

情報工学コース卒業研究報告

学術論文からの 研究資源情報の自動抽出 に関する研究

2013年2月

井上 慧

概要

研究活動には研究資源の利用が不可欠である。研究で用いられるデータや資料、ツールに加え、既存手法や評価法などの無形のものも広義の研究資源と捉えることができる。これまでに膨大な量の研究資源が構築され、その多くは作成者によって公開・提供されており、多種多様な用途に再利用することができる。このように研究資源を共有し、様々な用途に活用することは、研究者各々の効率的な研究活動を促すとともに、研究分野全体のレベル向上につながる。しかし、現状では既存の研究資源が有効に活用されているとは言い難い。その理由として、研究資源に関する情報が十分に共有されていないため、利用者が自らの目的に合致した研究資源を容易に見つけることができないという点が挙げられる。

そこで、本論文では、研究資源の名称や定義・用途のような、研究資源を検索・選択する際に有用となる情報（以下、**研究資源情報**）を網羅的に取得し提供することを目的に、学術論文からの研究資源情報の自動抽出手法を提案する。学術論文は、一般的な Web テキストと比べて、文章が整っており、記述内容の専門性・信頼性も高いため、有用な情報を精度よく抽出できる可能性がある。

学術論文の本文には、研究資源の名称（以下、**研究資源名**）のみならず、研究資源がどのようなものかを示す表現（以下、**定義表現**）、どのような目的・用途で使用されたのかを示す表現（以下、**用途表現**）などの研究資源情報が含まれている。特に、定義表現および用途表現の中には、利用者の研究活動によって新たに創出された用途も多く存在する。そのため、研究資源名とともに定義表現および用途表現を提供することは、研究資源の広範な利用に大きく貢献すると考えられる。

本研究では、研究資源名と定義表現および用途表現が文内で共起することに着目した。50 論文を用いた予備調査によって、定義表現および用途表現の 96.8% が研究資源名を含む文中に現れていることを確認した。また、研究資源名と、定義表現あるいは用途表現の間には、研究資源情報抽出の手がかりとなる表現（以下、**手がかり表現**）がよく現れることを確認した。

提案手法では、学術論文の分析によって得られた、(1) 研究資源名の形態素的特徴、(2) 定義表現および用途表現の持つ特徴、(3) 手がかり表現を利用して、学術論文の本文テキストから研究資源名の候補となる形態素列を取得し、SVM を用いて研究資源名を抽出する。一方、定義表現および用途表現の抽出には、手がかり表現および研究資源名を利用する。

また、提案手法を評価するために、研究資源名の抽出実験と定義表現および用

途表現の抽出実験を個別に実施した。実験には、2008 年度人工知能学会全国大会の講演論文集に収録されている 150 論文を使用した。研究資源名の抽出実験の結果、研究資源名の抽出において、既知の定義表現および用途表現の持つ特徴を活用することの有効性を確認した。また、定義表現および用途表現の抽出実験の結果、定義表現および用途表現の抽出において、手がかり表現を活用することの有効性を確認した。

目 次

第 1 章	はじめに	1
1.1	研究の背景と目的	1
1.2	本論文の構成	4
第 2 章	研究資源情報とその利用	5
2.1	研究資源情報とは	5
2.2	研究資源情報に関する分析	6
2.2.1	分析の概要	6
2.2.2	分析結果	8
2.3	研究資源情報の利用	16
第 3 章	研究資源情報の抽出手法	19
3.1	抽出手法の概要	19
3.2	PDF のテキストへの変換	20
3.3	研究資源名の自動抽出	20
3.3.1	研究資源名候補の取得	20
3.3.2	SVM による研究資源名の取得	26
3.4	研究資源情報の自動抽出	31
3.4.1	定義表現の抽出	31
3.4.2	用途表現の抽出	33
第 4 章	評価実験	37
4.1	実験の目的と方法	37
4.2	実験結果	39
4.2.1	研究資源名の抽出実験	39
4.2.2	定義表現および用途表現の抽出実験	40
第 5 章	関連研究	45
5.1	固有表現抽出	45
5.2	知識獲得	47

第 6 章	おわりに	49
6.1	まとめ	49
6.2	今後の課題	50
謝辞		53
参考文献		55

第1章 はじめに

1.1 研究の背景と目的

一般に、研究活動において研究資源は必要不可欠である。研究資源とは、研究で用いられるデータや資料、ツールのことであるが、汎用的な手法や評価法といった無形のものも広義の研究資源と捉えることができる。研究資源のうち、特にツールや実験データなどは、研究者が自身の研究目的のために自作することも少なくなく、これまでに膨大な量の研究資源が構築されてきた。また、構築された研究資源の多くは作成者によって公開・提供されており、作成者の研究目的に限らず、多種多様な用途に再利用することが可能である。このように研究資源を共有し、様々な用途に活用することは、研究者各々の効率的な研究活動を促すとともに、研究分野全体のレベル向上につながる。しかし、現状では既存の研究資源が有効に活用されているとは言い難く、その理由として、研究資源に関する情報が十分に共有されていないため、利用者が自らの目的に合致した研究資源を容易に見つけることができないという点が挙げられる。

研究経験の浅い、あるいは専門分野とは異なる分野の研究を新たに始める利用者が研究資源を探す方法としては、専門家への問い合わせ、先行研究の文献調査、関連キーワードによるウェブ検索などが考えられる。しかし、専門家に問い合わせることは相手にとってコストが高く、あらゆる利用者が容易に実行できる手段ではない。また、研究経験の有無に関わらず、一人の人間が研究資源をすべて把握することは困難であるため、問い合わせによって利用者が求める研究資源情報を必ず入手できるとも限らない。先行研究の文献調査は、研究者が研究活動において必ず行うべき作業であり、調査を通して研究資源情報を収集することが考えられる。しかし、この方法では利用者の研究分野外で用いられている研究資源の情報を入手することはできないため、本来様々な研究分野で利用できるはずの研究資源が有効に活用されない可能性がある。ウェブ検索では、学術論文やツールの公式ウェブサイト、個人のブログやスライドなど様々な媒体から最新の研究資源情報を入手することができると考えられるが、他の手段と比べて不正確な情報が多分に含まれるため、信頼性は必ずしも高くない。

一方で、「分子生物学研究用ツール集」^{*1}や「サイエンスポータル 研究ツール

^{*1}<http://www.yk.rim.or.jp/~aisoai/molbio-j.html>

データベース」^{*2}など、研究ツールやデータベースに関する情報をまとめたウェブサイトが構築されてきている。ウェブサイトの多くは、研究資源の概要や入手先のURLなどの情報が人手によって収集・整理されたものであり、言語資源メタデータベース「SHACHI」^{*3}[3]などのように、一部にはキーワードで検索できるものもある。しかし現状では、データの整備を人手に頼っているために、更新・管理のコストが高く、新しく構築された研究資源を容易に反映させることができない。また、個々のウェブサイトにおいて、掲載されている研究資源の情報は特定分野のものに限られており、情報の内容も概要や基本的な用途のみであることが多く、様々な研究分野での研究資源の再利用を促すものではない。研究資源に関する情報を機械的に大量に取得することができれば、取得した情報をデータベース化して公開することにより、研究資源に関する情報を利用者間で十分に共有できると考えられる。

そこで、本論文では、研究資源の名称や定義・用途のような、研究資源を検索・選択する際に有用となる情報（以下、**研究資源情報**と呼ぶ）を機械的に取得することを目的とし、学術論文からの研究資源情報の自動抽出手法を提案する。

近年、学術論文の電子化が進んでいる。多くの学会発表では発表論文集をCD-ROMで配布するようになり、Google Scholar^{*4}のようなWeb上の論文データを検索できるサービスや、CiNii^{*5}のような学術情報から検索することで学術論文の詳細な情報やPDFデータを取得できるデータベース・サービスも広く利用されている。学術論文は日々蓄積され続けており、比較的文章が整っていて記述内容の専門性・信頼性も高いため、学術論文から有用な情報を機械的に抽出することが可能だと考えられる。

学術論文の本文には、研究で用いられた研究資源の名称（以下、**研究資源名**と呼ぶ）のみならず、研究資源がどのようなものなのかを示す表現（以下、**定義表現**と呼ぶ）、どのような目的・用途で使われたのかを示す表現（以下、**用途表現**と呼ぶ）などの情報が含まれている。定義表現および用途表現の例として、形態素解析ツールのひとつであるChaSenには定義表現「形態素解析器」や用途表現「単語への分割と品詞の推定」などが、パターン認識手法のひとつであるSVM(Support Vector Machine)には定義表現「2値分類問題に対する有効な分類モデルの導出手法」や用途表現「記事を主観的な意見を含むレビュー記事と非レビュー記事に分類する」などがそれぞれ該当する。

学術論文中の定義表現および用途表現には、研究資源の作成者が想定した用途だけではなく、利用者の研究活動によって新たに創出された用途も多く含まれている。そのため、研究資源名とともに定義表現および用途表現を提供することは、

^{*2}<http://scienceportal.jp/database/D01/5/>

^{*3}<http://www2.shachi.org/>

^{*4}<http://scholar.google.co.jp/>

^{*5}<http://ci.nii.ac.jp/>

研究資源の広範な利用に大きく貢献すると考えられる。また、研究資源名や定義表現および用途表現などの研究資源情報によって、利用者の研究目的に合致した研究資源の発見や、複数の類似した研究資源の比較をより効率的に行うことができると考えられる。

研究資源情報の自動抽出としては、言語資源の用途表現を、その言語資源の利用者によって記されたテキストから自動抽出する手法が小澤らによって提案されている [9]。小澤らの手法では、構文構造を考慮した抽出ルールを人手によって作成し、言語資源の名称を既知の情報として利用することで用途表現を獲得している。それに対して、本研究で提案する抽出手法では、言語資源以外の研究資源の情報も抽出対象とし、研究資源名は既知でないことを前提とする。すなわち、本研究では、研究資源名および定義表現および用途表現を学術論文から機械的に抽出する。

