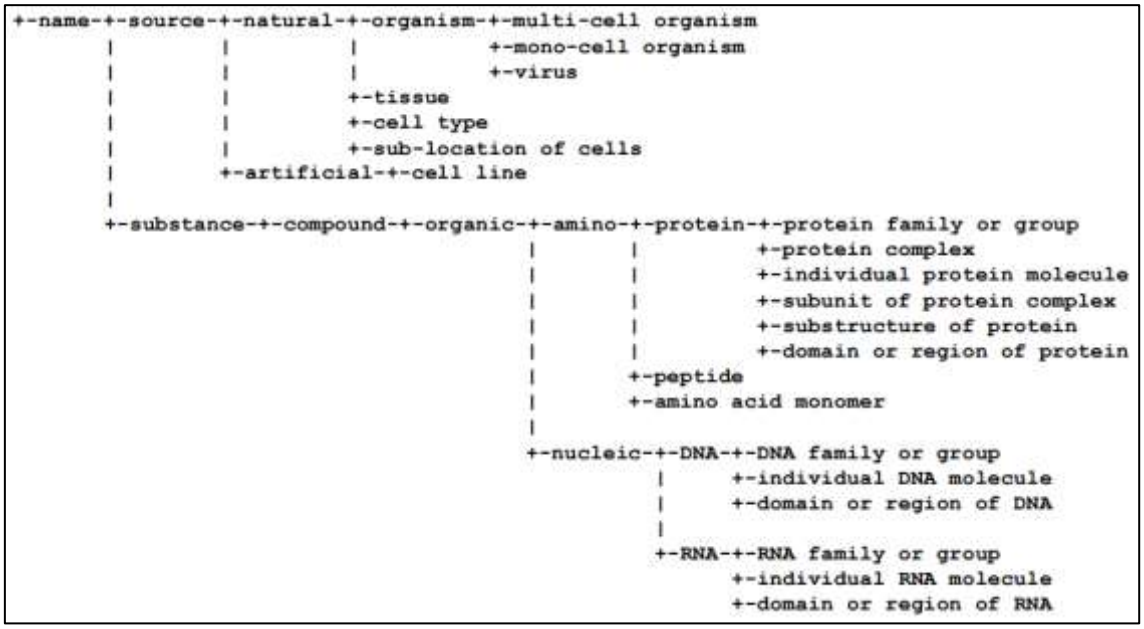
エンティティタグ	説明
OBJECT	システム名，国名，人名など実体を持つもの
MEASURE	評価，価値判断を示す語(「良い」，「高性能」など)，数値・数量表現，可能性(「できる」，「不可能」など)，必要性(「したい」，「必要」など)を表す語
TERM	上記以外の用語．動詞なども含む

表 2.5 (建石, 2013) [37]で使用された有向関係タグの説明

タグの説明	有向関係タグ
システムの振る舞いに関するもの	PERFORM (動作主体), APPLY_TO (アルゴリズム, 手段, 意図した結果, 目的), RESULT (意図しない結果, 副作用, 因果関係), INPUT (入力, 材料), OUTPUT (出力, 成果物), TARGET (動作対象でそれ自身が変化しないもの), ORIGIN (起点), DESTINATION(着点)
性質に関するもの	CONDITION (実験条件, 状況), ATTRIBUTE (属性, 特徴), STATE ((筆者以外の)他のものに関する評価)
評価に関するもの	EVALUATE ((筆者による)評価結果), COMPARE (評価の際の比較対象)
その他	SUBCONCEPT (上位-下位, 全体-部分, クラス-インスタンス), EQUIVALENCE (定義, 略称, 照応), SPLIT (カッコなどで分断された語句の前後をつなぐ)

(建石, 2013) [37]より抜粋

図 2.6 論文概要に表 2.4, 表 2.5 のタグを付与した例



(大田, 2005) [40]より抜粋

図 2.7 GENIA オントロジーの例

属性名	要素数	要素の例
店名	736	吉野家，ドトール
住所	248	目黒区自由が丘 X-X Y ビル B1F
電話	247	
FAX	51	
URL	47	
EMAIL	13	
地域	251	東京，自由が丘，首都圏
最寄路線	162	大井町線，東急東横線
最寄駅	230	田園調布駅，自由が丘駅正面口
最寄施設	79	大本木ヒルズ，みずほ銀行
距離	212	徒歩８分，歩いて１０分くらい
立地場所	68	素朴な商店街の中，路面店
ジャンル	307	トスカーナ料理，イタリアン
料理	2,064	帆立と三つ葉の挟み揚げ，日本酒
料理の質	1,188	かなりヘルシーな感じでした，辛い
食材	974	旬の素材，日本一のマグロ，豚
食事形式	261	ブリードリンク，ディナー
品揃え	156	充実しています，多数!!
形態	260	ファミレス，オープンカフェ風
歴史	81	老舗，今年４月にオープンしました
雰囲気	364	穴場的，ちょっと小洒落た
内装	69	凝っている，黒を基調とした空間だ
照明	23	暖かい灯り，夜はキャンドルが光る
清潔さ	16	厨房も清潔感がある，清潔感に溢れ
静けさ	7	中はＰＵＢならではの賑やかさ
賑め	2	賑めがいい
BGM	26	ジャズ，６０～８０年代の音楽
広さ	43	中も結構広いです，こじんまり
席数	63	テーブル３５席，１５席くらい
設備	211	テラス，席から見えるピザ窯
食器	22	真っ白な食器，手作りのデカンタ
外観	43	かわいらしい一軒家，看板が無い
駐車	35	駐車場サービス券有，Ｐなし

属性名	要素数	要素の例
営業日	397	不定休，無休，祝祭，水～土
営業時間	373	夜も遅くまで，５時半より１１時
メディア	15	グルメジャーナル２００４年５月号
経営者	99	下北沢 "X" の初代マスターだった店長
シェフ	76	本場上海から招いた点心職人
従業員	103	店員さんもとても親切
予約・貸切	55	貸切可能，当日予約で大丈夫
エンターテイメント	17	生演奏，野球観戦ができる
ドレスコード	2	普段着，普段着のまま
クレジットカード	223	AMMEX，カード使用不可
身体障害者	0	
子供	39	子どもを遊ばせることができ
喫煙	5	禁煙，喫煙可
ペット	13	犬を連れて行っても OK
その他のサービス	159	送迎バス有り
価格設定	474	一人５０００円
評判・知名度	90	お客さんが入れ替わるが常に満席
ビジネススタイル	27	全国チェーン
客層	182	スポーツ好き，主婦
その他の特徴	24	ベーカリーが併設されている
プロフィール	47	友だち同士，家族，５０代，男性
人数	107	５人，数人
希望日・利用日	31	１０月５日，正月
利用目的	193	飲み会，宴会，送別会
評価・印象	140	足繁く通いたくなる店です
もしくは	0	
できることなら	1	
でも可	0	
例えば	2	
のような	2	
除外	124	

(新里, 2006) [41]より抜粋

図 2.8 レストランに関する属性情報一覧

文１ 私の ＜**料理**＞＜**食材**＞ 紅茶 ＜/**食材**＞ のムースタルト ＜/**料理**＞ も ＜**料理の質**＞ おいしかった♪ ＜/**料理の質**＞

文２ ＜**予約**＞できましたら、事前に予約を入れて頂くことにより、より満足頂けるサービスをご提供出来るかと存じます。フォームをご用意させて頂きましたが、お電話やＦＡＸでもかまいませんので、よろしくお願い申し上げます。＜/**予約**＞

(新里, 2006) [41]より抜粋

図 2.9 レストランコーパスの例

拡張固有表現階層			
名前	名前 その他		
	人名		
	神名		
	組織名	組織名 その他	
		国際組織名	
		公演組織名	
		家系名	
		民族名	民族名 その他
			国籍名
		競技組織名	競技組織名 その他
			プロ競技組織名
			競技リーグ名
		法人名	法人名 その他
			企業名
			企業グループ名
		政治的組織名	政治的組織名 その他
			政府組織名
			政党名
			内閣名
			軍隊名
	地名	地名 その他	
		温泉名	
		ＧＰＥ	ＧＰＥ その他
			市区町村名
			郡名
			都道府県州名
			国名
		地域名	地域名 その他
			大陸地域名
			国内地域名
		地形名	地形名 その他
			山地名
			島名
			河川名
			湖沼名
			海洋名
			湾名
		天体名	天体名 その他
			恒星名
			惑星名
			星座名
		アドレス	アドレス その他
			郵便住所
			電話番号
			電子メイル
			ＵＲＬ

拡張固有表現階層			
名前	製品名	製品名 その他	
		材料名	
		衣服名	
		貨幣名	
		医薬品名	
		武器名	
		株名	
		賞名	
		勲章名	
		罪名	
		便名	
		等級名	
		キャラクター名	
		識別番号	
		乗り物名	乗り物名 その他
			車名
			列車名
			飛行機名
			宇宙船名
			船名
		食べ物名	食べ物名 その他
			料理名
		芸術名	芸術名 その他
			絵画名
			番組名
			映画名
			公演名
			音楽名
			文学名
		出版物名	出版物名 その他
			新聞名
			雑誌名
		主義方式名	主義方式名 その他
			文化名
			宗教名
			学問名
			競技名
			流派名
			運動名
			理論名
			政策計画名
		規則名	規則名 その他
			条約名
			法令名
		称号名	称号名 その他
			地位・職業名
		言語名	言語名 その他
			国語名
		単位名	単位名 その他
			通貨名

(橋本, 2008) [42]より抜粋

図 2.10 「関根の拡張固有表現階層」の例

上記で述べたコーパスの構築に関する取り組みから分かるように，研究者が設定するタスクによって抽出対象となる意味的な表現は異なる．しかし，学術研究としての(情報抽出技術を含む)言語処理技術の有用性や実用性を評価するためには，問題を定式化し，複数の研究グループ間で客観的な評価を比較することが必要である．そのため，近年では，言語処理技術の隆盛のための枠組みとして，大規模なタグ付けコーパスを研究グループ間で共有するための評価ワークショップが開かれている．その中でも情報抽出に関連した評価ワークショップとして，NTCIR-8 特許マイニングタスク[49]やMedNLP[32, 34, 35]などがある．以下で，各ワークショップの概観について述べる．

NTCIR-8 特許マイニングタスクは，特許と論文を対象とした情報処理のための研究プロジェクトであり，タスク内容として技術動向マップ作成サブタスクが設定されている．このサブタスクは，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ことを目的としている．例えば，図2.11に示すように，入力された文に対して，要素技術と効果を示す箇所にそれぞれTECHNOLOGYタグとEFFECTタグを自動的に付与する．ここで，EFFECTタグの中には，属性を表すATTRIBUTEと属性値を表すVALUEという2種類のタグが付与されている．図2.11に示すような要素技術と効果の対を抽出できれば，技術動向を可視化したマップを作成可能となる．MedNLPは，医療産業で利用できる実用的なツールの開発を最終的な目標とし，その第一歩として，日本語医療文書から重要な情報を抽出するための評価ワークショップを開催している．このワークショップでは，三種類のタスクが設定されており，その中でも，匿名化タスクと症状タスクでは，医療文書から患者の病歴に関する情報を抽出することを目的としている．匿名化タスクでは，図2.12に示すように，患者の個人情報を示す個所にタグが付与されている．また，病状と診断タスクでは，図2.12のように，患者の病状・診断名を示す個所にcタグが付与される．ここで，cタグには，医師の認識の程度などを表すモダリティ属性が付随している場合があり，肯定，否定，推量をそれぞれpositive, negation, suspicionで表す．また，病状が患者の家族の病歴である場合は，familyで表される．MedNLPでは，日本語文書を対象としているが，英語の医療文書を対象とした評価ワークショップとしては，i2b2[16]が挙げられる．このワークショップでは，患者の喫煙状態[50]や医薬品の使用状況[51]，医療上のコンセプト[52]に関する情報を自動的に抽出するといったタスクが設定されている．上記で述べたような評価ワークショップに参加することで，各ワークショップで構築しているタグ付きコーパスが配布され，利用することができる．また，一部の過去の評価ワークショップでは，利用手続きを行うことで，タグ付きコーパスを含むテストコレクションを利用することができる．

[16] https://www.i2b2.org/

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 >を施すため，<EFFECT><ATTRIBUTE>電力損失</ ATTRIBUTE>を<VALUE>最小化</VALUE></ EFFECT >できる．

図 2.11 技術動向マップ作成サブタスクに用いるデータの例

工場に勤めている<a>64 歳の<x>男性</x>．
<t>2025 年月 8 月 2 日(来院 5 日前)頃から</t><c>腹痛</c>が生じるとともに，<c>食欲不振</c>，<c>嘔気</c>・<c>嘔吐出現</c>した．
体幹は温かいが，末梢は<c>湿潤冷汗</c>で<c>ショック状態</c>．
明らかな<c modality="negation">運動麻痺</c>はみられず．
<t>翌日</t>，<c>意識障害出現</c>し，<c>腎機能障害</c>の増悪を認めて徐々に<c>尿量低下</c>し，<t>8 月 9 日 18 時 10 分</t>に<c>心肺停止</c>．
<t>8 月 9 日 21 時 44 分</t><c>死亡確認</c>．

図 2.12 MedNLP に用いるデータの例

2.3.2 意味的な情報の抽出技術に関する従来研究

意味的な情報の自動抽出に関する従来研究の概観を表2.6に示す．以下では，表2.6で示した従来研究について述べる．

Matsumuraら[53]は，ある概念を表すキーワードである「概念語」と，助詞や動詞のような，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．自然文に対して，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．そのため，Matsumuraらは，文書表題のような擬似的な自然文を対象に，係り受け関係を解析することでインデックスを作成している．三平ら[54]は，日本語論文を対象に，論文中の参照情報が出現する文の前後から，著者の主題表現を抽出している．主題表現の抽出には手がかり語を用いており，参照の出現する文の前後の文に手掛かり語が出現すれば，その箇所を主題表現として抽出している．Guptaら[31]は，研究アイディアの発展過程を調べるために，「FOCUS」「TECHNIQUE」「DOMAIN」という3種類のカテゴリに該当する語句を論文概要から自動的に抽出する手法を提案している．彼女らの手法はパターンマッチに基づいており，例えば，動詞「propose」の直後に出現する直接目的語を「FOCUS」を表す語句として抽出している．西山ら[55]は，特許公開公報と新製品発表デ

表 2.6 情報抽出タスクに関する従来研究

研究者・プロジェクト	抽出表現	抽出対象文書	抽出手法
Matsumuraら[53]	概念語，関係語	文書表題	係り受け解析
三平ら[54]	主題表現	論文	パターンマッチ
Guptaら[31]	FOCUS, DOMAIN TECHNIQUE,	論文概要	パターンマッチ
西山ら[55]	新技術, 新製品が持つ好ましい性質や新機能などの効果表現	特許公開公報と新製品発表データ	パターンマッチ
酒井ら[56]	出願目的情報, 技術課題情報	特許明細書	手掛かり語に基づくルールベース
近藤ら[57]	HEAD, METHOD, GOAL	論文表題	CRF，手掛かり語，不要語リスト
Tateisiら[58]	有向関係	論文概要	SVM
新里ら[41]	属性情報	レストランについて書かれた文書	SVM, レストランドメインに特化した専門辞書
橋本ら[59]	固有表現	BCCWJの白書, 書籍，Yahoo!知恵袋	CRF
NTCIR-8 特許マイニングタスク	要素技術, 効果（属性，属性値）	論文，特許	機械学習（SVM, CRF）ドメイン適応
MedNLP 匿名化タスク	年齢，日時，病院名，場所，個人名，性別	医療文書	ルールベース 機械学習 など
MedNLP 症状と診断タスク	患者の症状名, 医師の診断名	医療文書	機械学習, 医学分野を専門とした用語辞書など

ータ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．これらの文書の中に

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．例えば，「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，複数の手掛かり語を用いて，その手掛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．酒井ら[56]は，技術動向の可視化に用いる情報として，特許明細書から，出願目的および技術課題情報を自動的に抽出する手法を提案した．ここで，技術課題は，特許発明を使用することにより解決される課題を表し，出願目的は，特許を出願した目的を表す．出願目的情報は，特許明細書の「発明が解決しようとする課題」の項目から取得しており，「本発明は」「を提供する」「を課題」「を目的」のいずれかの表現を含む文を抽出している．技術課題情報は，特許明細書における「発明の効果」の項目から取得しており，「ができる」をはじめとする36種類の手掛かり語を含む文を抽出している．この36種類の手掛かり語は，手掛かり語に係る確率に基づくエントロピーを用いて自動的に収集している．

上記で述べた手法は，人手で作成したルールに基づいて解析している．一方で，近年では，機械学習を用いた統計的な解析手法が多く提案されており，タグ付きコーパスを訓練データとして用いることで，解析精度の向上を目指している．近藤ら[57]は，ある研究分野において，「どのような要素技術がいつ頃から使われているのか」という情報を網羅的に収集する手法を提案した．彼らは，機械学習手法であるCRFを用いて，論文表題から「HEAD(主題)」「METHOD(要素技術)」「GOAL(目的・目標)」に対応する表現を自動的に抽出している．機械学習に用いる素性として，手掛かり語と技術的な表現を示さない不要語を収集したリスト(例:「研究」,「検討」)を用いており，訓練データとして，各タグを人手で付与した論文表題集合を使用している．なお，この不要語リストは半自動的に構築されており，機械学習によりHEADとして抽出した文字列集合から人手で選定している．Tateisiら[58]は，論文中に出現するモノとモノの意味関係を同定するための手法を提案した．図2.6で示したエンティティタグと有向関係タグが付与されたコーパスを用いてSVMによる関係抽出器を作成し，その結果に基づいて論文中の文を解析している．新里ら[41]は，属性タグが付与されているレストランコーパス(図2.9)を用いて，レストランに関する属性情報を自動的に抽出する手法を提案した．彼らは，SVMに基づく機械学習手法を適用しており，素性として，レストランドメインに特化した専用の辞書を用いている．この辞書は，Web上から収集したHTML文書から自動的に構築しており，相互情報量に基づいてレストランに関連する固有表現を抽出している．橋本ら[59]は，白書，書籍，Yahoo!知恵袋のタグ付きコーパスを用いた固有表現認識手法を提案した．機械学習にはCRFを用い，素性として，一般的に使用される形態素，品詞，固有表現タグを用いた．彼らは，学習に用いる訓練データにおいて，各ドメインコーパスを単独で用いる場合とすべてのコーパスを同時に使用する

ことを試みている．その結果，すべてのコーパスを同時に学習した場合，単独でコーパスを用いる場合に比べて，再現率が向上するという傾向を示した．

次に，表2.6で挙げた各評価ワークショップで提出されたシステムの全体的な概要について述べる．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[49]に参加した研究グループの多くは，SVMやCRFなどの機械学習を用いており，最も精度が高かったシステムでは，CRFが用いられていた．このサブタスクにおいて，Nishiyama[60]らは，日本語論文および日本語特許を対象としてFEDA (Frustratingly Easy Domain Adaptation) [61]と呼ばれるドメイン適応手法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手法である．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．MedNLP[32, 34, 35]における「匿名化タスク」では，ルールベースに基づく手法がいくつかのチームで採用されていた．一方で，「症状と診断タスク」に参加したチームの多くは，個人情報，診断情報を固有表現とした固有表現抽出とみなし，医学分野を専門とした用語辞書と学習アルゴリズムCRFを組み合わせた機械学習手法を採用していた[62, 63, 64].

2.4 考察

構造解析技術および意味的な情報抽出に関する研究で共通する事柄として，以下が考えられる．

1. タグ付けコーパスの構築
2. 手掛かり語やパターンマッチによる自動解析
3. タグ付けコーパスを訓練データとした機械学習手法

手掛かり語やパターンマッチに基づく解析は，タグ付けコーパスのような大規模な訓練データを必要としない．しかし，文書を作成した著者や執筆者によって，同じの意味の手掛かり語やパターンでも表現がそれぞれ異なる場合がある(例:「を用いて」,「を使用して」). そのため，人手で作成するルールに対応付けてすべての文書を網羅的に解析することは難しいといえる．

近年では，統計的な解析モデルである機械学習が有望とされており，タグ付けコーパスを

訓練データとして用いて学習を行うことで，未知のデータを自動的に予測・分類する．機械学習には，各カテゴリの識別に用いる情報として素性が必要となる．素性を学習に用いることで，各カテゴリの特徴を効率的に捉えることができ，適切な素性選択を行うことで高精度な学習器を構築することができる．一般には，単語情報や品詞情報などが用いられる．一方で，手掛かり語やパターンは，特定の意味的な表現の識別に有用である．本論文では，要素技術とその効果の識別に有用な手掛かり語を素性として与えることで高精度な解析を目指す．

また，機械学習において，特定のドメインの解析には，そのドメインのコーパスを訓練データとして用いることが一般的である．例えば，機械学習により論文概要から意味的な情報を抽出する場合，論文用コーパスが学習に用いられる．一方で，橋本ら[42]や Nishiyama ら[60]のように，ドメイン適応により，異なるドメインコーパスを組み合わせて学習に用いることの有効性を示している．本論文でも同様に，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訓練データとして使用し，学習量を増加させ，解析精度の向上を目指す．また，本論文では，ドメイン適応に基づく新たな解析手法を提案する．

2.5 まとめ

本章では，まず，構造解析技術および意味的な情報の抽出に関する諸研究を紹介した．そして，それぞれの解析技術における共通する事柄について考察し，論文と特許から要素技術とその効果に関する表現をより高精度に抽出するための方針を述べた．

3章では，機械学習を用いた要素技術とその効果に関する表現の抽出手法，機械学習の素性として用いる手掛かり語の収集方法および本論文で提案するドメイン適応手法について詳細を述べる．そして，4章と5章で，要素技術とその効果に関する表現を利用した応用例を述べる．

第3章　特許と論文からの要素技術と効果の自動抽出

3.1 はじめに

本論文では，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，本論文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．本論文における「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出来ると考えられる．

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．本論文ではこの問題を「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．例えば，「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掛かり語のリストを作成しておき，各々のリスト中の手掛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．そこで本論文では，機械学習に用いる手掛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，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場合，論文用または特許用タグ付きコーパスを単独で用いることが一般的である．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．そこで本論文では，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．

3.2 論文と特許の表題および概要の構造解析

3.2.1 構造タグの定義

本論文では，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，本論文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掛かり語を示す．

- **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**：例えば，「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．

図 3.1 と図 3.2 に，論文特許に対して上記のタグを付与した例をそれぞれ示す．この例は，NTCIR-8 ワークショップ 特許マイニングタスクで提供された訓練データの一部から抜粋している．本論文では，論文の「表題」と「概要」に，特許では「発明の名称(以後，特許表題)」と「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(以後，特許概要)」に対してタグの自動付与を行う．

<TOPIC>
<TOPIC-ID>726</TOPIC-ID>
<IPC-LIST><IPC>G06F_15_18</IPC><IPC>G06F_17_30</IPC></IPC-LIST>
<TITLE><TECHNOLOGY>Support Vector Machine</TECHNOLOGY>によるテキスト分類</TITLE>
<ABSTRACT>本稿では，<TECHNOLOGY>Support Vector Machine (SVM)</TECHNOLOGY>を用いたテキスト分類法を提案する. テキスト分類問題に対して学習手法を適用する場合，出現頻度の小さい単語まで考慮して学習を行なわいと，分類精度が落ちることが知られている．このため<EFFECT><VALUE>高い</VALUE><ATTRIBUTE>分類精度</ATTRIBUTE></EFFECT>を実現するためには，高次元の単語ベクトルを用いなければならないが，過学習により分類精度が落ちてしまう危険性が生じる．<TECHNOLOGY>SVM</TECHNOLOGY> は <TECHNOLOGY>Kernel 関数</TECHNOLOGY>により非線形学習も可能であり, 高次元の入力ベクトルを用いても過学習なしに最適解が得られる. <TECHNOLOGY>SVM</TECHNOLOGY>をテキスト分類に適用し，1. 異なる次元の単語ベクトル，2. 異なる<TECHNOLOGY>Kernel 関数</TECHNOLOGY>，3.異なる目的関数，の 3 点について比較実験を行なった．その結果，<TECHNOLOGY>SVM</TECHNOLOGY>が<EFFECT><ATTRIBUTE>テキスト分類問題</ATTRIBUTE>に対して<VALUE>有効</VALUE></EFFECT>であることが確認された．</ABSTRACT>
</TOPIC>

図 3.1 論文表題および概要へのタグ付与例

<TOPIC>
<TOPIC-ID>514</TOPIC-ID>
<IPC-LIST><IPC>F02M_61_18</IPC></IPC-LIST>
<TEXT>
<TITLE>燃料噴射ノズル</TITLE>
<PATENT-PROBLEM>【発明が解決しようとする課題】しかしながら，前述の燃料噴射ノズルでは，噴射圧力，噴射量，噴射期間等が，燃料噴射ノズルへ燃料を送油する噴射ポンプによって決定されてしまう構造であり，噴孔の数が固定されており，噴射に有効な噴孔の総面積を増減できない構造となっている．このため，エンジン低速回転時には噴射圧が低くなってしまったり，エンジン低負荷時には，噴射時間が短くなってしまったりするなど，良好な燃焼状態を継続させることが困難になるという問題点があった．
</PATENT-PROBLEM>
<PATENT-MEAN>【課題を解決するための手段】しかして，この発明にかかる燃料噴射ノズルは，加圧燃料を噴射する噴孔がノズルボディの先端部に形成され，<TECHNOLOGY>前記ノズルボディの周囲に摺動自在に回動するカバー部材</TECHNOLOGY>を設け，<TECHNOLOGY>このカバー部材を回動させることにより噴射に有効な前記噴孔の総面積を可変させる噴孔面積可変機構</TECHNOLOGY>を備えたことにある(請求項１)．この<TECHNOLOGY>噴孔面積可変機構</TECHNOLOGY>は，例えば，カバー部材をコントロールユニットからの信号で回転する<TECHNOLOGY>マイクロモータ</TECHNOLOGY>によって回動させる機構が考えられる．</PATENT-MEAN>
<PATENT-EFFECT>【発明の効果】以上述べたように，請求項１乃至３にかかる発明によれば，噴射に有効な噴孔の総面積を<TECHNOLOGY>噴孔面積可変機構</TECHNOLOGY>によってカバー部材を回動させることにより調節することができるので，エンジンの負荷と回転数とに対応した噴射圧力，噴射期間，噴射量を得ることが可能となり，<EFFECT><ATTRIBUTE>ＮＯｘ</ATTRIBUTE>の<VALUE>低減</VALUE></EFFECT>や<EFFECT><ATTRIBUTE>燃費</ATTRIBUTE>の<VALUE>向上</VALUE></EFFECT>を図ることができる．</PATENT-EFFECT>
</TEXT>
</TOPIC>

図 3.2 特許に対する「発明の名称」と「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」へのタグ付与例

3.2.2　構造解析手段

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掛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，要素技術の手掛かり語表現は分野依存性が低く，人手での手掛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．そこで本論文では，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掛かり語リストを作成する．

3.2.3　手掛かり語リストの作成

以下に，手掛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[65]と NTCIR-2[66]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，属性値になり得ない表現を人手で削除し，最終的に，300 の手掛かり語から成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集する．係り受け解析器には CaboCha[17]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，属性になり得ない表現を人手で削除し，最終的に，700 の手掛かり語から成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同士の類似度によって語の類似度を数値化する方法がある[67, 68]．相澤[69]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．本論文では，(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一つの方法として，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，名詞毎に共起語ベクトル(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[18]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掛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[70]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 3.3 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掛かり語として収集している．その後，収集した手掛かり語集合に対して閾値を設定

[17] http://code.google.com/p/cabocha/
[18] http://geta.ex.nii.ac.jp/geta.html

し，高頻度で出現した語句のみを新たな手掛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．

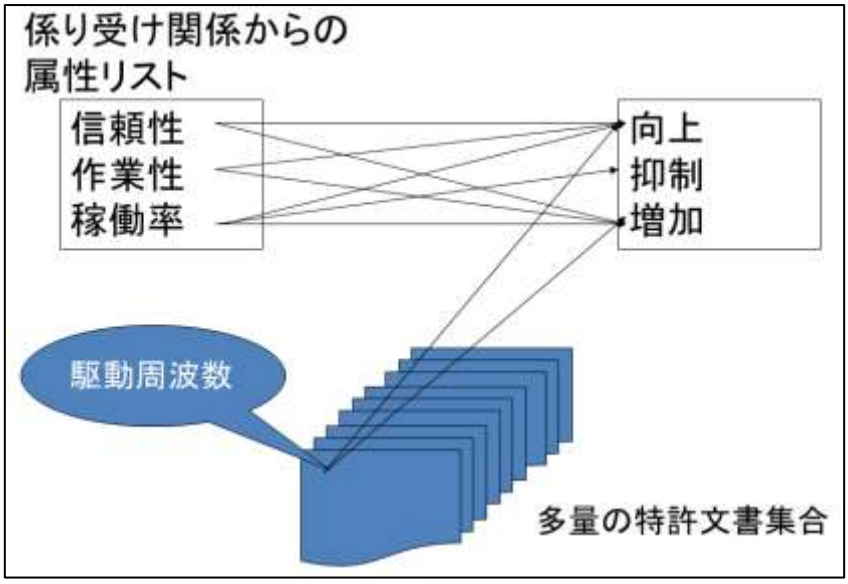


図 3.3 分布類似度により手掛かり語として"駆動周波数"を収集する例

(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，属性値における手掛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，本論文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．そこで本論文では，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 の手掛かり語から成る単位語リストを作成した．

3.2.2 節でも述べたように，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．例えば，図 3.1 における「<TECHNOLOGY>Support Vector Machine (SVM)</TECHNOLOGY>を用いたテキスト分類法を提案する」という文の場合，TECHNOLOGY タグが付与されている「Support Vector Machine (SVM)」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掛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．本論文では，要素技術になりうる専門用語(TECHNOLOGY-internal)を人手で収集すると共に，要素技術を示すような手掛かり語(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，手掛かり語を用いて判定する．例えば，"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．そこで，論文や特許の主題となるような手掛かり語の有無を素性のひとつとして用いる．

さらに本論文では，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．

3.2.4　機械学習に用いる素性，ツール，データ

本論文では，以下の 10 個の素性を機械学習に用いる．括弧内の数値は各リストの個数である．

概要中の各単語
品詞情報
ATTRIBUTE-internal(1210)：属性の手掛かり語の有無．(例: "処理量"，"精度")
EFFECT-external(21)：効果部の手掛かり語の有無．(例: "できる"，"実現する")
TECHNOLOGY-external(45)：要素技術の手掛かり語の有無．(例: "を用いた"，"に基づいた")
TECHNOLOGY-internal(17)：要素技術専門用語の有無．(例: "HMM"，"SVM")
VALUE-internal(408)：属性値の手掛かり語の有無．(例: "増加"，"抑止")
HEAD-exclusion(12)：主題となる不要語または主題の手掛かり語の有無．(例: "を提案"，"開発")
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．
UNIT-internal(274)：数値付き単位の有無．(例: "kg"，"cm")

論文および特許の構造解析に用いるツールには，SVM ベースのチャンキングツールである yamcha[19]を用い，形態素解析には MeCab[20]を用いる．機械学習で用いる入力データの例を図 3.4 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[71]でエンコードする．yamcha は，図 3.4 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．本論文では，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，本論文では，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，一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．

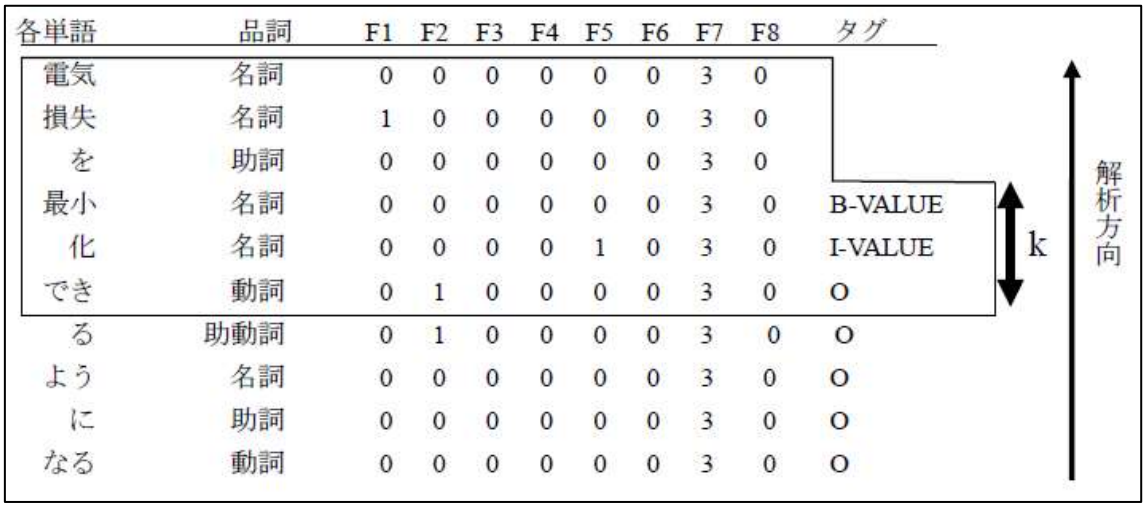


図 3.4 機械学習に用いる入力データ

[19] http://chasen.org/~taku/software/yamcha/
[20] http://mecab.sourceforge.net/

3.3 ドメイン適応を用いた情報抽出

本論文では，Nishiyama ら[60]が用いたドメイン適応手法である FEDA[61]を用いてその有効性を確認する．本論文でも同様に，ドメイン適応を利用することで解析精度の向上を行う．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，本論文では 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，このほかにも，あるドメインの素性を用いて一旦学習させ，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．本論文では，以下で述べる 2 種類の手法を提案する[21]．

[提案手法 1: SEQ]

1. 論文ドメインを訓練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 を用いて先程タグ付けされた論文にタグ付けを行う．

また，本論文では次の点を考慮する．特許における要素技術を表す表現句は，論文に比べて長く記述される傾向にある．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．本論文では上記の SEQ 手法に加え，この問題を考慮した手法を提案する．

[提案手法 2: SEQ(T)]

1. 論文ドメインを訓練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データとして用いてモデル B を獲得する．このとき，訓練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 を用いて先程タグ付けされた論文にタグ付けを行う．

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

3.4 実験

3.4.1 実験データ

本論文では NTCIR-8 特許マイニングタスク[49]で配布された論文および特許データを用いて実験を行った．1993 年から 2002 年まで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】に対して，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 件にも付与されている．このうち，300 件を訓練データ，200 件を評価データとして用いる．また，評価データにおいて，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 3.1 に示す．

表 3.1 評価データにおける人手で付与されたタグの数

	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

3.4.2 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．

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

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

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

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[49]．

3.4.3 比較手法

本論文では, 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におけ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): 3.2.4 節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(3.2.3 節における Step 1，Step 2)のみを用いる.
- J-ML: 3.2.4 節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: 3.2.4 節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，ドメイン適応手法を用いる.
- J-ML+SEQ: 3.2.4 節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，3.3 節で述べた提案手法 1 を用いる.
- J-ML+SEQ(T): 3.2.4 節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，3.3 節で述べた提案手法 2 を用いる.

ベースライン

- TRL7_1&TRL6_2 [60]: 表題および概要の解析を，機械学習(CRF[28])を用いて行う. 概要解析には CRF に加え，ドメイン適応手法を用いる. 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 [72]: 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 [73]: 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掛かり語(例: "用い", "備え")を用いる.
- HTC_1&HTC1_1 [74]: 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掛かり語リスト，日本語依存文法の構文解析を用いた修飾関係を用いる．

3.4.4　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 3.2 と表 3.3 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 3.2 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，本論文で提案した 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 改善している．一方，特許の解析結果について見ていくと，表 3.3 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 手法の性能が既に高く，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 3.4 と表 3.5 にそれぞれ示す．表 3.4 には論文の解析結果を，表 3.5 には特許の解析結果を示す．表 3.2，表 3.3 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, および F 値の平均値(AVERAGE)を示す．表 3.4，表 3.5 それぞれの結果から，本論文で提案したすべて手法は，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．

表 3.2 論文の表題および概要の解析結果

提案手法	評価	表題	概要				AVERAGE
		TECHNOLOGY	TECHNOLOGY	ATTRIBUTE	VALUE	EFFECT	
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 3.3 特許の表題および概要の解析結果

提案手法	評価	表題	概要				AVERAGE
		TECHNOLOGY	TECHNOLOGY	ATTRIBUTE	VALUE	EFFECT	
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

表 3.4 各ベースラインと比較した場合の論文の実験結果(AVERAGE)

	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 3.5 各ベースラインと比較した場合の特許の実験結果(AVERAGE)

	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

3.4.5 考察

分布類似度とドメイン適応手法の有効性

表 3.2，表 3.3 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．ゆえに本論文では，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，本論文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，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場合の結果を表 3.6 と表 3.7 に示す．なお，表 3.6 と表 3.7 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，特許共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，本論文の手法は，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，本論文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．

表 3.6 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）

	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 3.7 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）

	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

論文解析におけるドメイン適応手法の比較

表 3.2 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，本論文で提案した 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 3.2 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, 大幅に改善したことが分かる．本論文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，本論文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 3.8 に示す．この結果から，SEQ 手法，SEQ(T)手法を用いたことで改悪された件数はそれぞれ 2 件以下程度で留まっていることに対して，改善された件数は 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，本論文で提案した，「モデル A を用いて一旦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 の素性は, 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 を用いて既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わかった．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.