研究資源情報の抽出において、本研究では、研究資源名と定義表現および用途表現が文内で共起することに着目する。日本語で書かれた学術論文の分析の結果、研究資源の定義や用途について書かれた文には、研究資源名も書かれている場合が多いことが明らかとなったことから、研究資源名の抽出には、その形態的特徴の他に、定義表現および用途表現の持つ特徴が有用な手がかりとなると考えた。同様に定義表現および用途表現の抽出において、研究資源名が有用な手がかりとなる。したがって、研究資源名と定義表現および用途表現は、ブートストラップ的に抽出できると考えられる。

提案する抽出手法では、まず、前処理として日本語で書かれた学術論文の PDF データから本文テキストを抽出する。学術論文の分析によって得た研究資源名の形態的特徴や定義表現および用途表現を利用して、本文テキストから研究資源名の候補となる形態素列を取得し、SVM を用いて研究資源名を抽出する。

定義表現および用途表現はある程度の長さを持った形態素列であり、同一の表現が学術論文の集合中に複数回出現することは極めて稀であるため、そのままでは研究資源名の抽出に利用することができない。そこで、定義表現は末尾の名詞を、用途表現は末尾の動詞・サ変名詞を研究資源名抽出の手がかりとする。例えば、定義表現「2 値分類問題に対する有効な分類モデルの導出手法」の場合は「手法」が、用途表現「記事を主観的な意見を含むレビュー記事と非レビュー記事に分類する」の場合は「分類」が手がかりとなる。

定義表現および用途表現の抽出には、「を用いて」や「である」のような、定義表現および用途表現の抽出において手がかりとなる表現（以下、**手がかり表現**と呼ぶ）および研究資源名を利用する。抽出手法の評価は、学術論文から人手で作成した正解データを用いて行う。

1.2 本論文の構成

以下に本論文の構成を示す。第2章では、本研究における研究資源情報を定義し、学術論文中に登場する研究資源情報を分析した結果を述べ、研究資源情報の利用法について述べる。次に第3章で学術論文からの研究資源名の自動抽出手法、定義表現および用途表現の自動抽出手法について詳しく述べ、第4章で提案手法の評価実験と考察について述べる。第5章では本研究の関連研究について述べる。最後に第6章では、まとめと今後の課題について述べる。

第2章 研究資源情報とその利用

本研究では、研究で用いられるデータや資料、ツールだけではなく、汎用的な手法や評価法のような無形のものも広義の研究資源と見なし、それら研究資源の広範な利用を促すための研究資源情報を機械的に取得することを目的としている。

本章では、本研究における研究資源情報の定義と本研究で抽出対象とする研究資源情報の概要について述べ、研究資源名、定義表現、用途表現などの研究資源情報を学術論文から人手で抽出し、分析した結果について述べる。また、研究資源情報の利用法について述べる。

2.1 研究資源情報とは

本研究では、研究資源の検索・選択において有用となる情報を研究資源情報と呼ぶ。研究資源情報には、研究資源の名称、定義や用途などについて記述されたテキスト情報だけではなく、研究資源を説明するための画像や実際に研究資源を利用している様子を記録した映像など、様々な形式の情報が存在すると考えられる。同様に、研究資源情報を含む媒体も、研究資源の作成者が研究資源を説明するために記した文書や、研究資源の利用者による文書、プレゼンテーションスライド、講演の映像データなど、多種多様である。

情報検索を行う際には、目的の情報に関連するキーワードのようなテキストを入力することが一般的であるため、研究資源の検索・選択においてもテキスト形式の研究資源情報が最も有用であると考えられる。そこで本研究では、学術論文の本文テキストに着目する。研究資源の利用者によって書かれた学術論文の本文テキストには、利用者の研究活動によって新たに考案された利用法など、研究資源の作成者が提供できないような情報が多く含まれていると考えられる。また、そのような研究資源情報は大量の学術論文の中に散在しており、人手で取り出すことが困難であるため、学術論文から研究資源情報を機械的に抽出する手法が求められる。

学術論文の本文テキストに含まれる研究資源情報には、研究資源名、定義表現、用途表現などが存在する。研究で用いられたデータや資料、ツールや汎用的な手法は、その名称が論文中で言及されると予測される。また、使用した研究資源がどのようなものなのか、どのような目的・用途で使用されたのかを説明する必要

があるため、定義表現や用途表現も論文に含まれると考えられる。

小澤らは、言語資源 X の用途情報を「A のために X を利用する」という表現に言い換え可能な表現 A であると定義した [9]。本研究では、研究資源 X の定義表現を「X とは、A である」あるいは「X とは、A の一種である」という表現に言い換え可能な表現 A であると定義する。また、用途表現については、小澤らの定義と同様とする。

定義表現や用途表現の他にも、研究資源の機能を説明する表現や、研究資源の利用によって得られた成果物を示した表現、研究資源を適用した対象を示した表現、データや資料の場合はその規模を示した表現などの情報も、研究資源の検索・選択に有用な研究資源情報であると考えられる。学術論文に含まれるこれらの研究資源情報について分析した結果を次節で述べる。

2.2 研究資源情報に関する分析

2.2.1 分析の概要

学術論文に含まれる研究資源情報の実態の調査と、学術論文からの研究資源情報の自動抽出にあたって手がかりとなる特徴の発見を目的に、研究資源名、定義表現、用途表現などの研究資源情報を学術論文から人手で抽出して分析を行った。

分析には、2008 年度人工知能学会全国大会の発表論文集に収録されている 50 論文を使用した。また、研究資源情報を人手で取得するにあたって、PDF ファイルを HTML 形式に変換するツール `pdftohtml`^{*1}を用いて論文の PDF ファイルを HTML 形式に変換し、独自のプログラムによって HTML から本文テキストを抽出して文単位に分解した。この独自プログラムについては3章で詳しく述べる。

抽出する研究資源情報の内容としては、前節で述べた研究資源名、定義表現、用途表現に加えて、以下の4種の表現も取得対象とした。

- 機能・特徴表現：研究資源の機能や特徴を説明する表現
- 成果物表現：研究資源を使用することで得られる成果物を示した表現
- 対象表現：研究資源を適用する対象を示した表現
- 規模表現：研究資源の規模を示した表現

以下に、各研究資源情報の例を挙げる。

文(2.1)は、定義表現を含む例である。

(2.1) Radio Button は、拍打に基づくバトン型演奏プログラムである。

^{*1}<http://pdftohtml.sourceforge.net/>

この文では、「Radio Button」が研究資源名、「拍打に基づくバトン型演奏プログラム」が定義表現に相当する。

文 (2.2) は、用途表現を含む例である。

(2.2) 単語への分割と品詞の推定には ChaSen を利用している。

この文では、「ChaSen」が研究資源名、「単語への分割と品詞の推定」が用途表現に相当する。

文 (2.3)、(2.4) は、機能・特徴表現を含む例である。

(2.3) Simpleminer は、Windows に簡単にインストールができる。

(2.4) Vivid910 は、複数枚の静止画像と距離画像から 3 次元形状を構成する。

文 (2.3) では、「Simpleminer」が研究資源名、「Windows に簡単にインストールができる」が機能・特徴表現に相当する。なお、研究活動で使用された機能を説明する表現は、研究資源の用途を示す表現にもなりうると考えられるが、研究活動でその機能が使用されたということが一文中で明示的に表されていない場合は機能・特徴表現とする。文 (2.4) では、その研究でその機能が使われたかどうか不明のため、研究資源名「Vivid910」に対して、「複数枚の静止画像と距離画像から 3 次元形状を構成する」という部分が機能・特徴表現となる。

文 (2.5) は、成果物表現および対象表現を含む例である。

(2.5) 文末表現は質問と回答の中の全ての文から茶釜を用いて抽出する。

この文では、「茶釜」が研究資源名、「文末表現」が成果物表現、「質問と回答の中の全ての文」が対象表現に相当する。

文 (2.6) は、規模表現を含む例である。

(2.6) 治験薬構造データベース中に収載されている 114,501 件の薬物構造データを対象に、その NTG および関連する薬理活性情報を抽出し、相互の関係辞書ファイルを作成した。

この文では、「治験薬構造データベース」が研究資源名、「114,501 件の薬物構造データ」が規模表現に相当する。

研究資源名の分析では、研究資源の種類や、文字数、アルファベットや漢字などの文字列の種類、および、形態素情報などを調査した。その他の研究資源表現に対しては、研究資源名との出現関係について調査し、特に定義表現と用途表現においては出現に決まったパターンが存在しないか調査した。形態素解析には JUMAN^{*2}[13]、構文解析には KNP^{*3}[12] を使用した。

^{*2}<http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>

^{*3}<http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP>

2.2.2 分析結果

学術論文 50 論文の本文 4249 文から人手で研究資源情報を抽出した結果、異なり数 280 個の研究資源名を取得した。取得した研究資源名の内訳を表 2.1 に示す。なお、形態素解析を行う前にあらかじめ本文テキストを全角化したのが、本論文では、取得した研究資源名が英数字で構成されていれば半角文字で表記する。また、「茶釜」と「ChaSen」や「Wizard-of-Oz」と「WOZ システム」のように、同じ研究資源を指していても表記が異なる場合は本分析では区別して扱った。

表 2.1: 研究資源の種類ごとの研究資源名の取得数

研究資源の種類	取得数	取得した研究資源名の例
ツール・システム	72	LibSVM、茶釜、事故サーベイランスシステム
ウェブサービス	10	Flickr、Yahoo!知恵袋、Google Blog Search
コンピュータ言語	13	OWL、Java、Python
データ・言語資源	10	MDDR、類語検索大辞典日本語シソーラス
電子機器・電子部品	19	i-SOBOT、Vivid910
手法	117	双対尺度法、Ward 法、ベイジアンネットワーク
指標	35	F 値、kappa 統計量、赤池情報量基準
定理・法則	4	ベイズの定理、Hick-Hyman 法則
合計	280	

研究資源名の文字数や形態素数に特徴的な傾向があるか調べるため、取得した 280 個の研究資源名の文字数・形態素数を分析した。結果を図 2.1 と図 2.2 にそれぞれ示す。