表 3.8 論文概要解析における提案手法を用いた場合の正解件数の比較

	手法	正解件数	J-ML 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

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…

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[22]．

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，これを考慮した本論文の 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．

[22]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．

特許解析におけるドメイン適応手法の比較

表 3.3 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一つである，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，本論文で用いた特許用の訓練データに「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，タグの付与が出来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，論文を訓練データとして用いて，既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出来たと考えられる．

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，既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. 例えば,「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，既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．

3.5 まとめ

本論文では，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，本論文では単語や品詞に加えて，要素技術，属性，属性値の手掛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掛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．さらに本論文では，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．

実験の結果，論文の解析において，本論文で提案した「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．

第4章　要素技術と効果を考慮した学術論文の自動分類

4.1 はじめに

研究領域全般を横断した学術論文の分類は，網羅的かつ効率的な論文検索や技術動向分析などの支援を可能にする．一部の学術論文データベースでは，特定の研究分野を対象にした分類体系が考案されており，この分類体系に基づいてデータベース内の論文を人手で分類している．しかし，これから発表されていく論文や未分類のすべての論文を人手で分類することは，非常にコストがかかる．また，対象とする分類体系が改訂されたとき，改めて人手で論文を分類しなおすのは現実的でない．そこで本論文では，特定の分類体系に基づいた学術論文の自動分類手法を提案する．

文書分類は，自然言語処理などのデータ解析の分野における代表的な研究課題の一つであり，これを解決するための手段として，事前に与えられたデータ集合に基づき未知のデータを自動的に予測・分類する機械学習手法が提案されている．本論文では，学術論文固有の特徴を用いた機械学習による分類手法を提案する．理工系などの分野における多くの論文中では，研究課題に対して提案された新しい技術や既存技術を応用した技術，研究課題を解決するための手段などが記述されている．また，これらの技術や手段などを用いて得られた知見を，研究課題に対する成果として述べている場合が多い．そして，特定の研究課題において有用とされている技術や手段が確認されたとき，それらは同一あるいは近い分野の他の研究課題にも利用されることも少なくない．このような研究動向を示す技術や手段(要素技術)とそれらにより得られた知見(効果)に関する情報は，研究分野の特徴を表す重要な手掛かりになると考えられる．本論文では，要素技術とその効果を自動的に抽出し，論文を分類するための手掛かりとして利用することで， その有用性を示す．また，人文社会系のような要素技術や効果に関する情報があまり見られないような研究分野に対しても，本手法が有効に機能するかどうかを検証する．

特定の研究領域で考案された分類体系に基づく学術論文の分類に関しては，これまでにもいくつかの研究が行われている[75, 76, 77, 78]．しかし本論文では，すべての研究領域を網羅した学術論文の分類を目指している．これを実現するための第一歩として，本論文

では，科学研究費助成事業データベース(KAKEN)の分類体系を用いる．KAKENとは，国立情報学研究所が文部科学省，日本学術振興会と協力して作成・公開しているデータベースであり，過去に採択された67万件以上の研究課題を検索することができる．KAKENの分類体系は，理工系，人文社会系，生物系といったほぼすべての研究分野を網羅しており，研究領域によって「系・分野・分科・細目表」と呼ばれる4種類の階層に構造化されている．また，文部科学省において，分類体系の審議・改訂が年度ごとに行われており，研究分野の新設や統廃合，細分化などが実施されている．このように，KAKENの分類体系は，最新の研究領域の動向を考慮した横断的な学術論文の分類に適している．

4.2 学術論文の自動分類に関する研究

学術論文の自動分類に関して，これまでにいくつかの先行研究がある．国立情報学研究所が主催したNTCIR-7,8 特許マイニングタスクで実施された学術論文分類サブタスクでは， 国際特許分類(International Patent Classification: IPC)と呼ばれる分類体系に基づき，学術論文を自動分類するという課題が設定された[79, 49]. このタスクにおいて, Xiaoら[80]は，文書分類における機械学習手法のひとつであるk-NN (k-Nearest Neighbor)法を用いて，任意の論文抄録に対して候補となるIPCコードリストを作成した後， IPCコードリストをリランキングする手法を提案した．Akritidisら[75]は，計算機科学・情報技術分野を対象とした電子ジャーナルサービス，ACM Digital Library[23]で考案された分類体系ACM CCS (Computing Classification System)[24]に基づき，学術論文を自動的に分類する手法を提案した．このタスクにおいて彼らは，研究者欄，収録刊行物欄，キーワード欄といった論文のメタ情報に着目し，SVMを用いて分類を行った．今井ら[81]は，岩波情報科学辞典と呼ばれる，情報科学の分野に特化した索引用語辞典に基づく学術論文の分類手法を提案した．彼らの手法は，論文の表題構造解析に基づいており，「標準化」と「コード割当て」という2つの処理から構成されている．「標準化」では，文字列処理による不要部分の削除・分割を行い，木構造を変形する．その後，単語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，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」では，各部分要素内の専門用語を抽出し，その用語を岩波情報科学辞典の分類コードと対応付け，論文を分類する．上記で述べた研究で使用された分類体系とその研究領域，カテゴリ数，分類手法，論文項目および分類手法で用いられた手掛かり語を表4.1にまとめる．

[23] http://dl.acm.org/
[24] http://www.acm.org/about/class/

表 4.1 学術論文の自動分類における既存研究の概観

研究者	分類体系	研究領域	カテゴリ数	分類手法	論文項目	手掛かり語
Xiao ら [80]	IPC	生活必需品 処理操作 科学，繊維 機械工学 物理学	30,885 件 (第 5 階層)	k-NN 法 リランキング	表題 概要	Bag-of-Words
Akritidis ら[75]	ACM CSS	計算機科学 情報技術	276 件 (第 3 階層)	SVM	研究者欄 収録刊行物欄 キーワード欄	研究者名 学会・雑誌名 キーワード
今井ら [81]	岩波情報科学辞典	情報科学	約 4,500 件	標準化 コード割り当て	表題	表題中の専門用語

Akritidis らの研究では，論文への候補となる研究分野は，その論文を発表した研究者または学会・出版雑誌が扱う研究分野になる可能性が高いという考えに基づいて，研究者名や学会・雑誌名が手掛かり語として用いている．このようなメタ情報を用いた文書分類に関する研究はこれまでにもいくつか行われている[76, 77]．また，言語横断による論文の自動推薦[82]や異種・同種データコレクションからなるデータ群からのトピックの発見[83]など，文書分類以外の分野においても，メタ情報は有用な手掛かりとして活用されている．しかし近年では，工学分野と農学分野の研究者による農業用ロボットの研究開発など， 研究内容が大きく異なる分野間での共同研究が盛んに行われており，特定の研究分野を対象とした学術会議や論文雑誌においても様々な分野の研究課題が扱われることが多くなっている．このような専攻分野が異なる研究者により発表された論文を適切な研究分野に分類する場合，研究者名やその論文を発表した学会・雑誌名だけでは手掛かり語として不十分な可能性がある．Akritidis らは，論文に付与されているキーワードを手掛かり語として用いているが，その数や粒度は論文を執筆した研究者によって異なる．また，キーワードそのものが付与されていない論文も多く存在する．そのため，一般的には論文内容から判断する必要があるが，表題や概要においても，要素技術やその効果といったその論文を特徴付ける重要な手掛かり語が存在する．本論文では，論文内容から要素技術とその効果に関する表現を自動的に発見し，手掛かり語として用いることを行う．そして，研究者名や学会・雑誌名に加えた，文書分類に対する新たな手掛かり語としての有用性を示す．

論文中から要素技術とその効果に関する表現を発見する場合，Xiaoらが用いた

Bag-of-wordsによる方法や，今井らの手法では困難である．Bag-of-wordsは，単語の並び方や係り受け関係などは考慮せず，文書をモデル化する方法である．しかし，文書中のある単語が要素技術または効果を表すものかどうかを判断することはできない．さらに，効果に関する表現は論文表題では記述されないため，今井らの手法を用いて効果表現を発見することはできない．そこで本論文では，3章で提案した機械学習による抽出手法を用いる．

4.3 要素技術と効果を用いた分類手法

4.3.1 提案システム

本論文のタスクは，表4.2に示すようなKAKENの分類体系に基づき，各階層に対して論文を適切な研究分野に自動分類することである．KAKENの分類体系は，表4.1で示したIPCやACM CSS，岩波情報科学辞典のような特定の研究・技術分野を対象とした分類体系とは異なり，人文学，社会科学，生物学，農学，医歯薬学など，幅広い研究領域をカバーしている．なお，以降では，分類体系における「分野」を第1階層，「分科」を第2階層，「細目表」を第3階層と呼ぶ．また，この分類体系への分類対象として，国立情報学研究所(NII)[25]が運営するCiNii article[26]の論文データを用いる．CiNii articleでは，幅広い研究領域で発表された論文を500万件以上収録している．

これまでの文書分類タスクでは，事前に与えられたデータ集合に基づき未知のデータを予測して分類する機械学習手法が多く提案されている．本論文でも同様に，機械学習に基づいた手法を適用し，論文を特徴付けるための表現として，研究者名，学会・雑誌名，および論文表題・概要から形態素解析により得られる単語を用いる．そして，論文表題および概要から得られた単語が要素技術または効果(属性，属性値)を表す表現かどうかを判別し，一般的な単語との区別を行うことで，研究者名，学会・雑誌名に加えた要素技術とその効果による特徴表現の有効性を示す．

ここで，論文の適切な研究分野への自動分類に対する手掛かり語表現の有用度は，それぞれ異なると考えられる．例えば，主に情報科学の分野では，「SVM」が要素技術として用いられることが多いが，「精度が向上」という効果表現は，情報科学の他に工学などの分野でも使用される場合がある．そのため，論文の自動分類において，要素技術は効果よ

[25] http://www.nii.ac.jp/
[26] http://ci.nii.ac.jp/

り有用性が高いと考えられる．しかし，人文社会系の分野の論文では「精度が向上」という効果表現はほとんど使用されない傾向にあるため，一定の有用性はあると考えられる．そこで本論文では，要素技術，属性，属性値を表す表現を収集した重み付き手掛かり語リストを作成し，形態素解析により得られた単語の有用度を判別する．次節で詳細を述べる．

表 4.2 KAKEN の分類体系(2011 年度)の例

系	分野 (第一階層)	分科 (第二階層)	細目表 (第三階層)
総合・新領域系	総合領域	情報学	ソフトウェア，知能情報学
		科学教育・教育工学	科学教育，教育工学
	複合新領域	ナノ・マイクロ科学	ナノ構造科学 ナノクロ・ナノデバイス
		社会・安全システム科学	社会システム工学・安全システム 自然災害科学
人文社会系	人文学	文学	日本文学，英米・英語圏文学
	社会科学	人文地理学	人文地理学
理工系	数物系科学	数学	数学一般(含確率論・統計数学) 基礎解析学，大域解析学
		物理学	物性Ⅰ，物性Ⅱ 数理物理・物性基礎
	化学	基礎化学	物理化学，有機化学，無機化学
		複合化学	分析化学，機能物質化学
	工学	応用物理学・工学基礎	応用光学・量子光工学 応用物理学一般
		材料工学	金属物性，無機材料・物性 材料加工・処理
		総合工学	航空宇宙工学，原子力学
生物系	生物学	基礎生物学	遺伝・ゲノム動態 生態・環境
		生物科学	構造生物化学，機能生物化学
	農学	水産学	水産学一般，水産化学
		畜産学・獣医学	基礎獣医学・基礎畜産学 応用獣医学
	医歯薬学	薬学	化学系薬学，生物系薬学
		内科系臨床医学	消化器内科学，呼吸器内科学
		外科系臨床医学	消化器外科学，耳鼻咽喉科学

4.3.2　手掛かり語の収集方法

本節では，要素技術，属性および属性値を表す手掛かり語の収集方法について述べる．まず，3章で述べた抽出手法を用いて，KAKENの研究課題672,397件の表題および概要に対して<TECHNOLOGY>, <ATTRIBUTE>,<VALUE>タグを付与する．その後，タグ付けされた語句を抽出し，要素技術，属性，属性値リストとしてリスト化する．

上記で述べた要素技術，属性，属性値リストと同様に，本論文では，研究者名および学会・雑誌名を収集したリストの作成を行う．これは，4.2節でも述べたように，近年では研究内容が大きく異なる分野間での共同研究が盛んに行われており，特定の研究領域を対象とした学術会議や論文雑誌においても様々な分野の研究が扱われていることを考慮している．論文の適切な研究分野への分類において，ある特定の研究領域を専門としている研究者や学術会議・論文雑誌は，多岐にわたる研究領域を対象としているものと比べて，有用な手掛かりになると考えられる．本論文では，研究者名および学会・雑誌名に付随する研究分野の数が少ない場合，分野数が多いものより重要な手掛かり語になると考え，研究分野数に閾値を設ける．そして，研究者名および学会・雑誌名において，それぞれ閾値による2種類のリストを構築する．そして，閾値以下の手掛かり語により構築されたリストに対する重みを，閾値より大きい値の手掛かり語から構成されたリストより高くすることで，各研究者名および学会・雑誌名に対する手掛かり語としての有用度を表現する．まず，研究課題における研究者欄から代表者，研究分担者，連携研究者を含むすべての研究者名と，研究課題の発表文献欄における各文献が掲載されたすべての学会・雑誌名を正規表現により抽出する．そして，研究者名または学会・雑誌名と，それらを抽出した研究課題に付与されている(細目表に位置する)研究分野を対応付ける．その後，手掛かり語に付随する研究分野数に対して閾値を設ける．閾値は，4.4節で述べるチューニング方法を用いて1から設定し，最も高い性能を示したときの値を用いる．このとき，閾値より高い値を持つ手掛かり語は重み1を与えることで調整する．これにより本論文では，2分野以下に属する研究者名から研究者リスト1を構築し，それ以外は研究者リスト2として構築した．同様に，9分野以下に属する学会・雑誌名から学会・雑誌リスト1を作成し，それ以外は学会・雑誌リスト2として作成した．

7種類のリストにおける手掛かり語の例，収集した語句の数，各リストに対して与える重みを表4.3に示す．各リストの重みは，1から50までの範囲で決定し，4.4節で述べるチューニング方法により，1または50から1ずつ重みを変更していき，最も性能が高くなるように人手で調整を行った．

さらに本論文では，共同研究者リストを作成する．これは，4.2節で述べたような研究者間における共同研究の状況に加えて，一般的に， 研究者は同じ分野の研究者と共同研究を行う機会が多いことを考慮している．共同研究により発表された研究課題や論文が多いほど，その研究者らは同一の研究分野を専攻している可能性が高く，その分野を特徴付けるための重要な手掛かりとなるといえる[84, 85]．そこで，特定の研究者間において発表された研究課題および論文の数に閾値を設け，閾値以上の値を持つ研究者間を収集したリストを作成する．この共同研究者リストは，論文の適切な研究分野への分類において，論文中の研究者欄に記述されている研究者名と関連する研究者名を手掛かり語として追加する場合に用いる．まず，KAKENの研究者欄における代表者，研究分担者，連携研究者から，各研究者間における発表された研究課題の数を数え，その後，一定の閾値を設定する．閾値は，4.4節で述べるチューニング方法を用いて1から設定し，最も性能を示したときの値を用いる．このとき，閾値未満の値を持つ手掛かり語はリストから除外を行うことで調整する．その結果，研究課題の本数が1本以上ある共同研究者を対象にリストの作成を行う．また，本論文では，KAKENに加えてCiNii articleからの共同研究者リストの作成を行う．CiNii articleでは，共著論文の数が5本以上ある共同研究者を対象とする．本論文では，KAKENおよびCiNii articleからそれぞれ1,094,510対および3,268,625対を獲得した．

表 4.3 各リストに対する重み，手掛かり語数，手掛かり語の例

	重み	手掛かり語の数	手掛かり語の例
研究者リスト 1	50	144,108	荒巻英治
研究者リスト 2	1	15,567	川原稔
学会・雑誌リスト 1	40	125,232	教育情報学会
学会・雑誌リスト 2	4	4,352	情報処理学会
要素技術リスト	14	430,920	SVM, CRF
属性リスト	6	296,152	精度，安定性
属性値リスト	3	70,566	向上，抑制

4.3.3 システム構成

図4.1に本システムの構成を示す．提案システムは，「索引作成モジュール」と「文書分類モジュール」から構成される．以下では，各モジュールについてそれぞれ説明する．

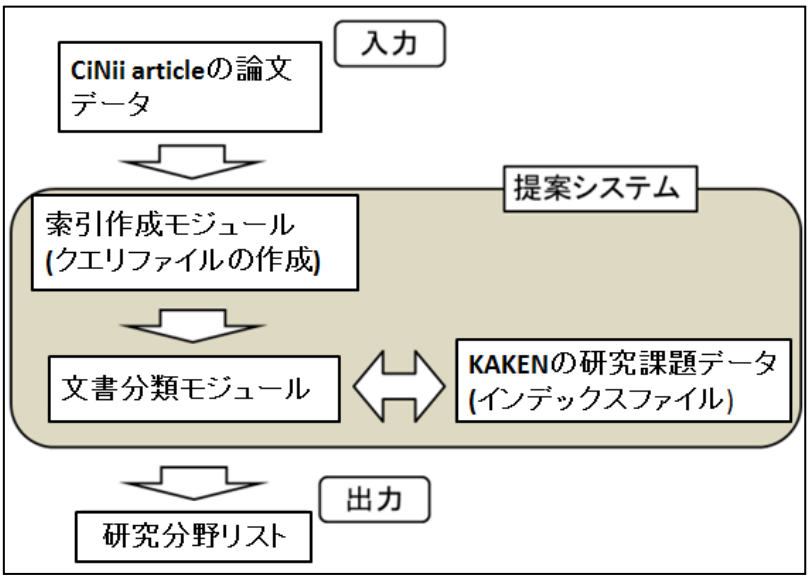


図 4.1 システム構成

索引作成モジュール

索引作成モジュールでは，入力論文からクエリファイル$\vec{q}$を作成する．$\vec{q}$は，論文内における接頭詞を含む名詞$(w_{q1},\dots,w_{ql})$，研究者名$(a_{q1},\dots,a_{qm})$，学会・雑誌名$(p_{q1},\dots,p_{qn})$を用いてベクトル化したデータを格納したものである$(\vec{q}=(w_{q1},\dots,w_{ql},a_{q1},\dots,a_{qm},p_{q1},\dots,p_{qn}))$．このとき，各ベクトルに対して，3.2 節で作成した手掛かり語リストにより重み付けを行う．

まず，名詞への重み付けについて説明する．論文概要に対して形態素解析を行い，抽出した名詞が，表 4.3 で示した要素技術，属性，属性値リストのいずれかに存在していれば，各リストに対応した重みを与える．もし，どのリスト内にも存在していなければ重み 1 を与える．その後，表題に対して形態素解析を行い，抽出した語句が表 4.3 の要素技術リストに存在していれば重み 17 を，存在していなければ 1 を与える．これは，語句が出現する文書内の項目に応じて重みを変えることは有効であることが報告されているためである[86, 87]．なお，重みの決定は，4.4 節で述べるチューニング方法により，要素技術リストの重みから 1 ずつ値を変更し，最も性能の高くなるように調整した．

次に，研究者名および学会・雑誌名への重み付けについて述べる．まず，論文の研究者欄から研究者名を，収録刊行物欄から学会・雑誌名を正規表現により抽出する．その後，抽出した研究者名と関連する研究者名を，KAKEN および CiNii article の共同研究者リストから抽出する．そして，研究者欄および共同研究者リストから抽出した研究者名が表 4.3 で示した研究者リスト 1 または 2 に含まれていれば，リストに対応した重みを与える．また，抽出した学会・雑誌名が学会・雑誌リスト 1 または 2 に含まれていれば，リストに対応した重みを与える．抽出した研究者名または学会・雑誌名がリストに存在しない場合，

その論文の特徴を表す手掛かり語ではないとみなし，クエリファイルの作成に用いない．KAKEN の研究課題からインデックスファイル$\vec{d}$を作成する際も，上記で述べた手順を適用する．このとき，学会・雑誌名は発表文献欄における各文献から抽出する．また，名詞を抽出するとき，表題，概要のほかにキーワードも対象とする．研究課題では，それに関連する技術用語(ツール，モデル)などがキーワードとして選定されており，要素技術となる手掛かり語が多く含まれていると考えられる．まず，各キーワードに対して形態素解析を行い，抽出した名詞が要素技術リストに存在していれば，表 4.3 に従い重み 14 を，そうでなければ重み 1 を与える．また，共同研究者リストは用いない．これは，インデックスファイルの作成において，共同研究者リストを用いた場合と用いない場合よる予備実験の結果から判断した．

<u>文書分類モジュール</u>

文書分類モジュールでは，文書分類タスクにおいて一般的に適用されており，また，表 4.1 で述べた既存研究でも使用されている k-NN 法および SVM 手法という 2 種類の分類手法を用いる．各分類手法について，以下で詳細を述べる．

- **k-NN 法**

k-NN 法とは，訓練データに含まれる情報を用いて，入力文書がどのようなカテゴリに属するのかを自動的に予測するための機械学習アルゴリズムである．k-NN 法は，「類似度計算」と「ランキング」という 2 つのステップから構成される．まず，入力論文と訓練データ内の研究課題との類似度を計算する．その後，類似度に基づき，候補となる上位k件の研究課題を選択する．そして，それらに付与されている研究分野をランキング手法により選択することで，入力論文に対する適切な研究分野を決定する．以下では，類似度計算およびランキング手法について詳細を述べる．