調査範囲内においては、研究資源名の最小文字数は 2 であり、最大文字数は 56 であった。2 文字の研究資源名の例としては「F 値」「法造」などが、56 文字の研究資源名の例としては「Self-Organizing Map considering False Neighboring Neuron」が挙げられる。また、3 文字の研究資源名が 35 個と最も多く、以降は文字数が増えるにつれて該当する研究資源名が減少していった。3 文字以上 13 文字以下の範囲では 228 個の研究資源名が該当し、全体の 81.4 % を占める。

形態素数については、最小形態素数は 1 であり、最大形態素数は 13 であった。形態素数が 1 である研究資源名の例としては「SVM」や「カブロボ」などが、形態素数が 13 である研究資源名の例としては、最大文字数と同様、「Self-Organizing Map considering False Neighboring Neuron」が挙げられる。形態素数が 1 である研究資源名が 85 個と最も多く、形態素数が増えるにつれて該当する研究資源名が減少していった。形態素数が 5 以下の研究資源名は 268 個であり、全体の 95.7 % を

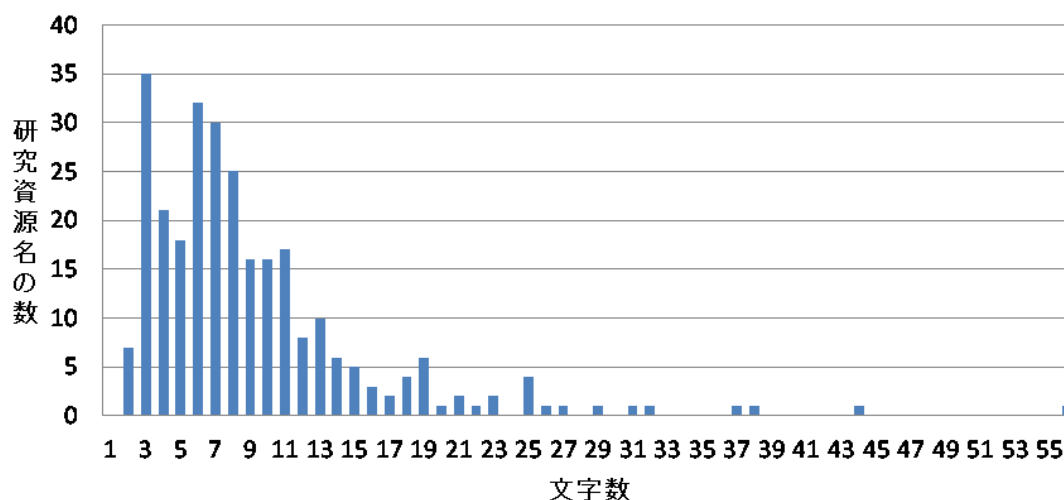


図 2.1: 文字数ごとの研究資源名の分布

占める。

また、研究資源名に使われている文字の種類についても分析を行った。アルファベット、数字、記号、空白、ひらがな、カタカナ、漢字などの特定の種類の文字のみで表記されている研究資源名の数および取得した研究資源名全体における割合、あるいは単に特定の種類の文字を含んでいる研究資源名の数および取得した研究資源名全体における割合を調査した。結果を表 2.2 に示す。

数字のみ、記号のみ、ひらがなのみで表記された研究資源名は、取得した 280 個の中には存在しなかった。また、アルファベットや記号、空白や数字のような、英単語列で用いられる文字のみで表記された研究資源名は、取得した研究資源名全体の約半数を占めていた。

同様に、研究資源名に使われている品詞の種類についても分析を行った。形態素解析に用いた JUMAN は、形態素を 15 種類の品詞に分類する。15 種類の品詞のうち特殊については、さらに空白と記号に分けた。また、JUMAN には、フリー百科事典 Wikipedia^{*4}のエントリなどのウェブ上のテキストから自動的に未知語を獲得して構築した辞書が存在する。このため、名詞については自動獲得されたものか否かを分類した。結果を表 2.3 に示す。

自動獲得された名詞、それ以外の名詞、未定義語は研究資源名に含まれることが多く、該当する研究資源名はそれぞれ全体の 40 % から 50 % であった。判定詞、助動詞、指示詞、副詞、接続詞、連体詞、感動詞を含む研究資源名は、取得した 280 個の中には存在しなかった。動詞や形容詞、助詞、接頭辞や接尾辞を含む研究

^{*4}<http://ja.wikipedia.org/wiki/>

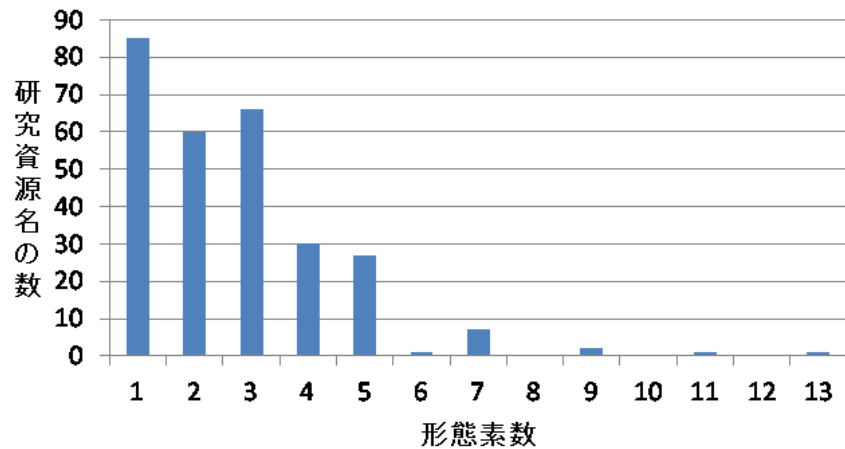


図 2.2: 形態素数ごとの研究資源名の分布

資源名がわずかに存在したが、動詞は「半教師あり学習」の「あり」のような基本連用形、形容詞は「食欲法」の「食欲」のような語幹、助詞は「ベイズの定理」などの接続助詞「の」、接頭辞や接尾辞は「非適合性フィードバック手法」の「非」や「性」など漢字一文字のものに限られていた。

ツールやシステムなどの研究資源は、作成者によって固有の名前を付けられるケースが多いため、Wikipediaのエントリなどのウェブ上のテキストから自動的に獲得した未知語や、その他の未定義語が研究資源名に含まれる可能性が高いと考えられる。

未定義語および自動獲得された名詞をまとめて未知語と見なして、研究資源名を未知語を含んでいるか否かで分類した。その結果を図 2.3 に示す。人手で取得した 280 個の研究資源名のうち、8 割以上が未知語を含んでいることが分かった。

研究資源名以外の研究資源情報としては、127 の定義表現、152 の用途表現、234 の機能・特徴表現、43 の成果物表現、10 の対象表現、4 の規模表現を 50 論文の本文 4249 文から人手で取得した。

これらの研究資源情報は、単独ではどの研究資源に関する情報なのかが分からない。研究資源情報を含む文で研究資源名が言及されていれば、その研究資源名の指す研究資源に関する情報であることはほぼ明白である。すなわち、ある種の研究資源情報が研究資源名と同じ文に現れることが多ければ、その研究資源情報は機械的な抽出が容易だと考えられる。同様に、研究資源名の抽出においても、研究資源名自体の特徴の他に、定義表現や用途表現などの研究資源情報の持つ特徴が有用な手がかりとなると考えられる。そのため、各表現を含む文中に研究資源名が含まれているかどうかという点で研究資源情報を分類した。結果を図 2.4 に示す。

表 2.2: 研究資源名に使われている文字の種類

文字の種類	研究資源名の数	全体の割合
アルファベット・記号・空白・数字のみ	136	0.49
アルファベットのみ	74	0.26
数字のみ	0	0.00
記号のみ	0	0.00
ひらがなのみ	0	0.00
カタカナのみ	24	0.09
漢字のみ	27	0.10
アルファベットを含む	176	0.63
数字を含む	12	0.04
記号を含む	37	0.13
空白を含む	46	0.16
ひらがなを含む	7	0.03
カタカナを含む	91	0.33
漢字を含む	104	0.37

定義表現および用途表現のほとんどが、研究資源名を含む文中に現れている。機能・特徴表現は取得数こと最も多いものの、研究資源名を含む文にもそうでない文にも同程度ずつ現れている。成果物表現、対象表現、規模表現については、他の研究資源情報に比べて取得数が大幅に少ない結果となった。

この結果を鑑みて、研究資源名、定義表現、用途表現の3種の研究資源情報を本研究における抽出対象と定めた。ただし、成果物表現は用途表現の一部と見なすことにした。例えば、文(2.5)では、研究資源名「茶釜」に対して、成果物表現「文末表現」と「抽出する」を合わせて一つの用途表現とする。