- 類似度計算

まず，汎用連想計算エンジン GETA[27]を用いて，入力論文と研究課題間の類似度を測定する．その後，類似度が高い順に研究課題をソートし，上位k件を選択する．類似度計算には，GETA のライブラリで提供されている SMART[70]を用いる．

[27] http://geta.ex.nii.ac.jp/geta.html

- ランキング

まず，類似度計算により選択された上位k件の研究課題に基づく研究分野のスコア$Score(c)$を計算する．$Score(c)$は，論文に研究分野cが付与される可能性の尺度を表す．そして，$Score(c)$が高い順に研究分野cをソートし，研究分野リストとして出力する．$Score(c)$の算出に対して，本論文では，Xiao らの研究[80]で用いられた 4 種類のランキング手法を適用する．そして，本システムにおいて有用なランキング手法を調べる．

(1) Naïve

Naïve は，最も類似度が高い研究課題に付与されている研究分野から順に決定していくランキング手法である．

(2) Sum

Sum は，研究分野cにおける，入力論文と研究課題間の類似度の総和を算出し，そのスコアに基づいて候補となる研究分野を決定するランキング手法である．以下の式を用いて算出する．

$$Score_{Sum}(c) = \sum_{i=1}^{k} occur(c, d_i)\, Sim(q, d_i)$$

$occur(c, d_i)$は，研究分野cが研究課題d_iに付与されているかどうかを示す関数を表す．もし，付与されていれば 1，付与されていなければ 0 となる．$Sim(q, d_i)$は，入力論文qとd_i間の類似度を表す．

(3) Listweak (List)

Listweak は，上記で述べた Sum 手法に基づくランキング手法であるが，この手法では，入力論文との類似度が低い研究課題はノイズであると仮定し，より類似度の高い研究課題を強調する．以下に述べる式により算出される．

$$Score_{Listweak}(c) = \sum_{i=1}^{k} occur(c, d_i)\, Sim(q, d_i) r_1^i$$

r_1は，より類似度の低い研究課題に対してペナルティを与えるパラメータを表す$(0 < r_1 < 1)$．本論文では，Xiao らの研究に従い，0.95 と設定する．

(4) Weak

k-NN法の欠点として，訓練データ内の各研究分野が持つ研究課題の数に偏りが大きいほど，入力論文に対する研究分野の予測において，その研究分野が選ばれやすくなることが挙げられる．Weakは，このような分野間の偏りを考慮するランキング手法である．スコア

は，以下の式で算出される．

$$Score_{Weak}(c) = \sum_{i=1}^{k} occur(c_i, d_i) Sim(q, d_i) r_2^{crank(c,i) \times \frac{size(c)}{k}}$$

$size(c)$は，上位k件の研究課題集合における研究分野cの数を表し，$crank(c,i)$は，その研究課題集合における上位$i-1$件までのcの出現頻度を表す．r_2は，より類似度の低い研究課題に対してペナルティを与えるパラメータを表す$(0 < r_2 < 1)$．本論文では，Xiaoらの研究に従い，0.90と設定する．

- SVM 手法

次に，SVMを用いた分類手法について述べる．SVM は，2値分類のための教師あり学習アルゴリズムである．マージン最大化による識別平面の決定により高い汎化性能を持ち，様々なソースデータをモデリングする場合において，その柔軟性が優れていることなどから，パターン認識の分野をはじめ，様々な分野で広く用いられてきている．

まず，索引作成モジュールで作成したインデックスファイル集合を用いて，各階層内の研究分野を表す分類器を SVM により作成する．このとき，すべてのインデックスファイル内に存在する名詞，研究者名，学会・雑誌名およびそれらに付随する重みを，SVM で用いる表現および重みとして使用する．次に，各分類器に対して，索引作成モジュールで作成した入力論文を表すクエリファイルを適用する．このとき，クエリファイル内の名詞，研究者名，学会・雑誌名およびそれらに付随する重みを，SVM で用いる表現および重みとして使用する．そして，各分類器において入力論文との超平面間の距離を測り，その距離の値が最も小さかった分類器が表す研究分野から順に出力する．

4.4 実験

4.4.1 実験データ

KAKEN

本論文では，KAKENの研究課題を訓練データとして用いる．このとき，表題，研究概要，キーワード，研究者欄，発表文献欄，および2011年度に使用可能な分類体系(表4.2)において，第3階層に位置する研究分野が付与されている研究課題を対象とする．また， 第3階層

の各研究分野におけるデータ数の偏りを無くすために，1種類の研究分野に対して200件の研究課題を用いる．そして，第3階層に位置する研究分野に基づき，その上位となる第1，第2階層の研究分野を各研究課題に付与する．例えば，表4.2の分類体系に基づき，知能情報学が付与されている研究課題に対して，総合領域，情報学という研究分野を新たに付与する．その結果，28,400件の研究課題および各階層における研究分野のうち，第1階層では10分野，第2階層では44分野，第3階層では142分野を本実験で使用する．ここで，第1，第2階層における各研究分野の研究課題の数には偏りがあることに注意する．

研究者リスト，学会・雑誌リスト，共同研究者リストの構築では，本実験で対象とする研究分野が付与されている283,686件の研究課題データを用いた．また，手掛かり語リストへの重み付けや閾値の決定について，本論文では，訓練データを除く2,000件の研究課題をチューニング用データセットとして作成し，KNN(Sum)手法に基づき， 精度@1(4.4.2節で述べる)において最も性能が高くなるように調整した．

CiNii article

CiNii articleは，2012年までに5,924,679件の論文データを収録しており，それらは主に，表題，概要，研究者欄，収録刊行物欄から構成されている．このうち，概要を含む1,000件の論文データ(Abstデータセット)，および概要を含まない1,000件の論文データ(Titleデータセット)を評価データとして使用する[28]．なお，訓練データ内の発表文献欄に存在する論文データは用いない．また，評価データで扱う研究分野は訓練データで用いているものを対象とし，第3階層における1種類の研究分野に対して20件の論文データ(Abst: 10件，Title: 10件)を用いる．そして，各データに付与されている研究分野の上位となる第1，第2階層の研究分野も付与する．その結果，評価データでは，第1，第2，第3階層において10分野，39分野，100分野を対象とする．なお，共同研究者リストの構築では，すべての論文データ(5,924,679件)を用いた．

4.4.2 評価尺度

システムが評価データに自動付与した研究分野と評価データに付与されている研究分野が一致したとき，正解とする．本論文では，これを精度として評価する．

$$\text{精度} = \frac{1}{K}\sum_{i=1}^{K} t_i$$

[28] CiNii article には，概要が存在しない論文データが約 21%含まれており，表題のみを含む論文を対象とした性能評価は必要であると考えられる．

ここで，Kは，評価データの数を表しており，t_iでは，システムにより付与された上位i件の研究分野のうち，正解となる研究分野が出現していればt_i=1，出現しなければt_i=0を表す．

また，MRR (Mean Reciprocal Rank)による評価も行う．そして，評価データがそれに付与されている研究分野に適切に分類されているかを上位i件までの研究分野候補を見ることで判断する．

本論文では，評価データに自動付与された研究分野のうち，上位 1 件(@1)，2 件(@2)，3 件(@3)までのものを対象として精度を算出する．また，MRR 値の算出では，システムが出力した上位 3 件までの研究分野を対象とする．

4.4.3 比較手法

以下に述べる5種類の提案手法とそれらの手法に対する比較手法を用いて実験を行った．なお，比較手法におけるk-NN法では，実験を通して全体的に精度が高かったB_KNN(List)およびB_KNN(Weak)手法を記載する．また，k-NN 法における上位k件の決定について，チューニングデータセットを用いてkの値を1から50まで1刻みで設定し，各実験条件において最も精度の高かったときの値を用いた．なお，形態素解析にはMeCabを用いた．また，SVMによる機械学習パッケージはTinySVM[29]を用い，カーネル関数は線形カーネルを使用した．

<u>提案手法</u>

- **KNN(Naive)**: 入力論文との類似度が高かった研究課題に付与されている研究分野から順に決定する．
- **KNN(Sum)**: 各研究分野における，入力論文と研究課題間の類似度の総和を算出し，そのスコアに基づいて候補となる研究分野を決定する．
- **KNN(List)**: 各研究分野における，入力論文と研究課題間の類似度の総和を算出し，最もスコアが高かった研究分野から順に決定する．このとき，入力論文との類似度が低い研究課題にペナルティを与える．
- **KNN(Weak)**: KNN(List)に基づく手法であるが，訓練データセットにおける研究分野間の文書数の偏りを考慮する．
- **SVM**: 各分類器において，入力論文に対する超平面の距離を測定し，最も距離が小さい結果を示した分類器が表す研究分野から順に決定する．

[29] http://chasen.org/~taku/software/TinySVM/

比較手法

- **B_KNN(List)**: KNN(List)手法において，要素技術とその効果(属性，属性値)に対応する表現を用いない.
- **B_KNN(Weak)**: KNN(Weak)手法において，要素技術とその効果に対応する表現を用いない.
- **B_SVM**: SVM 手法において，要素技術とその効果に対応する表現を用いない.

4.4.4 実験結果

第1，第2，第3階層の研究分野を対象にしたときの実験結果を表4.4，表4.5，表4.6にそれぞれ示す．Ave.は，各データセット(Title, Abst)を用いて算出した評価値を平均した値(マクロ平均)を示している．なお，すべての実験条件において，合計2,000件のデータセットに対して1種類以上の研究分野が付与された．表4.4から表4.6の結果から，第1階層から第3階層において，出力結果の上位1件までを正解とした場合，KNN(List)手法により，平均で最大0.853，0.712，0.615の精度が得られ，MRRでは，0.909，0.800，0.711の平均値を示した．また，k-NN法およびAbstデータセットを用いたときのSVMの結果から，研究者名および学会・雑誌名のみを手掛かり語として用いた場合では正しく分類できなかった論文を，要素技術とその効果を手掛かり語として加えることで改善できたことが分かる．

次に，提案手法と比較手法において全体的に性能が高かったKNN(List)手法とB_KNN(Weak)手法に対して，t検定による統計的有意差検定を行ったところ，Abstデータセットを対象としたとき，第3階層におけるすべての条件において有意水準1%で有意差があることが確認された．さらに，第2階層における上位2件の研究分野を正解対象としたとき，有意水準5%で有意差が確認された．これらの結果から，本手法における要素技術とその効果を用いることの有効性を示せたといえる．

表 4.4 第 1 階層における研究分野を対象にした場合の精度および MRR の結果

		精度									MRR		
		@1			@2			@3					
		Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.
提案手法	KNN(Naïve)	0.778	0.832	0.805	0.913	0.953	0.933	0.952	0.976	0.964	0.859	0.900	0.880
	KNN(Sum)	0.822	**0.878**	0.850	**0.930**	0.964	**0.947**	0.957	**0.986**	0.972	0.888	0.927	0.908
	KNN(List)	0.827	0.877	0.852	0.925	**0.966**	0.946	0.962	0.984	**0.973**	**0.889**	**0.928**	**0.909**
	KNN(Weak)	**0.828**	0.877	**0.853**	0.927	0.964	0.946	0.958	0.985	0.972	0.888	0.927	0.908
	SVM	0.737	0.815	0.776	0.835	0.892	0.864	0.873	0.938	0.906	0.799	0.869	0.834
比較手法	B_KNN(List)	0.815	0.852	0.834	0.921	0.944	0.933	0.958	0.976	0.967	0.880	0.909	0.895
	B_KNN(Weak)	0.813	0.853	0.833	0.924	0.944	0.934	**0.964**	0.980	0.972	0.882	0.911	0.897
	B_SVM	0.750	0.777	0.764	0.854	0.878	0.866	0.892	0.910	0.901	0.815	0.838	0.827

表 4.5 第 2 階層における研究分野を対象にした場合の精度および MRR の結果

		精度									MRR		
		@1			@2			@3					
		Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.
提案手法	KNN(Naïve)	0.623	0.688	0.656	0.782	0.849	0.816	0.852	0.905	0.879	0.726	0.787	0.757
	KNN(Sum)	0.667	0.753	0.710	0.811	0.871	0.841	0.879	0.918	**0.899**	0.760	0.824	0.792
	KNN(List)	0.676	0.744	0.710	**0.828**	**0.880**	**0.854**	0.872	**0.925**	**0.899**	**0.767**	**0.832**	**0.800**
	KNN(Weak)	0.670	**0.754**	**0.712**	0.810	0.872	0.841	**0.880**	0.916	0.898	0.763	0.829	0.796
	SVM	0.572	0.685	0.629	0.663	0.781	0.722	0.709	0.822	0.766	0.633	0.747	0.690
比較手法	B_KNN(List)	**0.677**	0.715	0.696	0.815	0.844	0.830	0.866	0.902	0.884	0.763	0.800	0.782
	B_KNN(Weak)	0.672	0.717	0.695	0.812	0.853	0.833	0.874	0.911	0.893	0.764	0.805	0.785
	B_SVM	0.591	0.637	0.614	0.710	0.753	0.732	0.764	0.803	0.784	0.669	0.712	0.691

表 4.6 第 3 階層における研究分野を対象にした場合の精度および MRR の結果

		精度									MRR		
		@1			@2			@3					
		Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.	Title	Abst	Ave.
提案手法	KNN(Naïve)	0.523	0.583	0.553	0.693	0.747	0.720	0.758	0.811	0.785	0.630	0.686	0.658
	KNN(Sum)	0.569	0.636	0.603	0.725	0.790	0.758	0.787	0.837	0.812	0.670	0.732	0.701
	KNN(List)	**0.588**	0.641	**0.615**	**0.744**	**0.800**	**0.772**	**0.790**	**0.852**	**0.821**	**0.683**	**0.738**	**0.711**
	KNN(Weak)	0.570	**0.642**	0.606	0.730	0.790	0.760	0.789	0.845	0.817	0.671	0.732	0.702
	SVM	0.527	0.607	0.567	0.634	0.708	0.671	0.666	0.765	0.716	0.591	0.677	0.634
比較手法	B_KNN(List)	0.574	0.603	0.589	0.724	0.733	0.729	0.776	0.795	0.786	0.667	0.690	0.679
	B_KNN(Weak)	0.572	0.607	0.590	0.730	0.739	0.735	**0.790**	0.809	0.800	0.671	0.697	0.684
	B_SVM	0.502	0.561	0.532	0.634	0.654	0.644	0.686	0.715	0.701	0.585	0.628	0.607

4.4.5考察

<u>各研究分野に対する要素技術とその効果の有効性</u>

各研究分野において，要素技術とその効果に関する表現を手掛かり語として用いることで，精度がどのくらい向上したのかについて述べる．ここでは，最も一般的な研究内容を扱う第1階層の10分野を対象に，表4.4における上位1件でのAve.が最も高かったKNN(Weak)手法と，同様のランキング手法を用いているB_KNN(Weak)手法の実験結果を調べた．各手法における詳細結果を表4.7に示す．表4.7では，評価データにおける各研究分野の論文数および正解件数を示している．

表4.7から，工学や化学において精度が向上していることが分かる．特に工学では，Abstデータセットにおける正解件数が231件から245件へと大幅に向上している．これは，3章で提案した抽出システムは元々，理工系の分野を対象に構築されており，研究課題から各分野の特徴となる要素技術とその効果に関する表現を多く抽出できたためと考えられる．

表 4.7 第 1 階層における研究分野ごとの正解件数(上位 1 件)

	工学		社会科学		総合領域		人文学		農学	
	Title	Abst	Title	Abst	Title	Abst	Title	Abst	Title	Abst
KNN(Weak)	**218**/260	**245**/260	**37**/60	**54**/60	**40**/70	**41**/70	**16**/20	**17**/20	68/80	**74**/80
B_KNN(Weak)	213/260	231/260	31/60	48/60	39/70	38/70	12/20	15/20	**71**/80	70/80
	医歯薬学		化学		複合新領域		数物系科学		生物学	
	Title	Abst	Title	Abst	Title	Abst	Title	Abst	Title	Abst
KNN(Weak)	**342**/360	333/360	**26**/30	**21**/30	7/20	12/20	67/80	70/80	7/20	**12**/20
B_KNN(Weak)	339/360	**340**/360	23/30	19/30	7/20	12/20	**68**/80	70/80	**10**/20	10/20

次に，人文学と社会科学の分野に対する正解件数を比較する．表4.7を見ると，要素技術とその効果に関する表現を手掛かり語として用いることで，人文学ではTitleおよびAbstデータセットにおいて正解件数がそれぞれ4件，2件増えており，同様に，社会科学では正解件数がそれぞれ6件増えていることが分かる．ここで，特に正解件数が増加した社会科学において，どのような語句が要素技術または効果であると判断されているのかについて調べた．その結果，「質問紙法」や「情報公開法」などが要素技術，「教育水準」や「回収率」などが効果とみなされていることが分かった．実際にこれらの要素技術が抽出された論文を見ると，主に，研究課題に対する問題を調査・解決するための手段として用いられていた．これらの結果から，本手法は，理工系だけでなく，人文社会系の分野においても，一定の効果があると考えられる．

また，総合領域と複合新領域における実験結果について考察する．総合領域および複合新領域には，人文社会系，理工系，生物系のうち2つ以上の系をまたがる学際的な研究分野が含まれている．表4.7を見ると，複合新領域では，比較手法と提案手法では性能に変化がなかった．しかし，総合領域では，TitleおよびAbstデータセットにおいて，正解件数がそれぞれ1件，3件増えている．これは，複合新領域は，2003年度の分類体系の大幅な改訂において新設された比較的新しい研究分野であり，複合新領域に対する特徴的な要素技術や効果が少なかったためと考えられる．一方で，総合領域では，2003年度以前の分類体系において，複合領域と呼ばれる分野に含まれていた研究分野が多く扱われている．そのため，KAKENにおける長い研究期間において多くの確立された技術や手法が開発され，学際的な研究領域に対する特徴的な知見が得られていたためと考えられる．

要素技術とその効果を用いることの有効性をさらに確かめるために，本論文では，より専門的な研究内容を扱う第3階層の研究分野を対象に，精度がどのように変化したのか調べた．ここでは，表4.6における上位1件でのAve.が最も高かったKNN(List)手法とそれに対応する比較手法であるB_KNN(List)手法の実験結果を比較した．その結果，本手法を用いることで精度が向上した研究分野は42分野であり，精度が低下した分野は16分野であること

が分かった．この結果は，TitleおよびAbstデータセットにおける各手法の比較結果を統合したものから判断している．また，表4.8に，各研究分野に対する詳細結果の一部を示す．上段では，比較手法より精度が向上した研究分野の例を示し，中段では，比較手法より精度が低下した例を示している．下段では，KNN(List)手法を用いたとき，最も精度が低かった研究分野の例を示している．ここで，B_KNN(List)手法より精度が低下した研究分野(無機材料・物性，基礎獣医学・基礎畜産学，細菌学(含真菌学))およびKNN(List)手法において最も精度が低かった研究分野(応用物理学一般，神経科学一般，生物系薬学) について詳しく見ていく．本論文では，論文に対してシステムが誤って付与した研究分野の傾向について調べた．上記の研究分野が付与されている評価データに対して，KNN(List)手法により誤って付与された研究分野の例を表4.9に示す．括弧内の数値は，システムが誤って付与した研究分野における論文の数を表す．

表 4.8 第 3 階層における研究分野毎の精度(上位 1 件)

KNN(List)手法とB_KNN(List)手法を　比較して精度が向上した研究分野の例						
	応用光学・量子光工学		衛生学		会計学	
	Title	Abst	Title	Abst	Title	Abst
KNN(List)	**0.60**	**1.00**	**0.70**	**0.60**	**0.80**	**0.80**
B_KNN(List)	0.40	0.70	0.50	0.30	0.50	0.70
KNN(List)手法とB_KNN(List)手法を　比較して精度が低下した研究分野の例						
	無機材料・物性		基礎獣医学基礎畜産学		細菌学(含真菌学)	
	Title	Abst	Title	Abst	Title	Abst
KNN(List)	0.40	0.20	0.10	0.40	0.50	0.60
B_KNN(List)	**0.60**	**0.30**	**0.30**	0.40	**0.60**	**0.70**
KNN(List)手法において最も精度が低かった研究分野の例						
	応用物理学一般		神経科学一般		生物系薬学	
	Title	Abst	Title	Abst	Title	Abst
KNN(List)	0.10	0.30	0.00	**0.30**	0.00	**0.20**
B_KNN(List)	0.10	0.30	0.00	0.20	0.00	0.10

表 4.9 KNN(List)手法においてシステムが誤って付与した研究分野の例

Correct field	B_KNN(List)と比較して精度が低下した研究分野（表 4.8 中段）	
	Title	Abst
無機材料・物性	● 金属物性（1） ● 材料加工・処理（1） ● 応用物性・結晶工学（1）	● 金属生産工学（2） ● 金属物性（1） ● 材料加工・処理（1）
基礎獣医学・基礎畜産学	● 応用獣医学（4） ● 応用動物科学（2） ● 解剖学一般（1）	● 応用獣医学（4） ● 畜産学・草地学（1） ● 農業土木学（1）
細菌学(含真菌学)	● 耳鼻咽喉科学（2） ● 応用獣医学（1） ● 無機材料・物性（1）	● 病態検査学（2） ● ウイルス学（1） ● 土木環境システム（1）
Correct field	KNN(List)手法において最も精度が低かった研究分野（表 4.8 下段）	
	Title	Abst
応用物理学一般	● 原子力学（2） ● 機械力学・制御（2） ● 流体工学（1）	● 応用光学・量子光工学（3） ● 原子力学（1） ● 航空宇宙工学（1）
神経科学一般	● 計測工学（4） ● 耳鼻咽喉科学（1） ● 麻酔・蘇生学（1）	● 環境生理学（2） ● 神経・筋肉生理学（2） ● 脳神経外科学（1）
生物系薬学	● 呼吸器内科学（2） ● 生理学一般（1） ● 薬理学一般（1）	● 循環器内科学（2） ● 構造生物化学（2） ● 薬理学一般（2）

まず，無機材料・物性の結果を見ると，金属生産工学，金属物性，材料加工・処理など，工学関連の様々な研究分野が誤って付与されていることが分かる．同様に，細菌学(含真菌学)では耳鼻咽喉科学や病態検査学，生物系薬学では呼吸器内科学や循環器内科学など，それぞれ医歯薬学に関連する研究分野が誤って付与されていることが分かる．一方で，基礎獣医学・基礎畜産学の論文に誤って付与された研究分野を見ると，主に応用獣医学であると判断されていることが分かる．同様に，応用物理学一般の論文には，応用光学・量子光工学が誤って付与されていることが分かる．なお，基礎獣医学・基礎畜産学と応用獣医学は，分類体系の第2階層における畜産学・獣医学に属しており，応用物理学一般と応用光学・量子光工学は，第2階層の応用物理学・工学基礎に属している．また，各分野間の研究内容は比較的類似している．

本論文では，基礎獣医学・基礎畜産学，応用獣医学，応用物理学一般および応用光学・量子光工学において，訓練データ内でどのような語句が要素技術または効果として用いられているのか調べた．表4.10に，4種類の研究分野における要素技術とその効果の例を示す．括弧内の数値は，訓練データ内において要素技術または効果が出現した研究課題数を示す．表4.10から，基礎獣医学・基礎畜産学と応用獣医学では，要素技術を表す用語として「マウ

ス」や「ウイルス」が，効果を表す用語として，「細胞」や「活性」がそれぞれ主に用いられていることが分かった．また，応用物理学一般と応用光学・量子光工学では，「レーザー」や「レンズ」が要素技術として，「特性」や「周波数」が効果として頻出していることが分かった．これらの結果から，研究内容が類似している分野間の特徴をより詳細に捉えるために，各研究分野における要素技術と効果の出現傾向を考慮した類似度計算を考案するといった新たな枠組みが必要と考えられる．

表 4.10 第 3 階層の研究分野において抽出された要素技術とその効果の例

基礎獣医学・基礎畜産学		応用獣医学	
要素技術	効果	要素技術	効果
ラット（65）	細胞（101）	血清（79）	細胞（106）
マウス（60）	活性（87）	マウス（66）	反応（93）
アミノ酸（38）	遺伝子（87）	リンパ（45）	成績（80）
ウイルス（37）	濃度（52）	ウイルス（43）	活性（70）
応用物理学一般		応用光学・量子光工学	
要素技術	効果	要素技術	効果
顕微鏡（58）	特性（88）	レーザー（87）	特性（114）
レーザー（40）	成果（83）	半導体（73）	波長（105）
半導体（37）	試料（79）	ファイバ（34）	成果（79）
レンズ（21）	周波数（58）	レンズ（31）	周波数（57）

<u>本手法の分類精度に対する要素技術とその効果の有効性</u>

次に，要素技術とその効果に関する表現をそれぞれ単独に用いた場合，どのように分類精度が変化するのか調べる．ここでは，表4.4から表4.6において，全般的に高い精度とMRRを示しているKNN(List)手法を対象に調査する．B_KNN(List)手法に対して，表題中の要素技術，概要中の要素技術および概要中の効果を表す表現をそれぞれ単独に加えたときの結果を表4.11，表4.12，表4.13に示す．各表は，Abstデータセットを用いて実験を行い，第1，第2，第3階層の研究分野を対象としたときの結果を記載している．また，KNN(List)およびB_KNN(List)手法の結果も記載する．各表から，概要中の効果を表す表現のみを用いた場合，ベースライン手法と比べて精度およびMRR値は上回っているが，一方で，要素技術を表す表現のみを用いた場合，ベースライン手法とほとんど変わらない結果を示す場合があることが分かる．しかし，要素技術とその効果を表す表現を手掛かり語としてすべて用いることで，効果表現のみを用いた場合の性能から，さらに上回る結果を示すことが分かった．本論文では，表題および概要中の語句が要素技術，属性または属性値であるかどうかを決定するために，3章で提案した情報抽出システムを用いて<TECHNOLOGY> ，

<ATTIBUTE> , <VALUE>タグをKAKENの研究課題に付与し，タグ付けされた語句の抽出・リスト化を行っている．そして，その語句が各リストのいずれかに存在するかどうかを調べることで手掛かり語として判定している．しかし，本論文で用いた分析システムの精度は決して高いとはいえない．そのため，要素技術または効果でない語句を誤って手掛かり語として抽出してしまい，その結果，誤った重みを与えた可能性がある．

実際に構築した各リストを調べると，要素技術リストでは，「導入」「解明」「再構築」など，属性または属性値である方がふさわしいと考えられる語句や，「予定」「観点」「こと」など，手掛かり語としてふさわしくない語句も含まれていた．同様に，属性リストでも，「ナノコンポジット」「支援」「それぞれ」など，要素技術，属性値またはそれらのいずれにも属さないと考えられる語句が含まれていた．属性値リストにおいても，「ニュース」「レシピエント」「スイッチ」「ビジュアル」など，属性値とは考えにくい用語が含まれていた．しかし，各リスト内における，これらの語句をタグ付けしている文書の数は低い傾向にある．実際，要素技術リストにおいて，「熱処理」という語句は870文書から抽出されているのに対し，「予定」や「再構築」といった要素技術とは考えにくい用語は，それぞれ3文書，13文書のみから抽出されていた．この問題は，技術動向分析システムの精度向上や，各リスト内の手掛かり語に対して一定の閾値を設定するといった処理により改善すると考えられる．

表 4.11 Abst データセットにおける表題中の要素技術, 概要中の要素技術および概要中の効果表現を単独で加えた場合の精度および MRR の結果（第 1 階層）

B_KNN(List)に追加する素性	精度			MRR
	@1	@2	@3	
ALL (KNN(List))	**0.877**	**0.966**	0.984	**0.928**
TITLE_TECHNOLOGY	0.856	0.939	0.977	0.910
ABST_TECHNOLOGY	0.855	0.958	0.979	0.912
ABST_EFFECT (ATTRIBUTE and VALUE)	0.870	0.960	**0.985**	0.923
NOTHING (B_KNN(List))	0.852	0.944	0.976	0.909

表 4.12 Abst データセットにおける表題中の要素技術，概要中の要素技術および概要中の効果表現を単独で加えた場合の精度および MRR の結果（第 2 階層）

B_KNN(List)に追加する素性	精度			MRR
	@1	@2	@3	
ALL (KNN(List))	**0.744**	**0.880**	**0.925**	**0.832**
TITLE_TECHNOLOGY	0.719	0.843	0.895	0.800
ABST_TECHNOLOGY	0.720	0.865	0.914	0.811
ABST_EFFECT (ATTRIBUTE and VALUE)	0.737	0.875	0.917	0.820
NOTHING (B_KNN(List))	0.715	0.844	0.902	0.800

表 4.13 Abst データセットにおける表題中の要素技術，概要中の要素技術および概要中の効果表現を単独で加えた場合の精度および MRR の結果（第 3 階層）

B_KNN(List)に追加する素性	精度			MRR
	@1	@2	@3	
ALL (KNN(List))	**0.641**	**0.800**	**0.852**	**0.738**
TITLE_TECHNOLOGY	0.613	0.741	0.802	0.698
ABST_TECHNOLOGY	0.609	0.761	0.820	0.705
ABST_EFFECT (ATTRIBUTE and VALUE)	0.631	0.775	0.830	0.724
NOTHING (B_KNN(List))	0.603	0.733	0.795	0.690

4.5 要素技術とその効果を利用した論文検索システム

本章では，提案手法を用いて構築したシステム，CiNii Miningついて説明する．本システムでは，CiNii論文検索API[30]から取得した論文データを対象に自動分類している．図4.2は，「音声認識」をクエリとして入力したときの検索結果を示している．このとき，本手法を用いて計算されたスコアが最も高かった研究分野に基づいて学術論文を分類している．

[30] http://support.nii.ac.jp/ja/cia/api/a_opensearch

また，近年において，KAKENの分類体系では網羅されていないような学際的な研究が増加していることを考慮し，最もスコアが高かった研究分野の他に，スコアが2番目および3番目に高かった研究分野をその他の研究分野候補として提示している．図4.2において，「音声認識」というキーワードで検索された論文が，知能情報学や教育工学など，様々な研究分野に分類されていることが分かる．また，本システムは，KAKENの分類体系に従い，論文の分類レベルを選択することができる．

図 4.2 クエリ「音声認識」を入力したときの検索画面結果

4.6 まとめ

本章では，研究領域全般を横断した学術論文の自動分類手法を提案した．本手法は機械学習に基づいており，研究者名や学会・雑誌名に加え，論文中で記述されている要素技術とその効果に関する表現を手掛かり語として用いた．

KAKENの分類体系である「分野・分科・細目表」を対象に評価実験を行った結果，各階層における上位1件の結果に対して，KNN(List)手法により，それぞれ平均0.853，0.712，0.615の精度が得られた．また，上位3件までの出力結果に対して，同様の手法により，平均で0.909，0.800，0.711のMRR値が得られた．これらの結果は，要素技術とその効果に関する表現を手掛かり語として用いない場合より高い値を示していることから，本手法の有効性が確認された．

第5章　要素技術と効果に基づいた技術動向情報の分析

ある技術分野において，「どのような要素技術がいつ頃から使われており，どのような効果が得られているのか」という情報を網羅的に収集し整理することは，その分野の技術動向を概観するのに必要不可欠である．しかし，このような動向調査は膨大な文書を対象に行われるため，分析作業に多くの時間と労力を要する．そこで本論文では，3 章で提案した情報抽出手法を用いて，論文と特許を対象に，特定分野の技術動向を効率的に把握するための可視化システムの構築を行う．このシステムにより，特定の分野を中心とした要素技術とその効果の変遷を効率的に知ることができる．

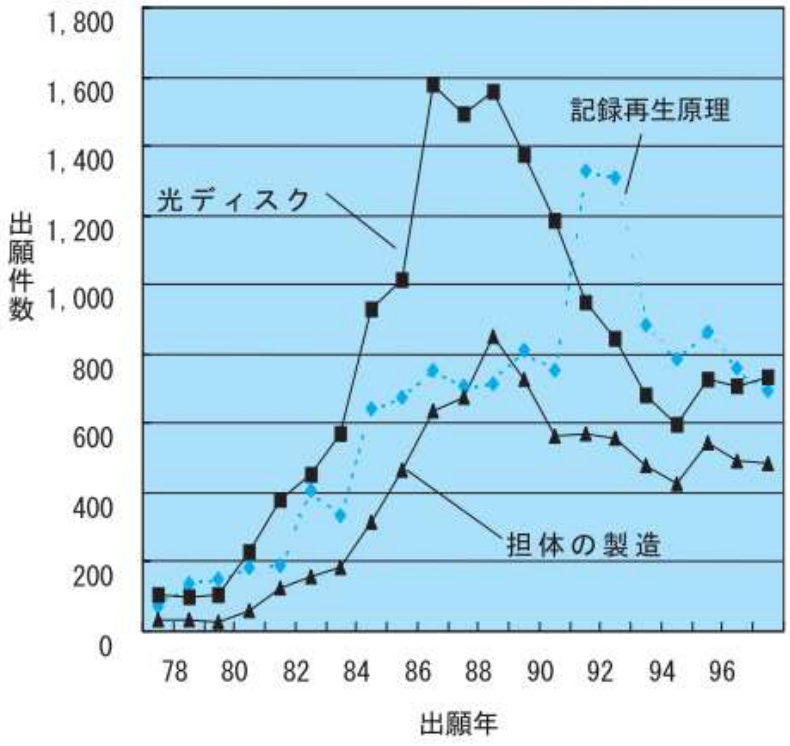
まず，5.1 節では，動向情報の可視化を行う上で基本となる出力形式について紹介する．5.2 節では，技術動向の可視化に関する関連研究をいくつか述べる．このとき，本論文と最も関連のある研究プロジェクトである NTCIR-7,8 特許マイニングタスク[79, 49]についての概要を述べる．そして，5.3 節では，本論文で構築した技術動向分析システムを示し，5.4 節で本章をまとめる．

5.1　動向情報の可視化

現在では，特許を中心とした多くの技術動向分析システムが構築されており，様々な可視化方法で分析結果の提示が行われている．技術動向マップの作成には，大きく分けて 2 種類の方法がある．一つ目は，技術別出願件数推移 (図 5.1)や，ある技術に対する出願人別公開件数 (図 5.2)といった，書誌事項などを用いて解析し，グラフで可視化する方法がである．図 5.1 のようなグラフは，検索キーワードにヒットした特許集合から，出願年毎に特許件数を数えることで作成することができる．また，出願人という特許項目を利用し，出願人別に特許件数を数えることで図 5.2 のようなグラフを作成できる．そして利用者は，膨大な文書集合から，特定の分野やキーワードに対する傾向や法則といった動向を，これらのグラフにより効率的に把握することができる．なお，論文を対象に図 5.1 や図 5.2 のようなグラフを作成する場合，論文の発行年や著者名，学会・雑誌名などの項目を用いることで作成することができる．

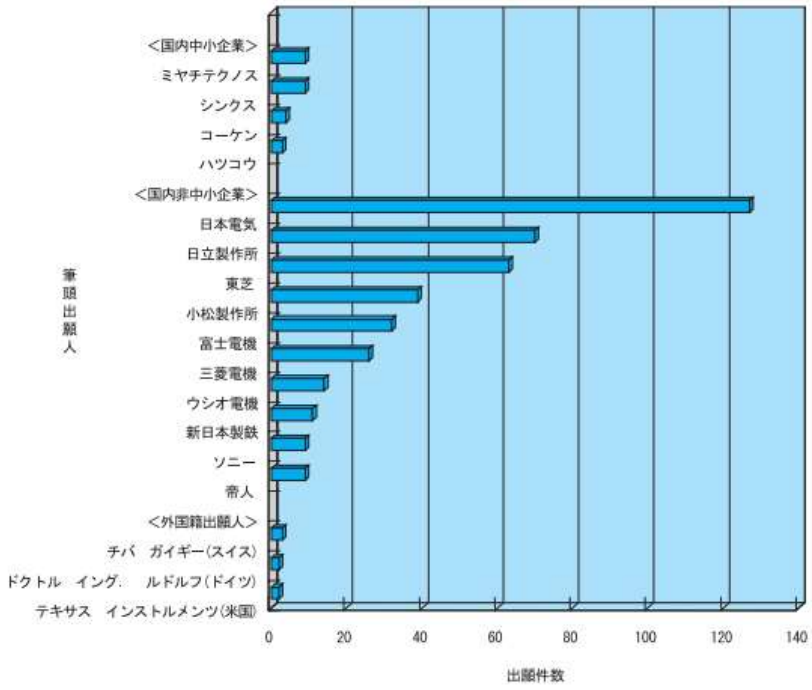
二つ目は，文書内容を解析することで，観点別に技術動向を提示する方法である．特許庁が公開している「技術分野別特許マップ 活用ガイドマップ」[8]で挙げられている例(図 5.3, 図 5.4)を参照すると，「方法・手段－目的・効果」や「技術開発課題－解決手段」といった，文書内容を理解しなければ得られないような情報を軸とした技術動向の可視化により，課題解決に必要な技術の把握や，目標達成への難易度などを見積もることができる．同様に，論文を対象として図 5.3 や図 5.4 のような技術動向マップの作成する場合，特定の分野における技術の将来性や研究の方向性の決定などに活用することができる．

図 5.1 から図 5.4 のような動向マップの作成は，「技術分野別特許マップ 活用ガイドマップ」によると，特許情報解析のエキスパートと各技術の専門家による共同作業により行われており，各技術に対して 2～3 万件の特許文書を対象としていると述べられている．このように，技術動向マップの作成には人手による多大なコストを要している．そのため，技術動向マップの自動作成は，分析作業の効率向上には欠かせない最も重要な課題である．