研究資源名を含む文中に現れる定義表現および用途表現について取得結果を分析したところ、どちらも文中での出現パターンがある程度決まっていることが確認された。

定義表現の出現パターンは、以下の4種に大きく分かれる。

- [研究資源名-定義表現] 型

研究資源名が定義表現よりも先に現れ、研究資源名から定義表現まででひとつの名詞句となるパターンである。文(2.7)は、[研究資源名-定義表現]型の例である。

(2.7) この出力結果はProtégéといったオントロジー構築ツールでの読み込みが可能であることは確認されている。

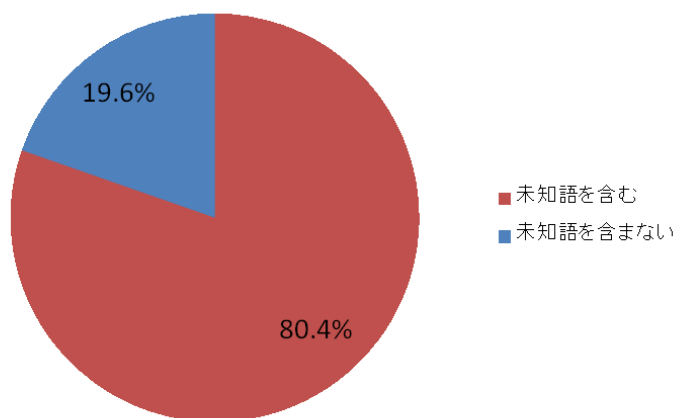


図 2.3: 未知語を含む研究資源名の割合

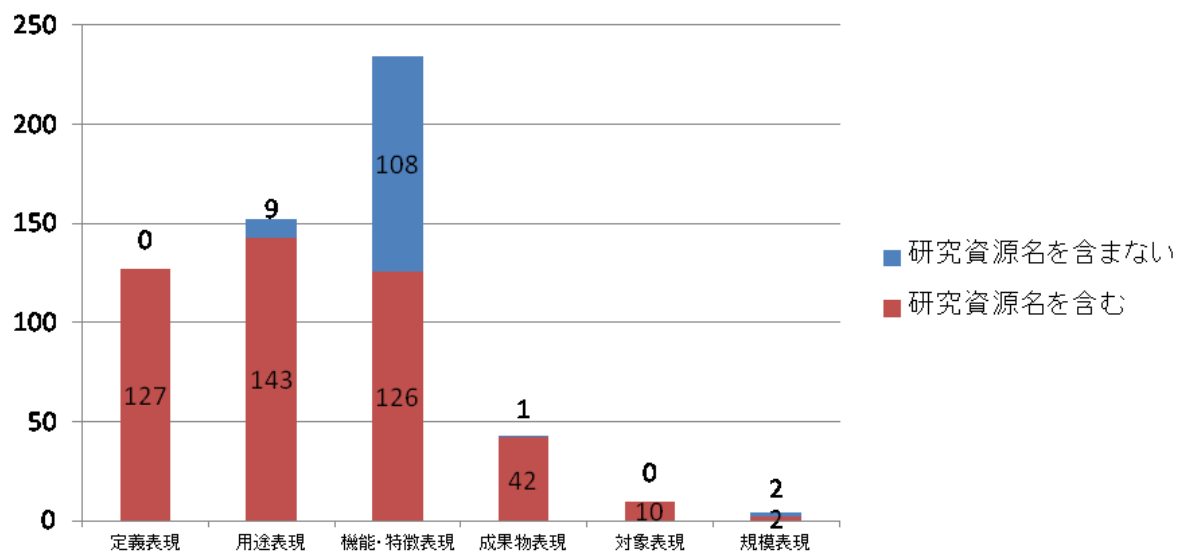


図 2.4: 研究資源情報を含む文

表 2.3: 研究資源名に使われている品詞の種類

品詞の種類	研究資源名の数	全体の割合
空白を含む	45	0.16
記号を含む	39	0.14
動詞を含む	3	0.01
形容詞を含む	5	0.02
判定詞を含む	0	0.00
助動詞を含む	0	0.00
自動取得した名詞を含む	144	0.51
上記以外の名詞を含む	136	0.48
指示詞を含む	0	0.00
副詞を含む	0	0.00
助詞を含む	3	0.01
接続詞を含む	0	0.00
連体詞を含む	0	0.00
感動詞を含む	0	0.00
接頭辞を含む	5	0.02
接尾辞を含む	17	0.06
未定義語を含む	116	0.41

この文では、研究資源名「Protégé」に対して、「オントロジー構築ツール」の部分が定義表現となる。研究資源名と定義表現の間には「といった」や「などの」のような表現が入ることが多い。

- [研究資源名-定義表現-である] 型

このパターンも研究資源名が定義表現よりも先に現れるが、定義表現の直後に判定詞「である」が現れる。文 (2.8) は、[研究資源名-定義表現である] 型の例である。

(2.8) F 値は再現率，適合率の調和平均を取ったものである。

この文では、研究資源名「F 値」に対して、「再現率，適合率の調和平均を取ったもの」の部分が定義表現となる。研究資源名の直後には「とは」や「は」などの表現が現れることが多い。

- [定義表現-研究資源名] 型

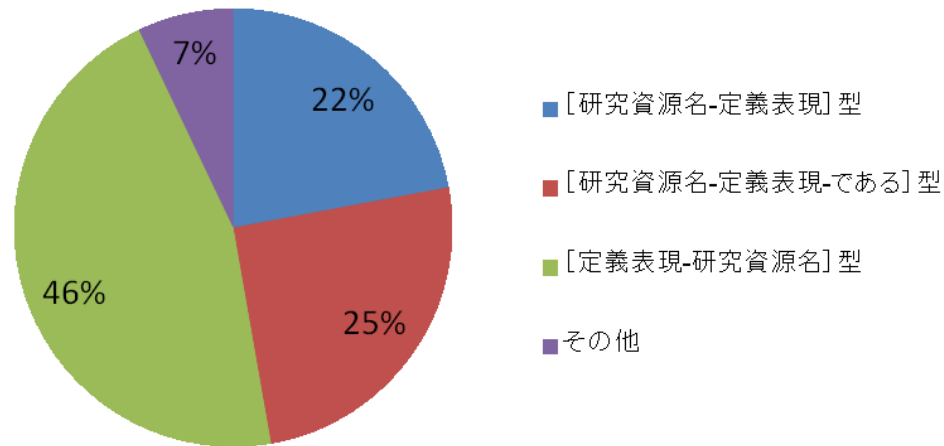


図 2.5: 定義表現の出現パターンの割合

定義表現が研究資源名よりも先に現れるパターンである。文 (2.9) と文 (2.10) は、[定義表現-研究資源名] 型の例である。

(2.9) そのために、高次元データの類似性を可視化する手法である自己組織化マップの改良を行った。

(2.10) 本論文では、周辺環境とユーザの心理状態を記録・再現するデジタルフォトメディア exPhoto を提案した。

文 (2.9) では、研究資源名「自己組織化マップ」に対して、「高次元データの類似性を可視化する手法」の部分が定義表現となる。また、文 (2.10) における研究資源名「exPhoto」と定義表現「周辺環境とユーザの心理状態を記録・再現するデジタルフォトメディア」のように、研究資源名と定義表現が接続している場合もある。

- その他

上記3つの型に属さないパターンである。

研究資源名を含む文中に現れる定義表現全体における、これらの定義表現の出現パターンの割合を図 2.5 に示す。[研究資源名-定義表現] 型、[研究資源名-定義表現-である] 型、[定義表現-研究資源名] 型の3種のパターンで、全体の90%以上を占めている。

用途表現の出現パターンは、以下の4種に大きく分かれる。

- [研究資源名-用途表現] 型

研究資源名が用途表現よりも先に現れるパターンである。文 (2.11) は、[研究資源名-用途表現] 型の例である。

(2.11) ユーザは、exPhoto Viewer を利用して写真の撮影と閲覧を行う。

この文では、研究資源名「exPhoto Viewer」に対して、「写真の撮影と閲覧を行う」の部分が定義表現となる。研究資源名と用途表現の間には「を利用して」や「を用いて」、「によって」のような表現が入ることが多い。

- [用途表現-研究資源名] 型

定義表現が研究資源名よりも先に現れるパターンである。文 (2.12) は、[用途表現-研究資源名] 型の例である。

(2.12) 単語への分割と品詞の推定には ChaSen を利用している。

この文では、研究資源名「ChaSen」に対して、「単語への分割と品詞の推定」の部分が用途表現となる。用途表現と研究資源名の間には「には」や「するための」のような表現が入ることが多い。

- [成果物表現-研究資源名-動詞・サ変名詞] 型

用途表現が成果物表現と動詞・サ変名詞に分かれて、研究資源名の前後に現れるパターンである。文 (2.13) は、[成果物表現-研究資源名-動詞・サ変名詞] 型の例である。

(2.13) 人の位置と ID は床面 RFID センサによって取得される。

この文では、研究資源名「底面 RFID センサ」に対して、成果物表現「人の位置と ID」とサ変名詞「取得」が合わせてひとつの用途表現となる。

- その他

上記 3 つの型に属さないパターンである。

研究資源名を含む文中に現れる用途表現全体における、これらの用途表現の出現パターンの割合を図 2.6 に示す。[研究資源名-用途表現] 型、[用途表現-研究資源名] 型、[成果物表現-研究資源名-動詞・サ変名詞] 型の 3 種のパターンで、全体の 85 % を占めている。

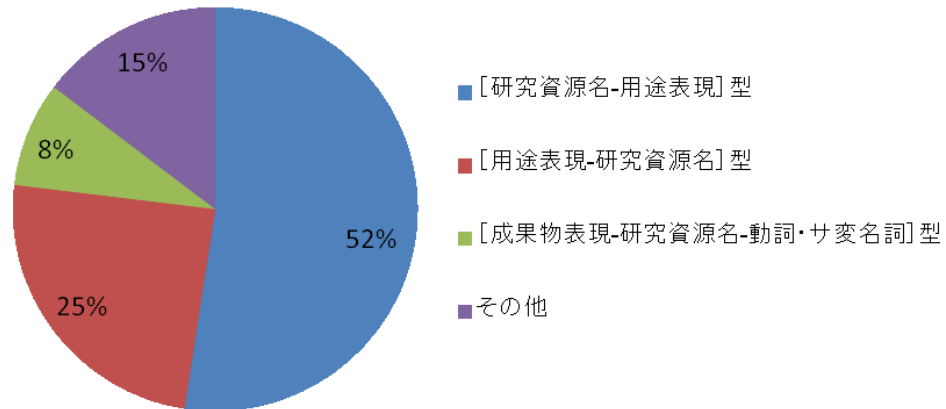


図 2.6: 用途表現の出現パターンの割合

2.3 研究資源情報の利用

研究資源情報を研究資源の検索・選択に利用することで、効率的な研究活動と研究資源の広範な活用を促す。大量の学術論文中に散在している状態の研究資源情報は一覧性に欠けており、検索することも困難である。学術論文から機械的に収集した研究資源情報をデータベース化すれば、キーワードを検索クエリとして、研究資源情報と文字列照合することによって、目的に合った研究資源を検索することができる。データベース化によって研究資源情報が研究資源の利用者間で共有されるようになるため、研究分野全体のレベル向上につながる。

同様の機能を持つ研究資源は同じような用途で使用されると考えられるため、検索によって複数の類似した研究資源が見つかる可能性も十分考えられる。複数の研究資源から最適のものを選択する際にも、データベース化された研究資源情報は有用である。各研究資源がどのようなものであるか、どのような用途で使われているか、などの情報がまとまっていれば、その量や内容を比べることによって複数の研究資源を容易に比較検討することができる。

また、本研究では研究資源情報の自動抽出にあたって学術論文に着目しているが、学術論文から研究資源情報を抽出することは、研究資源と学術論文（つまり、研究資源の活用事例）を紐付けするということでもある。研究資源情報から元の論文を辿れるようにすれば、論文を参照することで、機械的に整理してデータベース化できないような詳細な情報も得られると考えられる。

本研究では、新たな研究資源情報を抽出する際にも、既知の研究資源情報を活用することを考えている。研究資源名の抽出には、その形態的特徴の他に、定義

表現および用途表現の持つ特徴が有用な手がかりとなると考えられる。同様に定義表現および用途表現の抽出において、研究資源名が有用な手がかりとなる。

第3章 研究資源情報の抽出手法

3.1 抽出手法の概要

本稿で提案する抽出手法は、研究資源名の抽出、および、定義表現および用途表現の抽出、の2つのフェーズに分かれる。

研究資源名の抽出には、研究資源名の形態的特徴が手がかりとなる。2章で述べたように、研究資源には作成者が固有の名前を付けるケースが多いため、研究資源名には未知語が含まれる可能性が高い。また、「法」や「アルゴリズム」など、研究資源名の末尾によく登場する表現も抽出の手がかりとなると考えられる。しかし、「自己組織化マップ」のように、未知語や末尾表現を含まない研究資源名も存在するため、未知語や末尾表現のみを手がかりにするのでは不十分である。一方で、定義表現や用途表現を含む文には研究資源名が記述されている場合が多く、定義表現および用途表現の持つ特徴が有用な手がかりになると考えられる。

そこで、研究資源名の抽出手法としては、まず、研究資源名の候補となる形態素列を、研究資源名の形態的特徴や既知の定義表現および用途表現などを手がかりにして、論文の本文テキスト中から取得する。次に、取得した研究資源名候補の集合から、SVMを用いて正解となる研究資源名を選別する。

一方、定義表現および用途表現の抽出には、2章で述べたように、研究資源名が有用な手がかりとなる。また、研究資源名と定義表現が一文中に含まれている場合、研究資源名と定義表現の間には「とは」や「である」などの表現がよく出現する。同様に、一文中の研究資源名と用途表現の間には「を用いて」や「するための」などの表現がよく出現する。これらの表現（以下、それぞれを**定義の手がかり表現**、**用途の手がかり表現**と呼ぶ）は、研究資源名、定義表現および用途表現と、接続関係あるいは係り受け関係にあるため、研究資源情報の抽出において有用な手がかりとなる。

そこで、定義表現および用途表現の抽出手法としては、まず、本文テキストから研究資源名を含む文を選択する。研究資源名の前後に接続している手がかり表現を参考にして、接続関係あるいは係り受け関係を利用して定義表現や用途表現を抽出する。

3.2 PDFのテキストへの変換

抽出の前処理として、学術論文のPDFデータから本文テキストを抽出し、文への分割を行う。

まず、pdftohtmlを用いて論文のPDFファイルをHTML形式に変換する。変換後のHTMLには、文字列のフォントサイズや位置などの情報が含まれている。図3.1に、pdftohtmlの出力例を示す。

多くの学術論文、特に全国大会のような論文の形式があらかじめ指定されている場で発表された論文には、レイアウトに共通の特徴がある。例として、1ページ目の上部にタイトルや著者名、所属が記述されることや、章題は概ね本文よりも大きなフォントサイズであることが挙げられる。これらの特徴から作成したヒューリスティックルールを利用して、HTML中の文字列の内容・フォントサイズ・位置情報を手がかりにしてHTMLから本文テキストを抽出する。

また、研究資源情報の自動抽出には、形態素解析だけではなく構文解析も必要とするため、本文テキストを文単位に分割し、文献参照の番号のような括弧で囲まれた部分は削除する。

HTMLからの本文テキストの抽出・文単位の分割には、独自に作成したプログラムを用いる。2008年度人工知能学会全国大会の発表論文集に収録されている論文をpdftohtmlによって変換した結果を観察して、ヒューリスティックルールを作成した。図3.2に本文テキストの抽出例を示す。

プログラムの性能を評価するため、2006年度人工知能学会全国大会の発表論文集から無作為に選んだ20論文の本文テキストの抽出を行ったところ、抽出結果として取得した1546文のうち1440文がPDFファイルから人手で作成した正解データ(1584文)と一致し、適合率は93.1%、再現率は90.9%であった。

3.3 研究資源名の自動抽出

3.3.1 研究資源名候補の取得

学術論文の本文テキストから、以下の3種類の手法によってそれぞれ取得した形態素列を研究資源名の候補とする。

1. 未知語に基づく抽出

まず、形態素解析された本文テキストから未知語を取得する。形態素解析に用いたJUMANには、Wikipediaのエントリなどのウェブ上のテキストから自動的に未知語を獲得して構築した辞書(Wikipedia.dicおよびAuto.dic)が付属している。本手法では、JUMANで未定義語と判定された形態素・Wikipediaから抽出した辞書Wikipedia.dicに収録されている名詞・自動辞書Auto.dic

```

1 <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
2 <HTML>
3 <HEAD>
4 <TITLE>Page 1</TITLE>
5
6 <META http-equiv="Content-Type" content="text/html; charset=UTF-8">
7 <DIV style="position:relative;width:892;height:1263;">
8 <STYLE type="text/css">
9 <!--
10 .ft0{font-size:11px;font-family:Times;color:#000000;}
11 .ft1{font-size:19px;font-family:Times;color:#000000;}
12 .ft2{font-size:22px;font-family:Times;color:#000000;}
13 .ft3{font-size:15px;font-family:Times;color:#000000;}
14 .ft4{font-size:9px;font-family:Times;color:#000000;}
15 .ft5{font-size:10px;font-family:Times;color:#000000;}
16 .ft6{font-size:12px;font-family:Times;color:#000000;}
17 .ft7{font-size:11px;line-height:14px;font-family:Times;color:#000000;}
18 .ft8{font-size:10px;line-height:17px;font-family:Times;color:#000000;}
19 -->
20 </STYLE>
21 </HEAD>
22 <BODY bgcolor="#A0A0A0" vlink="blue" link="blue">
23 <IMG width="892" height="1263" src="100001001.png" alt="background image">
24 <DIV style="position:absolute;top:67;left:190"><nobr><span class="ft0">The 22nd Annual Conference of
the Japanese Society for Artificial Intelligence, 2008</span></nobr></DIV>
25 <DIV style="position:absolute;top:101;left:793"><nobr><span class="ft1">3J1-1</span></nobr></DIV>
26 <DIV style="position:absolute;top:132;left:145"><nobr><span class="ft2">人工知能学会全国大会プログラ
ム自動作成ツールの開発</span></nobr></DIV>
27 <DIV style="position:absolute;top:174;left:176"><nobr><span class="ft3">Development of a Tool for
Automatically Generating JSAI2008 Program</span></nobr></DIV>
28 <DIV style="position:absolute;top:224;left:143"><nobr><span class="ft3">西村 直史</span></nobr></DIV>
29 <DIV style="position:absolute;top:244;left:126"><nobr><span class="ft4">Naofumi
Nishimura</span></nobr></DIV>
30 <DIV style="position:absolute;top:224;left:273"><nobr><span class="ft3">徳永 亮</span></nobr></DIV>
31 <DIV style="position:absolute;top:244;left:262"><nobr><span class="ft4">Ryo
Tokunaga</span></nobr></DIV>
32 <DIV style="position:absolute;top:224;left:369"><nobr><span class="ft3">大田 直樹</span></nobr></DIV>
33 <DIV style="position:absolute;top:244;left:374"><nobr><span class="ft4">Naoki
Ohta</span></nobr></DIV>
34 <DIV style="position:absolute;top:224;left:471"><nobr><span class="ft3">櫻井 祐子</span></nobr></DIV>
35 <DIV style="position:absolute;top:244;left:471"><nobr><span class="ft4">Yuko
Sakurai</span></nobr></DIV>
36 <DIV style="position:absolute;top:224;left:587"><nobr><span class="ft3">岩崎 敦</span></nobr></DIV>
37 <DIV style="position:absolute;top:244;left:572"><nobr><span class="ft4">Atsushi
Iwasaki</span></nobr></DIV>
38 <DIV style="position:absolute;top:224;left:698"><nobr><span class="ft3">横尾 真</span></nobr></DIV>
39 <DIV style="position:absolute;top:244;left:686"><nobr><span class="ft4">Makoto
Yokoo</span></nobr></DIV>
40 <DIV style="position:absolute;top:280;left:311"><nobr><span class="ft3">九州大学大学院システム情報科
学府</span></nobr></DIV>
41 <DIV style="position:absolute;top:299;left:321"><nobr><span class="ft4">Graduate School of ISEE,
Kyushu University</span></nobr></DIV>
42 <DIV style="position:absolute;top:335;left:115"><nobr><span class="ft0">This paper develops an
automated tool for generating the JSAI2008 conference program. A lot of
conference</span></nobr></DIV>
43 <DIV style="position:absolute;top:350;left:102"><nobr><span class="ft7">programs are manually
generated based on informations, such as categories and affiliations, submitted from the<br>authors.
However, at a huge conference, it is very exhaustive to manually generate the conference program,
due<br>to various constraint condition or criteria, e.g., papers in each session must have consistent
research topics. Thus,<br>this paper formalizes this problem as a constraint
satisfaction/optimization problem, develops an automated tool,<br>and generates the JSAI2008