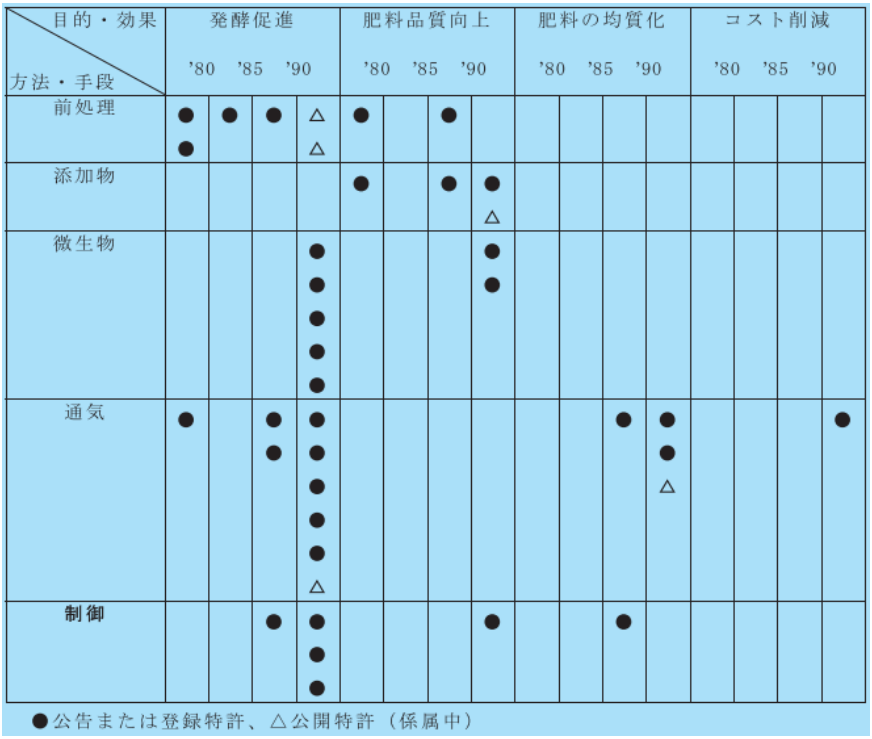
「技術分野別特許マップ 活用ガイドマップ」より抜粋

図 5.1 技術別出願件数推移の可視化例（光ディスクの主要技術分野別出願件数推移）



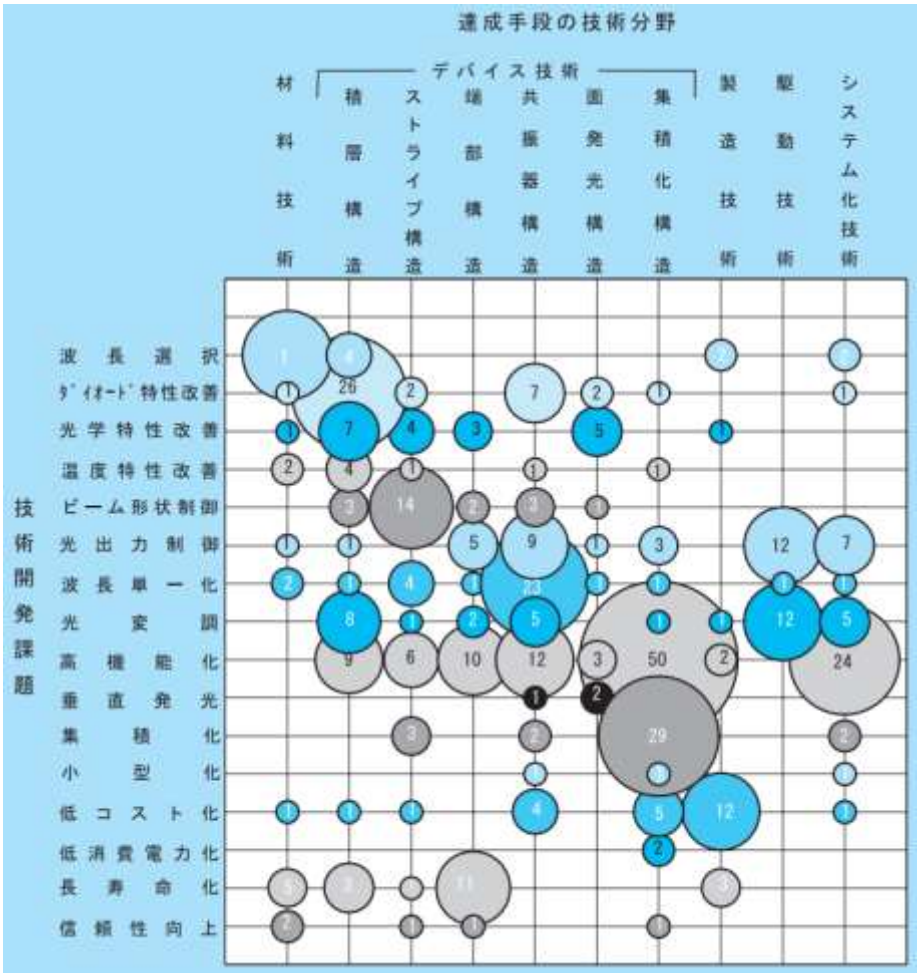
「技術分野別特許マップ 活用ガイドマップ」より抜粋

図 5.2 出願人別特許件数の可視化例（レーザマーキングの主要出願人）



「技術分野別特許マップ 活用ガイドマップ」より抜粋

図 5.3 観点別技術動向マップの例（農業廃棄物の肥料化処理における代表的特許の目的・効果と改良技術）



「技術分野別特許マップ 活用ガイドマップ」より抜粋

図 5.4 観点別技術動向マップの例（半導体レーザにおける技術開発課題と達成手段）

5.2 技術動向分析に関する研究

本論文では，「どのような要素技術がいつ頃から使われており，どのような効果が得られているのか」という技術内容に関する情報を解析することで，特定の分野における要素技術とその効果の変遷を分析している．そのため，以下では，文書を解析することで技術動向マップの作成を行うシステムや研究に焦点を当て，紹介する．

5.2.1　企業が提供している検索・分析システム

5.1 節でも述べたように，近年，技術動向分析システムの開発や整備が急速に進んでおり，特に，企業を中心に様々な可視化ツールが発表されている．表 5.1 に，企業が提供している分析システムの概要を示す．

ATMS/Analyzer[31]では，大量の特許に埋もれている関係性や特徴を，スケルトンマップにより可視化する機能を提供している．このスケルトンマップは，特許から抽出された単語間の関係性をネットワーク状で表しており，単語が出現する特許の数により文字を大きくし，2 つの単語が同時に出現する特許数に応じて線の太さを変更している．また，スケルトンマップでは，単語以外の情報を組み合わせることで拡張することができる．例えば，各企業が何を対象として特許出願をしているのかを分析したい場合，企業名をマップに追加することで，どの企業が何を得意としているのか，また，どの企業が競合しているのかを知ることができる．また，特許が何を課題とした技術であるかと，何を対象とした技術であるかを明細書から自動的に抽出し，それらの関係性をスケルトンマップで表すことができる．このマップにより，自社では気づいていない課題やまだ取り組んでいない課題を効率的に分析することができる[32]．

Biz Cruncher[33]では，分析対象技術の研究開発動向を，その特許群の課題と解決手段からマトリックス分析により可視化する機能を提供している．各セルは，特許の出願件数に応じてヒートマップ上に表示される．課題に関する表現は，要約書に記載された【発明が解決しようとする課題】から，解決手段に関する表現は，特許請求項からキーワード抽出により収集している[34]．

CsvAid[35]では，出願にあたり，他の発明者や企業がどのような開発課題に取り組んだかを把握することが本来の技術調査の基本情報であると考え，特定の課題に取り組んでいる特許の年代別出願件数を可視化する機能を提供している．このような時間推移でどのような動向があるのかを確認することができれば，特許分析のプロセスにおいて大きな判断材料となる[36]．

Text Mining Studio[37]は，構文解析により得られた単語間の係り受け情報や単語の頻度情報を分析し，属性(性別，年代)別にその傾向を可視化する機能を提供している．主に，推移

[31] http://www.fujitsu.com/jp/solutions/industry/manufacturing/ip/
[32] https://www.jpo.go.jp/shiryou/s_sonota/pdf/kigyou/fujitsu.pdf
[33] http://www.bizcruncher.com/
[34] https://www.jpo.go.jp/shiryou/s_sonota/pdf/kigyou/patentresult.pdf
[35] http://www.cks.co.jp/home/product-2-2.htm
[36] https://www.jpo.go.jp/shiryou/s_sonota/pdf/kigyou/chuou.pdf
[37] http://www.msi.co.jp/tmstudio/index.html

マップやバブルチャートなどにより，属性別の頻度情報を可視化している．また，「属性－単語」や「単語－単語」間の関連性の強さを共起確率により解析し，ネットワーク図として図示する機能も実装しており，関連の強いもの同士のクラスタなどを人手で効率的に分析することができる．

TrueTeller パテントポートフォリオ[38]では，特許中に出現する単語の共起情報を用いて単語間の距離を自動的に計算し，コレスポンデンス分析やマトリックス分析により可視化する機能を提供している．また，この単語マップをもとに，その単語が出現する特許数の多さをサーモグラフにより温度表示することで，より直感的な可視化を実現している．このとき，企業別に可視化結果を提示する機能を実装しており，各企業が保有する特許群の強みや弱みを効率的に解析することができる．

表 5.1 企業が提供している分析ツールの概観

システム名	提供	概要	主な可視化方法
ATMS/Analyzer	富士通株式会社	スケルトンマップと呼ばれるネットワーク分析により，単語間や観点間などの関係性を可視化	スケルトンマップ
Biz Cruncher	株式会社パテントリザルト	分析対象技術の研究開発動向を，その特許群の課題と解決手段から分析	マトリックス分析
CsvAid	中央光学出版株式会社	技術課題を軸に特許を整理して可視化	マトリックス分析
Text Mining Studio	株式会社NTTデータ 数理システム	係り受け頻度や特徴語抽出などにより，属性毎の傾向を分析	推移マップ マトリックス分析 ネットワーク分析
TrueTeller パテントポートフォリオ	野村総合研究所	単語同士の出現情報を用いて単語間の距離を分析し，配置の距離を自動計算 単語が出現する特許の数をサーモグラフで可視化	サーモグラフ分析 コレスポンデンス分析 マトリックス分析

[38] http://www.trueteller.net/patent/

5.2.2　技術動向分析に関する従来研究

技術動向の分析および可視化に関連する研究の概要を表 5.2 にまとめる．以下では，表 5.2 における従来研究について述べる．

安藤[88]は，膨大な特許情報の中からエンドユーザにとって必要な情報を迅速に抽出して分析するための枠組みとして，オープンソースのテキストマイニングツール termmi[39]と統計解析言語 R[40]を組み合わせた可視化方法を検討した．termmi は，Windows 用テキストマイニングツールであり，複数の文書間の特徴語を重要度付きで抽出することが可能であり，ベクトル空間法による文書間の類似度を計算する機能を実装している．安藤は，これらのツールを用いていくつかの特許の分析方法を紹介しているが，ここでは，重み付けと類似度計算によるクラスタリング方法について述べる．分析手順として，まず，特許公開公報からテキストマイニングツール termmi を用いて特徴キーワードを抽出し，コサイン類似度による文書間の相互類似度計算を行う．次に，文書間の類似度を距離に変換し，多次元尺度法により文書間の相対的位置関係を求める．そして，統計解析言語 R を用いて 2 次元平面上にプロットする．このとき，類似している文書ほど距離が近くにプロットされる．図 5.5 に，テキストマイニングツールと R を用いて可視化した例を示す．図 5.5 の左側が特徴キーワードの分析結果を示しており，中央と右側は，特徴キーワードから出願人を表すものに対して重みを与えたときの分析結果を示している．中央と右側において，出願人に与える重みを変更しており，重みを大きくすることで，特定の出願人が一箇所に集中する傾向があることが分かる．なお，出願人だけでなく，分析対象としている観点や項目に重みを与えることで，同様の分析を行うことができる．

西山ら[55]は，特許公開公報と新製品発表データから 2.3.2 節で述べた手掛かり語に基づく抽出手法により特徴表現を自動的に抽出し，ユーザに提示するシステムを構築している．このシステムを Advantage Phrase Annotator と名付けており，より斬新な技術応用を示している可能性の高いものを上位に配置し，リストの形式にして出力している．図 5.6 に，西山らが提案した技術動向分析システムの概要図を示す．

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，pLSA (Probabilistic Latent Semantic Analysis)[89]やLDA (Latent Dirichlet Allocation)[90]など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，あ

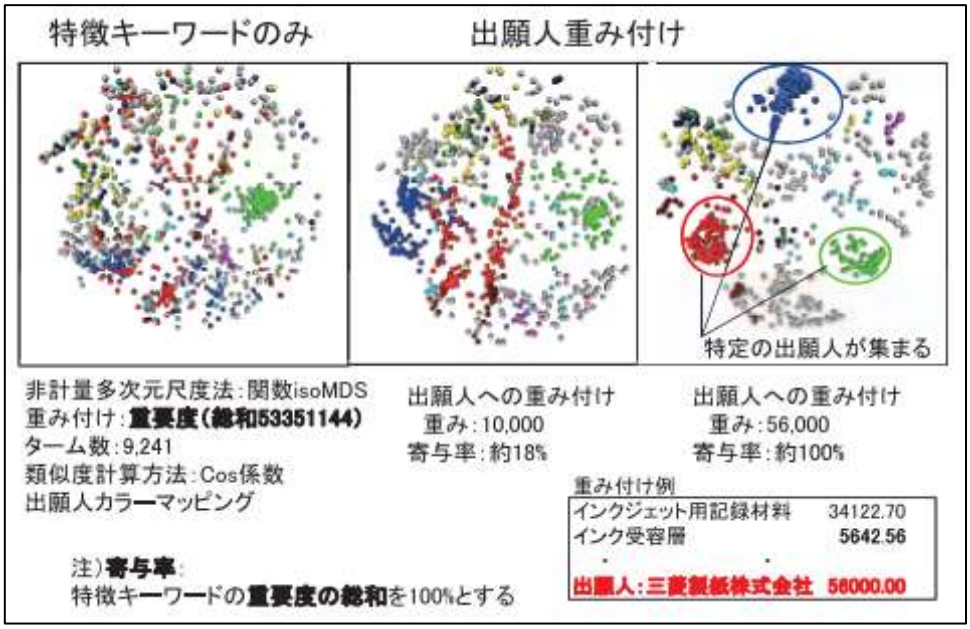
[39] http://gensen.dl.itc.u-tokyo.ac.jp/termmi.html
[40] https://www.r-project.org/

るトピックに対して特徴的な語句を抽出することができる．情報科学の分野では，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[91, 92, 93]．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[94]や著者情報[95, 96, 97]など，多様な情報と統合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重要なメタ情報などを考慮することができる．

岩田ら[98]は，Probabilistic Latent semantic Visualization (PLSV)という離散データの非線形可視化のためのトピックモデルを提案した．彼らは，文書およびトピックが2次元または3次元ユーグリッド可視化空間に座標を持つと仮定し，それらの座標から文書が生成される過程をモデル化することにより文書群を可視化している．モデル推定にはEMアルゴリズムを使用している．PLSVの特徴として，分析対象となる文書とともに画像や時間など他の情報も同時に可視化することができる．図5.7に，PLSVによる映画評点データの可視化結果を示す．このとき，いくつかの映画タイトルを示している．図5.7から，同じジャンルの映画は近くに配置されていることが分かる．例えば，クラシック映画は右下部に配置されており，外国映画は上部に配置されている．

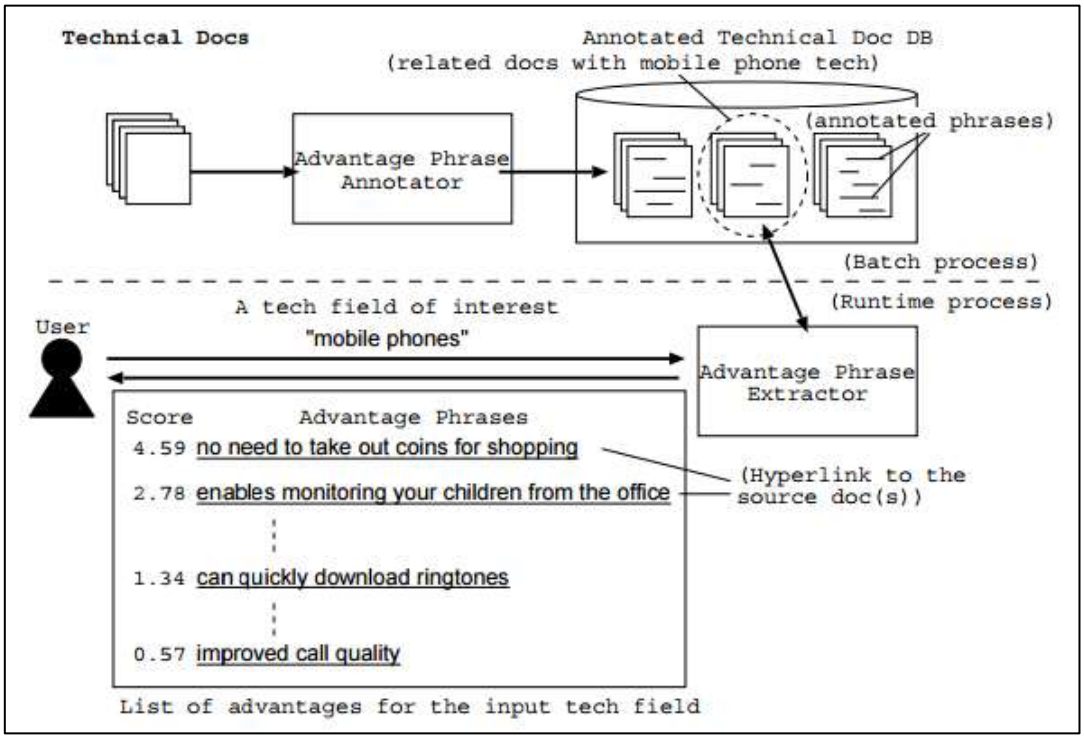
表 5.2 技術動向の分析および可視化に関連する従来研究の概観

研究者名	対象文書	分析手法
安藤[88]	特許明細書	テキストマイニング， 統計解析言語 R
西山ら[55]	特許明細書， 製品発表データ	パターンマッチ
岩田ら[98]	論文，ニュース記事， 映画評点データ	トピックモデル



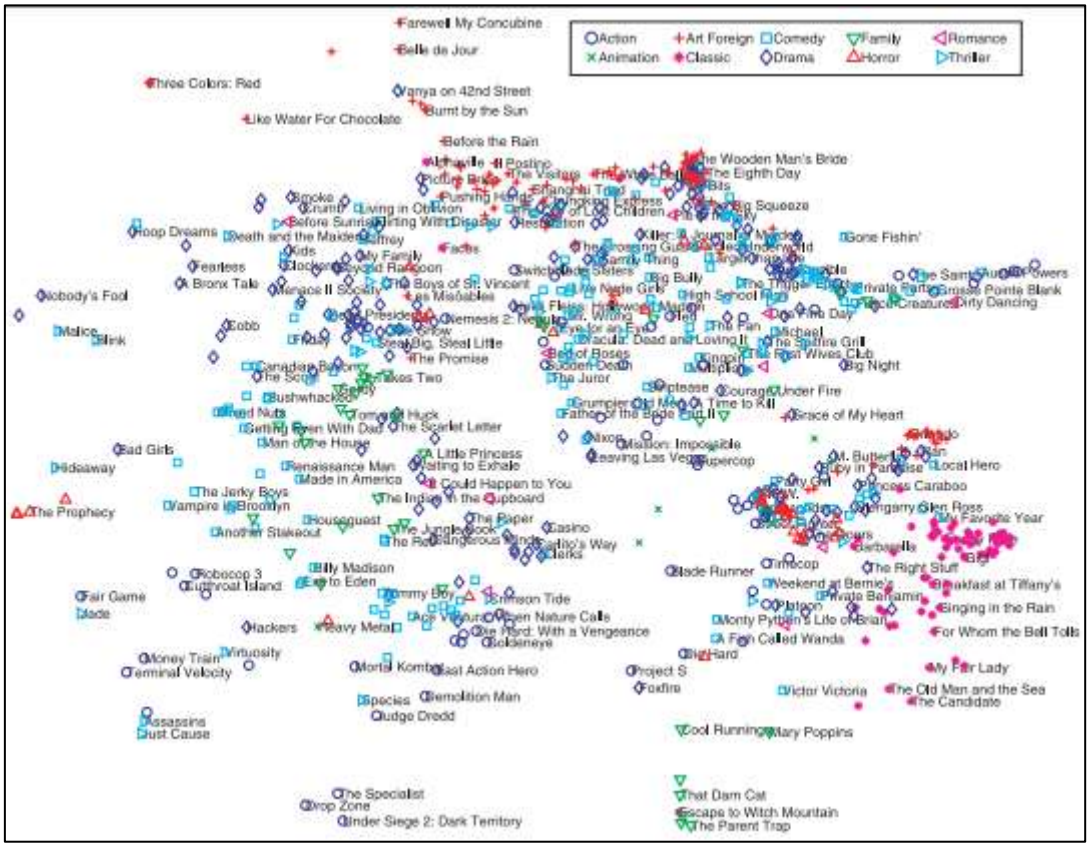
(安藤, 2009) [88]より抜粋

図 5.5 テキストマイニングツールと R を用いた可視化例



(西山, 2009) [55]より抜粋

図 5.6 西山らが提案した技術動向分析システムの概要図



(岩田, 2009) [98]より抜粋

図 5.7 PLSV による映画評点データの可視化結果

5.2.3　技術動向分析に関する研究プロジェクト

論文と特許を対象にした技術動向分析に関する研究プロジェクトとして，国立情報学研究所が主催した第 7 回，および第 8 回 NTCIR ワークショップ(NTCIR-7, NTCIR-8)で実施された特許マイニングタスクがある[79, 49]．このタスクでは，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 コードを自動的に付与し，特定の分野に分類された論文と特許から, 図 5.8 に示すような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生成するツールは，先行技術調査や無効資料調査の支援ツールとして活用できる．

	効果 1	効果 2	効果 3
要素技術 1	[論文 AAA] [特許 XXXXX]		[論文 BBB]
要素技術 2	[論文 CCC]		
要素技術 3		[特許 YYYYY]	[特許 WWWWW] [特許 ZZZZZ]

図 5.8 技術動向マップの作成例

図 5.8 のようなマップを作成するために, 特許マイニングタスクでは以下の二つのサブタスクを設定している.

(1) 学術論文分類サブタスク

(2) 技術動向マップ作成サブタスク

学術論文分類サブタスクでは，特定の分野の特許と論文を網羅的に収集することを目的としており，技術動向マップ作成サブタスクでは，(1)で収集された特許と論文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ことを目的としている．図 5.9 に特許マイニングタスクの概観をまとめる．以下では，各サブタスクの概要について述べる．

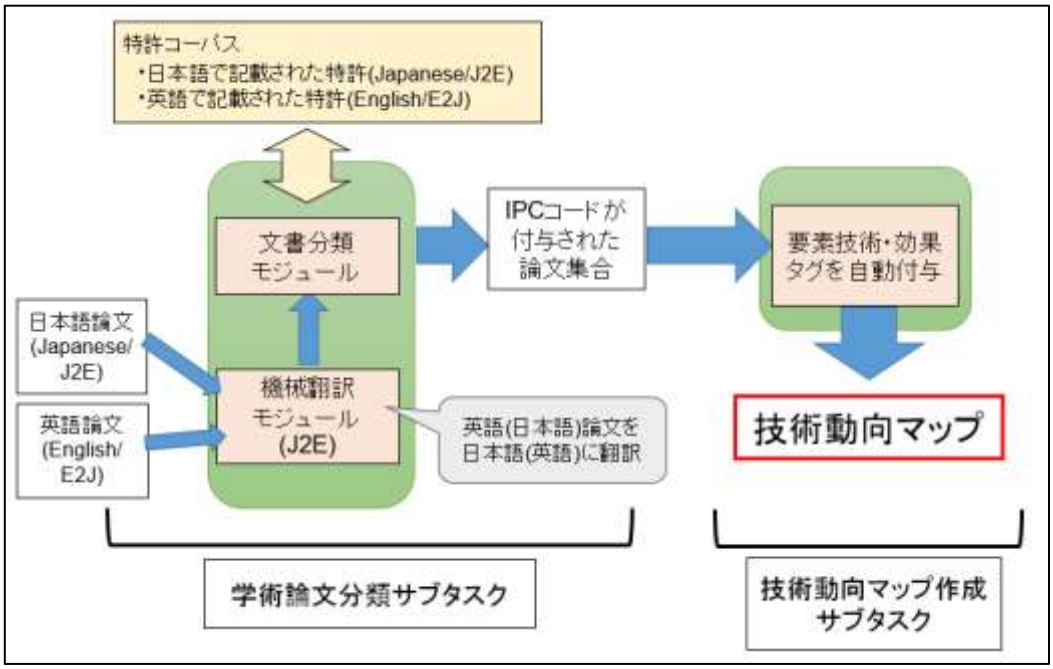


図 5.9 特許マイニングタスクの概観

学術論文分類サブタスク

前述したように，学術論文分類サブタスクは，日本語または英語論文抄録に IPC コードを自動的に付与することを目的としている．IPC は，特許文献の技術内容に基づき，「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 階層から構成されており，「サブグループ」レベルで約 50,000 種類の IPC コードが存在する．なお，このサブタスクでは，「サブクラス」「メイングループ」「サブグループ」レベルの IPC コードを自動付与の対象としており，「サブグループ」レベルで学術分野と関連性の低い分野を除外した 30,885 種類の IPC コードを対象としている．

学術論文分類サブタスクでは，以下に示す四種類の課題が実施された．

- 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訓練データとして用いて分類する．
- 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訓練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データとして用いて分類する．

このサブタスクに参加したチームのほとんどは，k-NN (k-Nearest Neighbor)法を分類手法として採用していた[80, 99]．これは，付与対象となる IPC コード数がサブグループレベルで 30,885 種類もあるため，計算コストの面から最も適していたことが理由として挙げられる．一方で，機械学習のひとつであるロジスティック回帰モデルを採用したチームが一つだけあり，成績は 2 位であったが，トップと僅差の性能を得ていた[100]．また，特筆すべきシステムとして，英語文書分類サブタスクに参加した Xiao らのものが挙げられる[80]．このシステムでも k-NN 法を使用しているが，他の参加システムと異なる点に，出力する IPC コードの順位付けにリランキング手法を利用したことが挙げられる．k-NN 法は，SVM のような分類関数の学習を必要としないため，大規模なデータでも比較的高速に処理を行うことができる反面，データ中のノイズの影響を受けやすいことが問題点として挙げられる．そこで Xiao らは，k-NN に基づいてランク付けされた IPC コードリストを，順位付け学習のひとつである RankSVM[101]および重み付き線形和の 2 種類の方法で組み合わせ，精度の向上を試みた．その結果，いずれの方法においても，分類精度が向上することを示し，英語文書分類タスクにおいて最も良い成績を収めた．

技術動向マップ作成サブタスク

技術動向マップ作成サブタスクは，特許と論文から要素技術とその効果を示す表現を自動的に抽出することを目的としている．例えば，「英語の単名詞句とその他の句の同定問題に SVM を適用し，実際のタグ付けデータを用いて解析を行ったところ，従来手法に比べて高い精度を示した」という文が入力されると，図 2.11 に示したように，要素技術と効果を示す個所に，それぞれ TECHNOLOGY および EFFECT タグが自動的に付与される．なお，EFFECT タグの中では，さらに ATTRIBUTE と VALUE という 2 種類のタグが付与される．効果に関する表現は分野やタスクによって多様であり，そのすべてを処理対象とするのは，現在の言語処理技術では非常に困難である．このため，例えば，「処理速度(ATTRIBUTE)が向上する(VALUE)」や「ノイズ(ATTRIBUTE)が減少する(VALUE)」のように，技術の効果が「属性(ATTRIBUTE)」と「属性値(VALUE)」の対で表現できるもののみを対象としている．また，属性値となる表現が数値である場合，図 5.10 に示すような，属性値を抽出した文書の発行年を横軸，抽出した数値を縦軸とした，要素技術に関する性能の推移を可視化することができる．

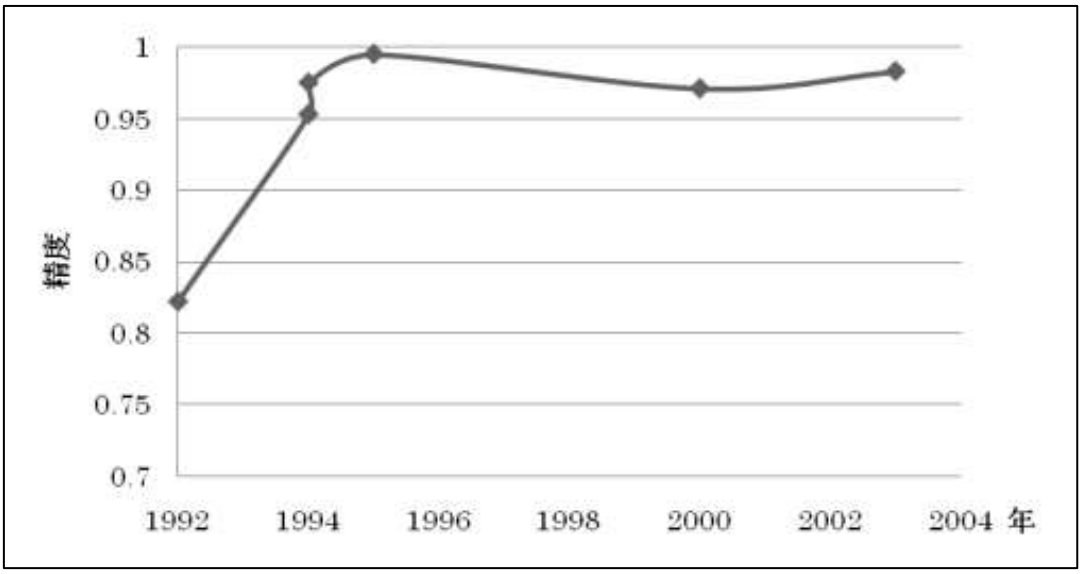


図 5.10 技術動向を数値として可視化した出力例

技術動向マップ作成サブタスクでは，以下の二つの課題が実施された．

- 日本語情報抽出(Japanese)： 日本語の論文と特許から，要素技術と効果に関する箇所を自動抽出する．
- 英語情報抽出(English)： 英語の論文と特許から，要素技術と効果に関する箇所を自動抽出する．

なお，論文では表題および概要を，特許では「発明の名称」「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」の各項目を，要素技術とその効果の抽出対象個所とした．

このサブタスクに参加した研究グループのシステムの多くは教師あり学習を用いており，日本語論文，英語論文からの情報抽出において，最も精度が高かったシステム[60, 102]では，CRF が用いられていた．特に Nishiyama らは，このタスクを転移学習としてとらえ，FEDA という手法で，特許と論文の訓練データを両方用いて学習を行うことを試みた．その結果，特許または論文の訓練データを単独で用いた場合よりも高い精度で情報抽出が可能であることを示した．

5.3 技術動向分析システムの動作例

本節では，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 article に収録されている論文データ, および NTCIR-8 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 5.11 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 5.11 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，その右側に各技術が使われた年が表示されている. 例えば図 5.11 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 5.11 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 5.12 は，図 5.11 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．中央には,「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．なお，図 5.12 の左側で表示されている「Research Fileds」の情報は，近藤らの主題抽出手法により抽出している[57].

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 5.11 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 5.12 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．

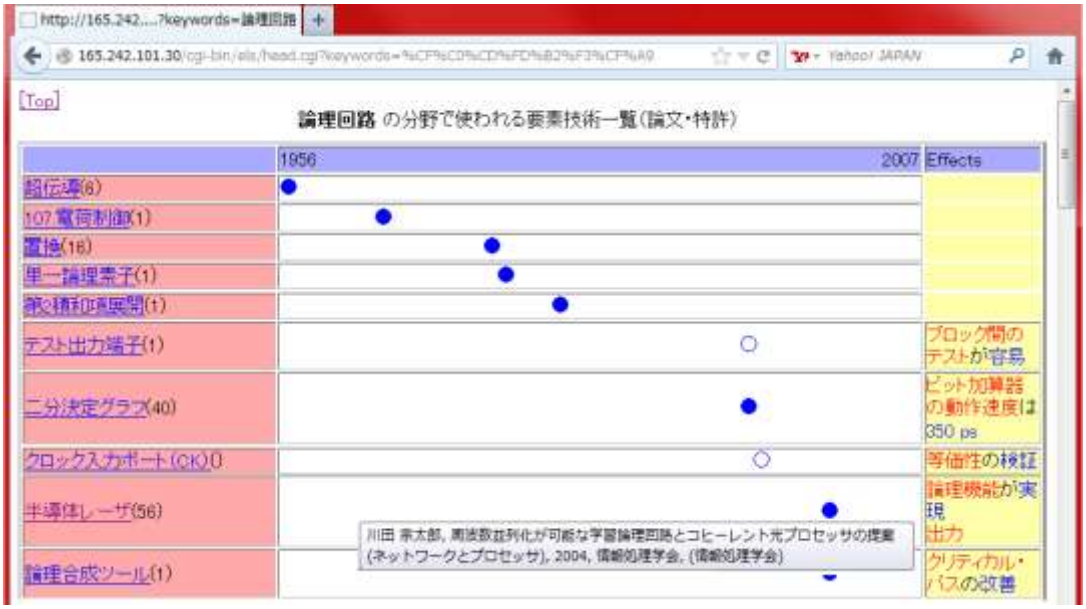


図 5.11 「論理回路」で使われる要素技術と効果の一覧表示

図 5.12 「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

5.4 まとめ

本章では，まず，技術動向の可視化に用いられるマップの形式を紹介した．次に，企業が提供している技術動向分析システムをいくつか紹介し，動向分析の可視化に関連する研究を述べた．そして，本論文と最も関連のある研究プロジェクトである NTCIR-7, 8 特許マイニングタスクの概要を述べ，最後に，要素技術とその効果を用いた，技術動向分析システムを示した．

本システムを用いることで，特定の分野を中心とした要素技術とその効果の変遷を可視化ことができる．これにより，膨大な文書を対象に分析作業を行うことなく，ある技術分野における「どのような要素技術がいつ頃から使われており，どのような効果が得られているのか」という情報を効率的に知ることができるといえる．

第6章　結論

本論文では，論文と特許から要素技術とその効果を抽出する手法を提案した．そして，抽出した手掛かり語を用いた KAKEN の分類体系に基づく論文の自動分類手法を提案した．さらに，要素技術とその効果を観点とした技術動向分析システムを構築した．

要素技術とその効果の抽出では，機械学習を用い，素性として，単語や品詞に加えて，要素技術，属性，属性値の手掛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掛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．さらに本論文では，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，本論文で提案した「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．

KAKEN の分類体系に基づく論文の自動分類手法では, k-NN 法と 4 種類のランキング手法を適用した．また，分類の際に用いる手掛かり語として，研究者名や学会・雑誌名に加え，論文中で記述されている要素技術とその効果に関する表現を用いた．そして，KAKEN の分類体系である「分野・分科・細目表」を対象に評価実験を行った結果，各階層における上位 1 件の結果に対して k-NN 法+Listweak 手法を適用したとき，それぞれ平均 0.853，0.712，0.615 の精度が得られた．また，上位 3 件までの出力結果に対して，同様の手法により，平均で 0.909，0.800，0.711 の MRR 値が得られた．これらの結果は，要素技術とその効果に関する表現を手掛かり語として用いない場合より高い値を示していることから，本手法の有効性が確認された．

そして，要素技術とその効果を示す表現を抽出することで，論文と特許を「要素技術」と「効果」という 2 つの観点で分析する動向分析システムを構築し，その動作例を示した．本システムでは，特定のキーワードにより収集した論文と特許から抽出した要素技術を縦軸に，各文書の著作年や出願年を横軸に取り，要素技術を用いて得られた効果をあわせて

提示する．これにより，そのキーワードを中心とした要素技術とその効果の変遷を可視化することができ，効率的な技術動向分析を実現することができる．

本論文では，日本語論文と特許を対象とした情報抽出および動向分析システムの構築を行った．今後は，英語論文および特許を対象も対象とすることで，海外との技術動向を比較したシステムの構築などを検討している．また，技術や発明の評価を行う際，論文や特許は，技術的な側面からの分析に有用であるといえる．一方で，ニュース記事や評価報告書などの文書には，技術や発明がもたらした経済効果や社会的評価に対する分析や評価が記述されている場合が多い．そして，研究としての有用性だけでなく，それにより得られる多大な利益や社会貢献の可能性を兼ね備えている技術や発明は，世の中において必要とされているものだと考えられる．今後は，技術的側面および経済的・社会的側面の両方の側面からによる技術や発明を評価した動向分析システムの構築も検討している．

謝辞

本研究を行うに当たり，格別なる御指導ならびに御鞭撻を賜りました竹澤寿幸教授，難波英嗣准教授に深甚なる感謝の意を表します．

高濱徹行教授，日浦慎作教授には，お忙しい中，博士論文の審査をお引き受けいただき，公聴会などで研究に関する助言をいただきました．深く御礼申し上げます．

黒澤義明助教，目良和也助教には，日頃から有益なご助言をいただき，多面に渡って励ましていただきました．有難うございました．

本論文をまとめるに当たって御協力いただいた言語音声メディア工学研究室の皆様に厚く御礼申し上げます．

最後に，私を支えてくれた家族に感謝致します．

参考文献

[1] 富浦洋一, 石田栄美. 学術論文検索の高度化のための論文アブストラクトのアノテーション, テキストアノテーションワークショップ, 2012.

[2] Takafumi Yamamoto, and Yoichi T Yamamoto. Constructing Corpus of Scientific Abstracts Annotated with Sentence Roles, In Proceedings of the the 5th International Congress on Advanced Applied Informatics, 2016.

[3] William C. Mann, and Sandra A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text, Vol.8, No.3, pp.243-281, 1988.

[4] Françoise Salanger-Meyer. Discoursal Flaws in Medical English Abstracts: A Genre Analysis Per Research- and Text-type, Text, Vol.10, No.4, pp.365-384, 1990.

[5] John M. Swales. Genre Analysis: English in Academic and Research Settings, Chapter 6. Cambridge University Press, UK, 1990.

[6] Constantin Orasan. Patterns in Scientific Abstracts, In Proceedings of Corpus Linguistics 2001 Conference, Lancaster University, Lancaster, UK, pp.433-443, 2001.

[7] Harold P. Edmundson. New Methods in Automatic Extracting. Journal of the Association for Computing Machinery, Vol.16, No.2, pp.264-285, 1969.

[8] Chris D. Paice. The Automatic Generation of Literature Abstracts: an Approach Based on the Identification of Self-indicating Phrases. In Proceedings of the 3rd annual ACM conference on Research and Development in Information Retrieval, pp.172-19, 1981.

[9] 三池誠司, 住田一男. 文の意味役割解析に基づく全文検索, 情報処理学会研究報告情報学基礎, pp.17-24, 1994.

[10] Daniel Marcu. Discourse Trees are Good Indicators of Importance in Text, In Inderjeet Mani and Mark T. Maybury, Editors, Advances in Automatic Text Summarization, 1999, MIT Press.

[11] Simone Teufel. Argumentative Zoning: Information Extraction from Scientific Text, Ph.D Thesis, University of Ediburgh, 1999.

[12] Douglas Biber, and Edward Finegan. Section 13: Intra-textual Variation within Medical Research Articles, Corpus-Based Research into Language, Oostdijk & de Haan(eds.), Amsterdam, Rodoph, pp.201-221, 1994.

[13] Noriko Kando. Text-level Structure: Implications for Information Retrieval and the

Potential for Genre Analysis, British Computer Society IR SG Annual Colloquium, 1997.

[14] Noriko Kando. Text Structure Analysis as a Tool to Make Retrieved Document Usable, In Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages, pp.126-135, 1999.

[15] 新森昭宏, 奥村学, 丸川雄三, 岩山真. 手がかり句を用いた特許請求項の構造解析, 情報処理学会論文誌, Vol.45, No.3, pp.891-905, 2004.

[16] 田中規久雄. 法律効果規定部の意味機能について, 自然言語処理, 1998.