```

図 3.1: pdftohtml の出力例

HTML化された論文

```

44 <DIV style="position:absolute;top:482;left:72"><no><span class="ft3">1.</span></no></DIV>
45 <DIV style="position:absolute;top:482;left:108"><no><span class="ft3">はじめに</span></no></DIV>
46 <DIV style="position:absolute;top:492;left:86"><no><span class="ft5">大規模な学術会議では、複数の論文発表が</span></no></DIV>
47 <DIV style="position:absolute;top:492;left:329"><no><span class="ft0">1</span></no></DIV>
48 <DIV style="position:absolute;top:492;left:339"><no><span class="ft5">つのセッション</span></no></DIV>
49 <DIV style="position:absolute;top:510;left:72"><no><span class="ft8">と呼ばれる単位にまとめられ、複数のセッションが並行して実<br>施される。従
来、会議プログラムは、論文投稿時に入力される<br>登録情報（論文タイトル、キーワード等）に基づいて手作業で<br>決定されていた
が、</span></no></DIV>
50 <DIV style="position:absolute;top:564;left:181"><no><span class="ft5">、様々な制約条件／価値基準を満足する</span></no></DIV>
51 <DIV style="position:absolute;top:582;left:72"><no><span class="ft5">プログラムを作成することは非常に困難であった。例えば、</span></no></DIV>
52 <DIV style="position:absolute;top:582;left:400"><no><span class="ft0">2007</span></no></DIV>
53 <DIV style="position:absolute;top:600;left:72"><no><span class="ft8">年度人工知能学会全国大会では、論文キーワードをベースに手<br>作業でセッショ
ンが構成されており、セッション毎に発表論文<br>数が大きく異なる、発表者の殆どが同じ所属となるセッション<br>が生じる等の不都合が生じている。また、
会議プログラム作成<br>問題は、スケジューリング問題の一つと捉えることができる。<br>スケジューリング問題は従来より盛んに研究されてきた分野<br>であり</span></no></DIV>
54 <DIV style="position:absolute;top:708;left:114"><no><span class="ft0">[</span></no></DIV>
55 <DIV style="position:absolute;top:708;left:117"><no><span class="ft5">松井</span></no></DIV>
56 <DIV style="position:absolute;top:708;left:149"><no><span class="ft0">89, Easton 99</span></no></DIV>
57 <DIV style="position:absolute;top:708;left:234"><no><span class="ft5">、様々な手法が提案されてきたが、</span></no></DIV>
58 <DIV style="position:absolute;top:725;left:72"><no><span class="ft5">会議プログラム作成を対象とした研究はほとんどなかった。</span></no></DIV>
59 <DIV style="position:absolute;top:743;left:86"><no><span class="ft5">そこで本論文では、制約充足／最適化のテクニックを用い</span></no></DIV>

```



本文の抽出・文への分割

大規模な学術会議では、複数の論文発表が1つのセッションと呼ばれる単位にまとめられ、複数のセッションが並行して実施される。従来、会議プログラムは、論文投稿時に入力される登録情報に基づいて手作業で決定されていたが、様々な制約条件／価値基準を満足するプログラムを作成することは非常に困難であった。

例えば、2007年度人工知能学会全国大会では、論文キーワードをベースに手作業でセッションが構成されており、セッション毎に発表論文数が大きく異なる、発表者の殆どが同じ所属となるセッションが生じる等の不都合が生じている。

また、会議プログラム作成問題は、スケジューリング問題の一つと捉えることができる。

スケジューリング問題は従来より盛んに研究されてきた分野であり、様々な手法が提案されてきたが、会議プログラム作成を対象とした研究はほとんどなかった。

図 3.2: 本文テキストの抽出例

に収録されている名詞を未知語とみなす。Wikipedia.dic や Auto.dic に収録されている語の中には、「茶釜」や「キャプチャ」など、研究資源名あるいはその一部となるような、標準の辞書には載っていない専門的な用語が存在すると考えられる。

次に、取得した未知語の前後に、研究資源名の一部になりうる形態素が存在するか否かを判断する。以下の8種類の形態素を、研究資源名の一部になりうる形態素とみなす。

- 未知語
- 名詞
- 接頭辞
- 接尾辞
- 形容詞の語幹
- 動詞の基本連用形
- 空白
- 記号（句読点・括弧類を除く）

...	おいて	Vicon	モーション	キャプチャ	システム	を	...
	動詞 (タ行連用テ形)	未知語	名詞	名詞 (Wikipedia.dictに収録)	名詞	助詞	

図 3.3: 未知語に基づく研究資源名候補の抽出

...	に	遺伝	的	アルゴリズム	を	...
	助詞	名詞	接尾辞	既知の末尾表現	助詞	

図 3.4: 既知の末尾表現に基づく研究資源名候補の抽出

取得した未知語の前後に研究資源名の一部になりうる形態素が存在すればそれも取得し、得られた形態素列を研究資源名候補とする。取得した未知語の前後に研究資源名の一部になりうる形態素が存在しない場合は、取得した未知語のみを研究資源名候補とする。

図 3.3 の場合、未知語「Vicon」を含む形態素列「Vicon モーションキャプチャシステム」を研究資源名候補とする。

2. 既知の末尾表現に基づく抽出

まず、本文テキストから、「法」や「アルゴリズム」など、研究資源名の末尾によく登場する既知の表現を取り出す。末尾表現の前に、名詞・接頭辞・接尾辞など研究資源名の一部になりうる形態素が存在すればそれも取得し、得られた形態素列を研究資源名候補とする。

図 3.4 の場合、既知の末尾表現「アルゴリズム」が末尾となる形態素列「遺伝的アルゴリズム」を研究資源名候補とする。

3. 既知の定義表現および用途表現に基づく抽出

同一の定義表現および用途表現は、どちらもある程度の長さを持った形態素列であり、学术论文の集合中に複数回出現することは極めて稀であるため、定義表現および用途表現そのものを研究資源名の抽出に利用することはできない。そこで、定義表現は末尾の名詞（以下、**定義名詞**と呼ぶ）を、用途表現は末尾の動詞・サ変名詞（以下、**用途動詞**と呼ぶ）を研究資源名抽出の手がかりとする。例えば、定義表現「2 値分類問題に対する有効な分類モデルの導出手法」の場合は定義名詞「手法」が、用途表現「記事を主観的な意見

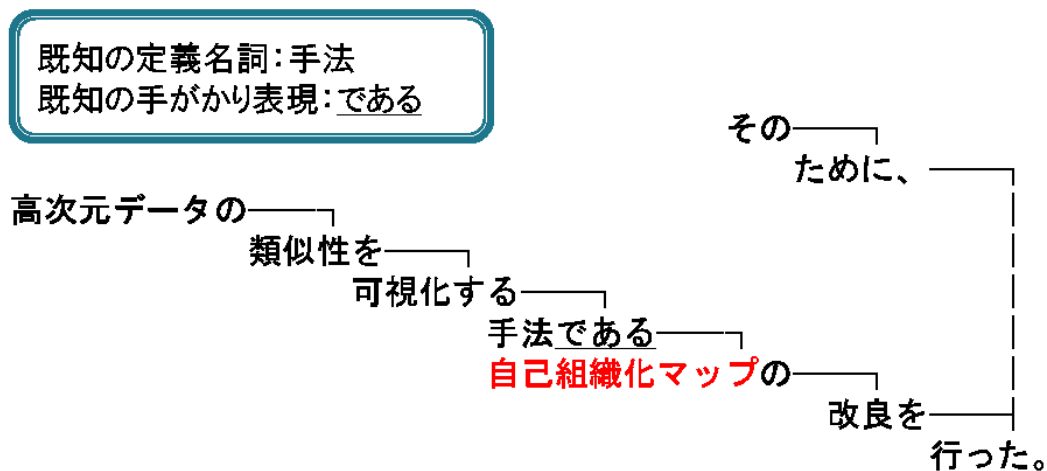


図 3.5: 定義表現に基づく研究資源名候補の抽出

を含むレビュー記事と非レビュー記事に分類する」の場合は用途動詞「分類」が手がかりとなる。

また、3.1 節で述べたように、定義の手がかり表現および用途の手がかり表現は、研究資源名・定義表現・用途表現と、接続関係あるいは係り受け関係にあるため、研究資源情報のみならず、研究資源名の抽出においても有用な手がかりとなると考えられる。

抽出手法としては、まず、構文解析された本文テキストから既知の手がかり表現を発見する。手がかり表現の直前の形態素が既知の定義名詞あるいは用途動詞であった場合は、手がかり表現を含む文節の係り先の最後の形態素から、名詞・接頭辞・接尾辞など研究資源名の一部となりうる形態素を取得していき、最終的に得られた形態素列を研究資源名候補とする。ただし、最終的な取得結果が未知語でない名詞1形態素のみであった場合は、「データ」や「人手」のような、学术论文において一般的に使われる用語である可能性が高いため、研究資源名候補に含めない。

図 3.5 中の木構造は、「そのために、高次元データの類似性を可視化する手法である自己組織化マップの改良を行った。」という文の構文解析結果である。図 3.5 の場合は、手がかり表現「である」の直前の形態素が定義名詞「手法」であるので、「である」を含む文節の係り先から、研究資源名候補「自己組織化マップ」を取得する。

定義の手がかり表現「として」や用途の手がかり表現「するために」などの場合は、手がかり表現と定義名詞あるいは用途動詞は接続関係にあるが、研究資源名は手がかり表現を含む文節の係り先にはなく、係り先の動詞に対し

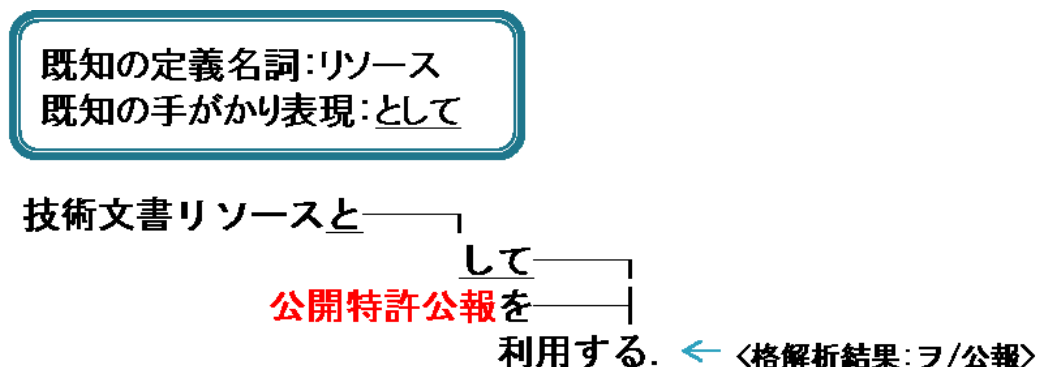


図 3.6: 格解析結果を利用した研究資源名候補の抽出

て格関係がヲ格となる位置に存在する。この場合は、KNP による格解析の結果を用いて、手がかり表現を含む文節の係り先に対してヲ格となる形態素から、名詞・接頭辞・接尾辞など研究資源名の一部となりうる形態素を取得していき、最終的に得られた形態素列を研究資源名候補とする。ただし、最終的な取得結果が未知語でない名詞 1 形態素のみであった場合は、学术论文において一般的に使われる用語である可能性が高いため、研究資源名候補に含めない。

図 3.6 中の木構造は、「技術文書リソースとして公開特許公報を利用する。」という文の構文解析結果である。図 3.6 の場合は、手がかり表現「として」の直前の形態素が定義名詞「リソース」であるので、手がかり表現「として」の係り先の格解析結果を利用して、係り先の「利用する」に対して格関係がヲ格となる「公報」から、研究資源名候補「公開特許公報」を取得する。

手がかり表現を含む文節の係り先に、既知の定義名詞あるいは用途動詞が含まれている場合は、手がかり表現の直前の形態素から研究資源名の一部となりうる形態素を取得していき、最終的に得られた形態素列を研究資源名候補とする。ただし、最終的な取得結果が未知語でない名詞 1 形態素のみであった場合は、研究資源名候補に含めない。

図 3.7 中の木構造は、「不完備情報ゲームを用いてエージェントの意思決定モデルを構築する。」という文の構文解析結果である。図 3.7 の場合は、手がかり表現「を用いて」の一部「用いて」の係り先に用途動詞「構築」があるので、「を用いて」の直前の形態素から、研究資源名候補「不完備情報ゲーム」を取得する。

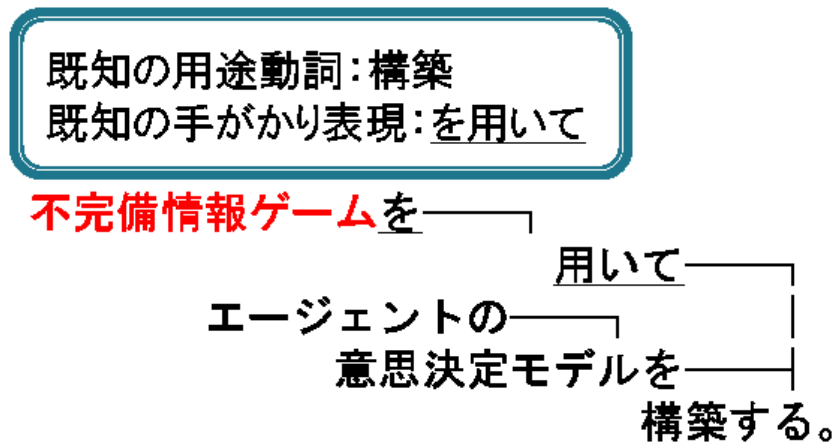


図 3.7: 用途表現に基づく研究資源名候補の抽出

3.3.2 SVMによる研究資源名の取得

前項で述べた3種類の抽出手法では、研究資源名ではないものも多く取得されてしまう。例えば、未知語に基づく抽出では、「メタ認知」や「NTCIR」などの専門用語・固有名詞、「fpk」などの文字式が取得される。既知の末尾表現に基づく抽出では、「解法」や「近似アルゴリズム」などの一般的な用語が取得される。既知の定義表現および用途表現に基づく抽出では、「大脳基底核」のように、それがどのようなものなのか論文中で説明が必要な専門用語や、「提案手法」や「画像データ」のように何らかの用途に使われるが研究資源名とは言えないものが取得される。このように、研究資源名の抽出において有用な手がかりは多数あるが、どれも単独では不十分である。そこで、提案手法では、素性としてこれらの手がかりを利用し、SVM (Support Vector Machine) [1] を用いて研究資源名候補の集合から研究資源名を選別する。

なお、SVMによる研究資源名の取得の際には、3種類の手法で抽出した研究資源名候補をまとめて学習に用いるのではなく、3種類の手法で抽出した研究資源名候補の集合それぞれにおいて学習を行い、研究資源名を選別する。すなわち、未知語に基づく抽出手法によって取得した研究資源名候補の集合、既知の末尾表現に基づく抽出手法によって取得した研究資源名候補の集合、定義表現および用途表現に基づく抽出手法によって取得した研究資源名候補の集合それぞれにおいて学習モデルを生成し、研究資源名の選別を行う。それぞれの学習モデルで抽出した研究資源名の和集合を出力とする。

本研究では、libSVM^{*1}を利用して学習を行う。学習用データの作成には、学術論文から人手で抽出した研究資源名・定義表現・用途表現・手がかり表現を用い

^{*1}<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

る。前節で述べた手法によって研究資源名候補を取得し、人手で抽出した研究資源名に該当するものを正例、該当しないものを負例とする。

学習には、以下の 15 種類の素性を用いる。

1. 研究資源名候補の文字数に関する素性（次元数:1）

研究資源名候補の文字数に関する素性では、2 文字以下であれば 1、3 文字から 13 文字であれば 2、14 文字以上であれば 3 とする。この素性は、2 章で述べた研究資源情報に関する分析において、3 文字から 13 文字の研究資源名が全体の 81.4 % を占めたという調査結果に基づく。

2. 研究資源名候補を構成する文字の種類に関する素性（次元数:6）

- ひらがなの有無
- カタカナの有無
- 漢字の有無
- アルファベットの有無
- 数字の有無
- 記号の有無

研究資源名候補に、ひらがな・カタカナ・漢字・アルファベット・数字・その他の記号が含まれているかというものである。それぞれ、含まれていれば 1、含まれていなければ 0 とする。

3. 研究資源名候補の形態素数に関する素性（次元数:1）

1 形態素であれば 1、2 形態素であれば 2、3 形態素以上であれば 3 とする。この素性は、予備実験として 2008 年度人工知能学会全国大会の発表論文集中に収録されている 50 論文から提案手法を用いて研究資源名候補を抽出した結果、2 形態素の研究資源名候補中の研究資源名の割合が、他に比べて低かったことに基づく。

事前実験の抽出結果における形態素数ごとの研究資源名候補中の研究資源名の割合を表 3.1 に示す。

4. 未知語に関する素性（次元数:3）

- KNP の判定結果における未知語の有無
- JUMAN に付属する Wikipedia.dic 内の語の有無
- JUMAN に付属する Auto.dic 内の語の有無

表 3.1: 研究資源名候補中の研究資源名の割合

形態素数	研究資源名数	研究資源名候補数	割合
1	69	604	0.11
2	51	980	0.05
3以上	99	889	0.11

研究資源名候補が、KNP の判定結果における未知語・JUMAN に付属する Wikipedia.dic 内の語・Auto.dic 内の語を含むか否かというものである。それぞれ含まれていれば1、含まれていなければ0とする。

5. 既知の末尾表現に関する素性（次元数:3）

研究資源名候補が既知の末尾表現を含むか否かという2値である。事前実験として2008年度人工知能学会全国大会の発表論文集に収録されている50論文から提案手法を用いて研究資源名候補を抽出した結果から、末尾表現ごとの研究資源名候補中の研究資源名の割合の大小に基づいて、末尾表現を3つのグループに分けた。

6. 数学記号に関する素性（次元数:1）

研究資源名候補を含むすべての文において、研究資源名候補を含む文節の中に「=」や「 $\in$ 」などの数式や論理式に使われる記号が一つも含まれていないかというものである。未知語に基づく研究資源名候補の抽出では、「fp」や「cos」のような文字式や数学記号も取得されてしまうが、これらは「=」や「 $\in$ 」などの数式や論理式に使われる記号と同じ文節中に現れることが多いと考えられる。

事前実験として2008年度人工知能学会全国大会の発表論文集に収録されている50論文から提案手法を用いて研究資源名候補を抽出した結果では、研究資源名候補中の研究資源名のうち、同じ文節に数学記号が含まれているものは1つも存在しなかった。

7. 括弧表現に関する素性（次元数:3）

- 直後に括弧『（』が接続されている文があるか
- 直後に格括弧『[』が接続されている文があるか
- 直後に鉤括弧閉じ『』が接続されている文があるか

研究資源名候補を含むすべての文において、研究資源名の直後に、括弧・角括弧・鉤括弧閉じが接続している文が1文でもあれば1、1文もなければ0と

する。これは、学術論文では、ツールやシステムなどの研究資源名の直後に研究資源名の解説や文献の引用番号が書かれることや、研究資源名が鉤括弧によって強調して表記されることが多いという特徴を反映したものである。

8. アスタリスクに関する素性（次元数:1）

研究資源名候補を含むすべての文において、研究資源名の直後にアスタリスクが接続している文が1文でもあれば1、1文もなければ0とする。これは、学術論文では、ツールやシステムなどの研究資源名の直後に注釈を示す記号が書かれるという特徴を反映したものである。

9. 定義の手がかり表現に関する素性（次元数:1）

研究資源名候補の直前に定義の手がかり表現が接続している文・研究資源名候補の最後の形態素の係り先に定義の手がかり表現がある文・定義の手がかり表現の係り先のヲ格が研究資源名候補の一部である文のいずれかが存在すれば1、存在しなければ0とする。