[17] 小林良輔, 中村誠, 島津明. 修辞構造に基づく法令文の解析, 言語処理学会 第 14 回年次大会, pp.608-611, 2008.

[18] Paul Willot, Kazuhiro Hattori, and Akiko Aizawa. Extracting Structure from Scientific Abstracts using Neural Networks, In Proceedings of the 17th International Conference on Asia-Pacific Digital Libraries (ICADL 2015), pp.329-330, 2015.

[19] Simone Teufel, and Marc Moens. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status, Computational Linguistics, Vol.28, No.4, pp.409-445, 2002.

[20] Patrick Ruch, Celia Boyer, Chistine Chichester. Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, and Violaine Pillet. Rebholz-Schuhmann, Dietrich., Lovis, Christian. and Veuthey, Anne-Lise. Using Argumentation to Extract Key Sentences from Biomedical Abstracts. International Journal of Medical Informatics, Vol.76, No.2-3, pp.195-200, 2007.

[21] Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. Generative Content Models for Structural Analysis of Medical Abstracts, In Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06), pp.65-72, 2006.

[22] Jien-Chen Wu, Yu-Chia Chang, Hsien-Chin Liou, and Jason S. Chang. Computational Analysis of Move Structures in Academic Abstracts, In Proceedings of the COLING/ACL on Interactive Presentation Sessions, pp.41-44, 2006.

[23] Hong-Jie Dai, Wei-Chi Tsai, Tzong-Han Richard, and Wen-Lian Hsu. Enhancing Search Results with Semantic Annotation Using Augmented Browsing. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, pp.2418-2423, 2011.

[24] Larry McKnight, and Padmini Arinivasan. Categorization of Sentence Types in

Medical Abstracts. In Proceedings of the AMIA 2003 Symposium, pp.440-444, 2003.

[25] Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto, Using Sectioning Information for Text Retrieval: a Case Study with the Medline Abstracts, In Proceedings of Second International Workshop on Active Mining (AM'03), pp.32-41, 2003.

[26] Takahiko Ito, Masashi Simbo, Takahiro Yamasaki, and Yuji Matsumoto. Semi-supervised Sentence Classification for Medline Documents. In IPSJ SIG Technical Report, pp.141-146, 2004.

[27] Yasunori Yamamoto, and Toshihisa Takagi. A Sentence Classification System for Multi-document Summarization in the Biomedical Domain, In Proceedings of the InternationalWorkshop on Biomedical Data Engineering (BMDE2005), pp.90-95, 2005.

[28] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning (ICML-2001), pp.282-289, 2001.

[29] Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. Identifying Sections in Scientific Abstracts Using Conditional Random Fields, In Proceedings of the 3rd International Joint Conference of Natural Language Processing (IJCNLP2008), pp.381-388, 2008.

[30] Ryan T.K. Lin, Hong-Jie Dai, Yue-Yang Bow, Justin Liang-Te Chiu, and Richard Tzong-Han Tsai. Using Conditional Random Fields for Result Identification in Biomedical Abstracts, Journal of Integrated Computer-Aided Engineering, Vol.16, No.4, pp.339-352, 2009.

[31] Sonal Gupta, and Christopher D. Manning. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers, In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2011.

[32] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. Overview of the NTCIR-10 MedNLP Task, In Proceedings of the 10th NTCIR Conference, pp.696-701, 2013.

[33] Peter Anick, Marc Verhagen, and James Pustejovsky. Identification of Technology Terms in Patents. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014.

[34] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkumura. Overview of the NTCIR-11 MedNLP-2 Task. In Proceedings of the 11th NTCIR Workshop

Meeting on Evaluation of Information Access Technologies, pp.147-154, 2014.

[35] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkumura. Overview of the NTCIR-12 MedNLPdoc Task, In Proceedings of the 12th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pp.71-75, 2015.

[36] Michael Roth, and Ewan Klein. Parsing Software Requirements with an Ontology-based Semantic Role Labeler. In Proceedings of the 1st Workshop on Language and Ontologies, pp.15-21, 2015.

[37] 建石由佳, 仕田原容, 宮尾祐介, 相澤彰子. 情報科学論文からの意味関係抽出に向けたタグ付けスキーマ, 言語処理学会第 19 回年次大会講演論文集, pp.702-705, 2013.

[38] Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. Relation Annotation for Understanding Research Papers, In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp.140-148, 2013.

[39] Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. The GENIA corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain, In Proceedings of the second international conference on Human Language Technology Research, pp.82-86, 2002.

[40] 大田朋子, 建石由佳, 金進東, 薬師寺あかね, 辻井潤一. 生命科学分野のタグ付きコーパス：GENIA コーパスの設計と作成, 言語処理学会第 11 回年次大会発表論文集, 2005.

[41] 新里圭司, 関根聡, 吉永直樹, 鳥澤健太郎. 固有表現抽出手法を用いたレストラン属性情報の自動認識, 言語処理学会第 12 回年次大会発表論文集, pp.93-96, 2006.

[42] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築, 情報処理学会研究報告 自然言語処理, Vol.2008, No.113, pp.113-120, 2008.

[43] 前川喜久雄. 代表性を有する大規模日本語書き言葉コーパスの構築, 人工知能学会誌, Vol.24, No.5, pp. 616-622, 2009.

[44] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy, In Proceedings of the LREC 2002, 2002.

[45] Satoshi Sekine, and Chikashi Nobata. Definition, Dictionary and Tagger for Extended Named Entities, In Proceedings of the Forth International Conference on Language Resources and Evaluation, pp.1977-1980, 2004.

[46] Satoshi Sekine. Extended Named Entity Ontology with Attribute Information, In Proceedings of the 5th International Conference on Language Resources and Evaluation, pp.52-57, 2008.

[47] Ralph Grishman, and Beth Sundheim. Message Understanding Conference - 6: A

Brief History, In Proceedings of the COLING 1996, pp.466-471, 1996.

[48] Satoshi Sekine, and Hitoshi Isahara. IREX: IR and IE Evaluation project in Japanese, In Proceedings of the LREC 2000, pp.7-12, 2000.

[49] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto. Overview of the Patent Mining Task at the NTCIR-8 Workshop. In Proceedings of the 8th NTCIR Workshop Meeting, 2010.

[50] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying Patient Smoking Status from Medical Discharge Records, Journal of the American Medical Informatics Association (JAMIA), Vol.15, No.1, pp.14-24, 2008.

[51] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting Medication Information from Clinical Text, Journal of the American Medical Informatics Association (JAMIA), Vol.17, No.5, pp.514-518, 2010.

[52] Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text, Journal of the American Medical Informatics Association (JAMIA), Vol.18, No.5, pp.552-556, 2011.

[53] Atsushi Matsumura, Atsuhiro Takasu, and Jun Adachi. Structured Index System at NTCIR1: Information Retrieval using Dependency Relationship between Words, In Proceedings of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.117-122, 1999.

[54] 三平善郎, 山本喜一. 文献間の引用関係を用いた主題抽出とその検索システム, 日本ソフトウェア科学会 第 12 回全国大会論文集, pp.77-80, 1995.

[55] 西山莉紗, 竹内広宜, 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.

[56] 酒井浩之, 野中尋史, 増山繁. 特許明細書からの技術課題情報の抽出, 人工知能学会論文誌, Vol.24, No.6, pp.531-540, 2009.

[57] 近藤友樹, 難波英嗣, 竹澤寿幸. 論文と特許からの技術動向マップの自動構築, 言語処理学会 第 16 回年次大会, pp.114-117, 2010.

[58] Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. Annotation of Computer Science Papers for Semantic Relation Extraction, In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp.26-31, 2014.

[59] 橋本泰一, 中村俊一. 拡張固有表現タグ付きコーパスの構築 －白書, 書籍, Yahoo!知恵袋コアデーター,言語処理学会 第 16 回年次大会, pp.916-919, 2010.

[60] Risa Nishiyama, Yuta Tsuboi, Yuya Unno, and Hironori Takeuchi. Feature-Rich

Information Extraction for the Technical Trend-Map Creation, In Proceedings of the 8th NTCIR Workshop Meeting, pp.318-324, 2010.

[61] Hal Daumé III. Frustratingly Easy Domain Adaptation, In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp.256-263, 2007.

[62] Osamu Imaichi, Toshihiko Yanase, and Yoshiki Niwa. A Comparison of Rule-based and Machine Learning Methods for Medical Information Extraction, In Proceedings of the 1st Workshop on Natural Language Processing for Medical and Healthcare Fields, pp.38-42, 2013.

[63] Pierre-François Laquerre, and Christopher Malon. NECLA at the Medical Natural Language Processing Pilot Task (MedNLP), In Proceedings of the 10th NTCIR Conference, pp.725-727, 2013.

[64] 東山翔平, 関和広, 上原邦昭. 医療用語資源の語彙拡張と診療情報抽出への応用, 自然言語処理, Vol.22, No.2, pp.77-105, 2015.

[65] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. 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.

[66] Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka. 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, 2001.

[67] Lillian Lee. Measures of Distributional Similarity, In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp.25-32, 1999.

[68] Dekang Lin. Automatic Retrieval and Clustering of Similar Words, In Proceedings of the 36th Annual Meeting of the Associ- ation for Computational Linguistics and the 17th International Conference on Computational Linguistics, pp.768-774, 1998.

[69] 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.

[70] Gerand Salton. The SMART Retrieval System – Experiments in Automatic Document Processing, Prentice- Hall, Inc., Upper Saddle River, NJ, 1971.

[71] Erik F. Sang, Tjong Kim, and Jorn Veenstra. Representing Text Chunks, In Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics, pp.173-179, 1999.

[72] Hironori Mizuguchi, and Dai Kusui. An Information Extraction Method for Multiple Data Sources, In Proceedings of the 8th NTCIR Workshop Meeting, pp.348-353, 2010.

[73] Yusuke Suzuki, Hirofumi Nonaka, Hiroki Sakaji, Akio Kobayashi, Hiroyuki Sakai, and Shigeru Masuyama. NTCIR-8 Patent Mining Task at Toyobashi University of Technology, In Proceedings of the 8th NTCIR Workshop Meeting, pp.364-369, 2010.

[74] Yusuke Sato, and Makoto Iwayama. Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi, In Proceedings of the 8th NTCIR Workshop Meeting, pp.359-363, 2010.

[75] Leonidas Akritidis, and Panayiotis Bozanis. A Supervised Machine Learning Classification Algorithm for Research Articles, In Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp.115-120, 2013.

[76] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. A Machine Learning Approach to Building Domain-Specific Search Engines, In Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp.662-667, 1999.

[77] 宮田洋輔, 石田栄美, 神門典子, 上田修一. NDC の階層構造を利用した図書の自動分類の試み, 日本図書館情報学会春季研究集会発表要綱, pp.51-54, 2006.

[78] Leonardo Rocha, Fernando Mourão, Hilton Mota, Thiago Salles, Marcos André GonçAlves, and Wagner Meira Jr. Temporal Contexts: Effective Text Classification in Evolving Document Collections, Information Systems, Vol.38, No.3, pp.388-409, 2013.

[79] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop, In Proceedings of NTCIR-7 Workshop Meeting, pp.16-19, 2008.

[80] Tong Xiao, Feifei Cao, Tianning Li, Guolong Song, Ke Zhou, Jingbo Zhu, and Huizhen Wang. KNN and Re-ranking Models for English Patent Mining at NTCIR-7, In Proceedings of the 7th NTCIR Workshop Meeting, pp.333-340, 2008.

[81] 今井俊, 佐藤理史. 表題解析による科学技術論文の詳細分類, 情報処理学会第 57 回全国大会講演論文集, pp.211-212, 1998.

[82] Kiyoko Uchiyama, Hidetsugu Nanba, Akiko Aizawa, and Takeshi Sagara. OSUSUME: Cross-lingual Recommender System for Research Papers, In Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation, pp.39-42, 2011.

[83] 坂本剛彦, ホーツーバオ. データコレクション間の類似度を利用したトピック発見手

法の提案, 電子情報通信学会第 19 回データ工学ワークショップ, 2008.

[84] Sumanne Hoche, and Peter Flach. Predicting Topics of Scientific Papers from Co-Authorship Graphs: a Case Study, In Proceedings of the 2006 UK Workshop on Computational Intelligence (UKCI2006), pp.215-222, 2006.

[85] Xiaodan Zhang, Xiaohua Hu, and Xiaohua Zhou. A Comparative Evaluation of Different Link Types on Enhancing Document Clustering, In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pp.555-562, 2008.

[86] Leah S. Larkey. Some Issues in the Automatic Classification of U.S. Patents, Working Notes for the AAAI-98 Workshop on Learning for Text Categorization, pp.87-90, 1998.

[87] 間瀬久雄, 辻洋, 絹川博之, 石原正博. 特許テーマ分類方式の提案とその評価実験, 情報処理学会論文誌, Vol.39, No.7, pp.2207-2216, 1998.

[88] 安藤俊幸. テキストマイニングと統計解析言語 R による特許情報の可視化, 情報管理, Vol.52, No.1, pp.20-31, 2009.

[89] Thomas Hofmann. Probabilistic Latent Semantic Analysis, In Proceedings of 15th Conference on Uncertainty in Artificial Intelligence, pp.289-296, 1999.

[90] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Diriclet Allocation, Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[91] David M. Blei, and John D. Lafferty. A Correlated Topic Model of Science, The Annals of Applied Statistics, Vol.1, No.1, pp.17-35, 2007.

[92] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the History of Ideas Using Topic Models, In Proceedings of EMNLP 2008, pp.363-371, 2008.

[93] Fabian Mörchen, Mathäus Dejori, Dmitriy Fradkin, Julien Etienne, Bernd Wachmann, and Markus Bundschus. Anticipating Annotations and Emerging Trends in Biomedical Literature, In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and DataMining, pp.954-962, 2008.

[94] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and C. Lee Giles. Detecting Topic Evolution in Scientific Literature: How Can Citations Help?, In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09), pp.957-966, 2009.

[95] Levent Bolelli, Şeyda Ertekin, Ding Zhou, and C. Lee Giles. Finding Topics Trends in Digital Libraries, In Proceedings of Joint Conference on Digital Libraries

(JCDL'09), pp.69-72, 2009.

[96] Levent Bolelli, Şeyda Ertekin, and C. Lee Giles. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation, In Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, pp.776-780, 2009.

[97] Ding Zhou, Xiang Ji, Hongyuan Zha, and C. Lee Giles. Topic Evolution and Social Interactions: How Authors Effect Research, In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp.391-399, 2006.

[98] 岩田具治, 山田武士, 上田修功. トピックモデルに基づく文書群の可視化, 情報処理学会論文誌, Vol.50, No.6, pp.1234-1244, 2009.

[99] Stéphane Clinchant, and Jean-Michel Renders. XRCE's Participation to Patent Mining Task at NTCIR-7, In Proceedings of the 7th NTCIR Workshop Meeting, pp.351-353, 2008.

[100] Hisao Mase, and Makoto Iwayama. NTCIR-7 Patent Mining Experiments at Hitachi, In Proceedings of the 7th NTCIR Workshop Meeting, pp.365-368, 2008.

[101] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support Vector Learning for Ordinal Regression. In Proceedings of the 9th International Conference on Artificial Neural Networks, pp.97-102, 1999.

[102] Jingjing Wang, Han Tong Loh, and Wen Feng Lu. Extracting Technology and Effect Entities in Patents and Research Papers, In Proceedings of the 8th NTCIR Workshop Meeting, pp. 325-330, 2010.

発表論文一覧

受賞

- データサイエンス・アドベンチャー杯 入賞
  難波英嗣, 福田悟志, 飯沼俊平, 竹澤寿幸. ニュース記事と特許を利用した科学技術の重要性の評価, 2014.
- 最優秀修士論文賞
  福田悟志. 要素技術とその効果に着目した技術文書の自動分類および動向分析, 平成24年度 広島市立大学大学院情報科学研究科 修士論文, 2013.
- 国立大学共同利用機関法人情報・システム研究機構 国立情報学研究所賞
  福田悟志. CiNii Mining, Mashup Awards 8, 2012.

論文誌

1. 福田悟志, 難波英嗣, 竹澤寿幸. 要素技術とその効果を用いた学術論文の自動分類, 日本図書館情報学会誌, Vol.63, No.3, pp.145-162, 2016.
2. Shumpei Iinuma, Satoshi Fukuda, Hidetsugu Nanba, and Toshiyuki Takezawa. Evaluation of the Industrial and Social Impacts of Academic Research using Patents and News Articles, International Journal of Computers & Information Sciences, Vol.16, No.1, pp.12-21, 2015.
3. 福田悟志, 難波英嗣, 竹澤寿幸. 論文と特許からの技術動向情報の抽出と可視化, 情報処理学会論文誌データベース, Vol.6, No.2, pp.16-29, 2013.

国際会議

1. Satoshi Fukuda, Hikaru Nakahashi, Hidetsugu Nanba, and Toshiyuki Takezawa. Quick Evaluation of Research Impacts at Conferences using SNS, In Proceedings of the 12th International Workshop on Text-based Information Retrieval, in conjunction with DEXA 2015, 2015.

2. Shumpei Iinuma, Satoshi Fukuda, Hidetsugu Nanba, and Toshiyuki Takezawa. Evaluation of the Industrial and Social Impacts of Science and Technology Using Patents and News Articles, In Proceedings of the 5th International Conference on E-Service and Knowledge Management (ESKM 2014), 2014.
3. Satoshi Fukuda, Hidetsugu Nanba, and Toshiyuki Takezawa, and Akiko Aizawa. Classification of Research Papers Focusing on Elemental Technologies and Their Effects, In Proceedings of the 6th Language & Technology Conference (LTC'13), 2013.
4. Satoshi Fukuda, Hidetsugu Nanba, and Toshiyuki Takezawa. Extraction and Visualization of Technical Trend Information from Research Papers and Patents, In Proceedings of the 1st International Workshop on Mining Scientific Publications, collocated with JCDL 2012, 2012.

その他の発表論文

1. 福田悟志, 難波英嗣, 竹澤寿幸, 乾孝司, 岩山真, 橋田浩一, 藤井敦. F タームに基づいたオントロジーの構築, 言語処理学会第 21 回年次大会, pp.1056-1059, 2015.
2. 福田悟志, 難波英嗣, 竹澤寿幸. 要素技術とその効果を用いた学術論文の自動分類, 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2015), 2015.
3. 福田悟志, 難波英嗣, 竹澤寿幸, 乾孝司, 岩山真, 橋田浩一, 藤井敦. F タームに基づいたオントロジーの構築, 第 3 回特許情報シンポジウム, 2014.
4. 飯沼俊平, 福田悟志, 難波英嗣, 竹澤寿幸. ニュース記事と特許を利用した科学技術の重要性の評価, 人工知能学会 全国大会(第 28 回), 2014.
5. 福田悟志, 難波英嗣, 竹澤寿幸, 橋田浩一. 用語の属性を考慮した上位, 下位概念辞書の構築, 言語処理学会第 20 回年次大会, pp.967-970, 2014.
6. 福田悟志, 難波英嗣, 竹澤寿幸, 武田英明, 相澤彰子, 大向一輝, 宮尾祐介, 内山清子. 学術情報データベースを用いた学術情報検索システムの構築, NLP 若手の会 第 7 回シンポジウム, 2012.
7. 福田悟志, 難波英嗣, 竹澤寿幸, 武田英明, 相澤彰子, 大向一輝, 宮尾祐介, 内山清子. CiNii データベースを用いた研究動向分析システムの構築, 言語処理学会第 18 回年次大会, pp.539-542, 2012.
8. 福田悟志, 難波英嗣, 竹澤寿幸. 技術文書からの動向情報の抽出と可視化, 言語処理学会第 17 回年次大会, pp.276-279, 2011.