10. 用途の手がかり表現に関する素性（次元数:n）

研究資源名候補の直前に用途の手がかり表現が接続している文・研究資源名候補の最後の形態素の係り先に用途の手がかり表現がある文・用途の手がかり表現の係り先のヲ格が研究資源名候補の一部である文のいずれかが存在しなければ0、1文存在すれば1、複数文存在すれば2とする。

また、次元数 n は、既知の用途の手がかり表現を、手がかり表現中の動詞・名詞などでグループ分けしたグループ数である。例えば、用途の手がかり表現「を用い」「を用いて」「に用いられる」は、すべてに動詞「用いる」が含まれるので、同じグループとする。

11. 用途の手がかり表現の種類数に関する素性（次元数:1）

研究資源名候補の直前に用途の手がかり表現が接続している文・研究資源名候補の最後の形態素の係り先に用途の手がかり表現がある文・用途の手がかり表現の係り先のヲ格が研究資源名候補の一部である文が存在した場合の、用途の手がかり表現に該当するグループ数に関する素性である。グループが1種類しかなければ1、複数種類あれば2とする。研究資源名候補の直前に用途の手がかり表現が接続している文・研究資源名候補の最後の形態素の係り先に用途の手がかり表現がある文・用途の手がかり表現の係り先のヲ格が研究資源名候補の一部である文が存在しない場合は0とする。

2008年度人工知能学会全国大会の発表論文集に収録されている50論文から提案手法を用いて研究資源名候補を抽出した結果における、用途の手がかり表現の種類数ごとの研究資源名候補中の研究資源名の割合を表3.2に示す。

表 3.2: 用途の手がかり表現の種類数ごとの研究資源名の割合

種類数	研究資源名数	研究資源名候補数	割合
0	134	2385	0.06
1	65	349	0.19
2以上	24	58	0.41

12. 「使用」を意味する動詞に関する素性（次元数:m）

研究資源名候補の最後の形態素の係り先に、「用いる」「使う」「利用」などの、「使用」を意味する動詞・サ変名詞が含まれる文が存在しなければ0、1文存在すれば1、複数文存在すれば2とする。「使用」を意味する動詞としては、用途の手がかり表現をグループ分けする際に用いた、用途の手がかり表現中の動詞・サ変名詞を用いる。

また、次元数 n は、用途の手がかり表現をグループ分けする際に用いた、用途の手がかり表現中の動詞・サ変名詞の種類数である。

13. 判定詞「だ」に関する素性（次元数:1）

研究資源名候補の最後の形態素の係り先に、判定詞「だ」が含まれる文が存在すれば1、存在しなければ0とする。このとき、活用形は問わない。この素性は、定義表現の出現パターンの一つである「研究資源名-定義表現である」型のKNPによる構文解析結果において、研究資源名の係り先が判定詞「だ」のデアル列基本形「である」であることに基づく。

14. 定義名詞に関する素性（次元数:2）

- 文中で直前の形態素が定義名詞であるか
- 係り先が定義名詞であるか

研究資源名候補の直前に定義名詞が接続している文が存在するか否かと、研究資源名候補の最後の形態素の係り先に定義名詞が含まれる文が存在するか否かという2点において、それぞれ存在すれば1、存在しなければ0とする。

15. 用途動詞に関する素性（次元数:1）

研究資源名候補の最後の形態素の係り先に用途動詞が含まれる文が存在すれば1、存在しなければ0とする。

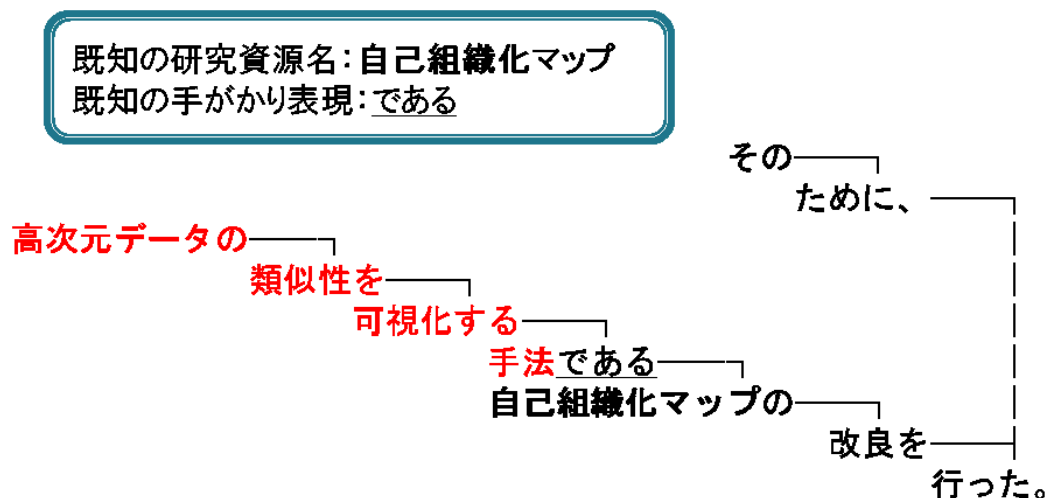


図 3.8: 手がかり表現「である」に基づく定義表現の抽出

3.4 研究資源情報の自動抽出

3.4.1 定義表現の抽出

まず、本文テキストから既知の研究資源名を含む文を選択し、研究資源名の前後に「である」や「などの」のような定義の手がかり表現が接続しているかどうかを判断する。

研究資源名の直前に手がかり表現が接続していたら、構文木において、手がかり表現の最初の形態素を含む文節を根とした部分木に対応する形態素列を定義表現として抽出する。ただし、手がかり表現の部分は定義表現に含めない。このケースは、2.2.2 項で述べた定義表現の出現パターン〔定義表現-研究資源名〕型に対応する。

図 3.8 中の木構造は、「そのために、高次元データの類似性を可視化する手法である自己組織化マップの改良を行った。」という文の構文解析結果である。図 3.8 の場合は、研究資源名「自己組織化マップ」の直前に定義の手がかり表現「である」が接続しているため、「である」を含む文節を根とした部分木に対応する形態素列「高次元データの類似性を可視化する手法である」から、手がかり表現の部分を削除した「高次元データの類似性を可視化する手法」の部分を定義表現として抽出する。

研究資源名の直後に手がかり表現が接続していたら、手がかり表現の直後の形態素から、手がかり表現の最後の形態素を含む文節の係り先の主辞までの形態素列を定義表現として抽出する。主辞とは、文節の中で主となる意味を持つ形態素のことである。主辞の判定は KNP によって行う。このケースは、2.2.2 項で述べ

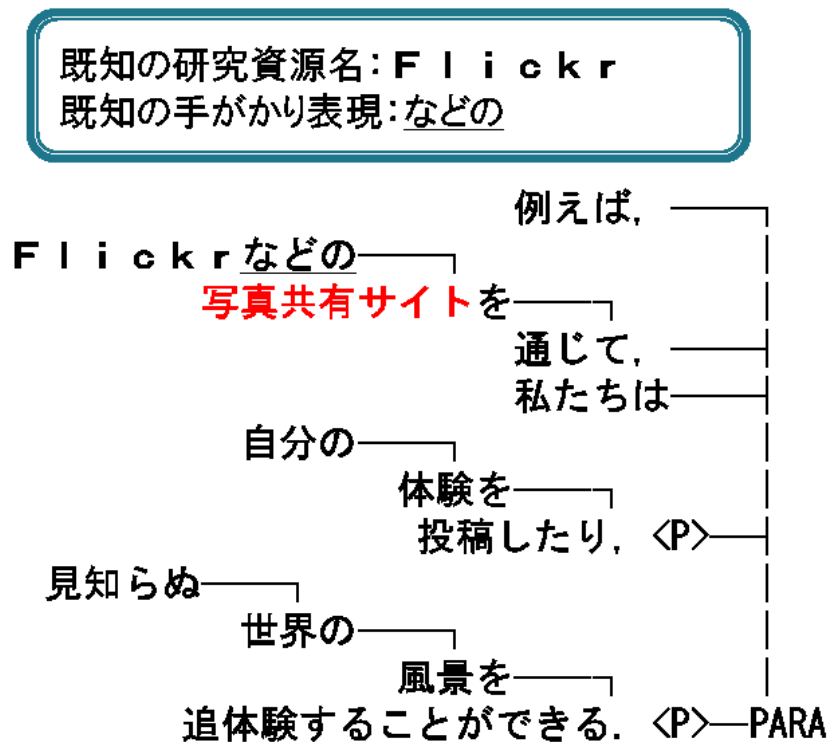


図 3.9: 手がかり表現「などの」に基づく定義表現の抽出

た定義表現の出現パターン〔研究資源名-定義表現〕型および〔研究資源名-定義表現である〕型に対応する。

図 3.9 中の木構造は、「例えば、Flickr などの写真共有サイトを通じて、私たちは自分の体験を投稿したり、見知らぬ世界の風景を追体験することができる。」という文の構文解析結果である。図 3.9 の場合は、研究資源名「Flickr」の直後に定義の手がかり表現「などの」が接続しているため、手がかり表現「などの」の直後の形態素「写真」から、手がかり表現の最後の形態素「の」を含む文節の係り先の主辞「サイト」までの形態素列「写真共有サイト」を定義表現として抽出する。

研究資源名の前後に定義の手がかり表現が接続していない場合は、研究資源名の直前に既知の定義名詞が接続しているかどうか判断する。研究資源名の直前に既知の定義名詞があれば、研究資源名を含む文節を根とした部分木に対応する形態素列から研究資源名を削除したものを定義表現として抽出する。このケースは、2.2.2 項で述べた定義表現の出現パターン〔定義表現-研究資源名〕型の特殊例に対応する。

図 3.10 中の木構造は、「本論文では、周辺環境とユーザの心理状態を記録・再現するデジタルフォトメディア exPhoto を提案した。」という文の構文解析結果である。図 3.10 の場合は、研究資源名「exPhoto」の直前に定義名詞「メディア」があ

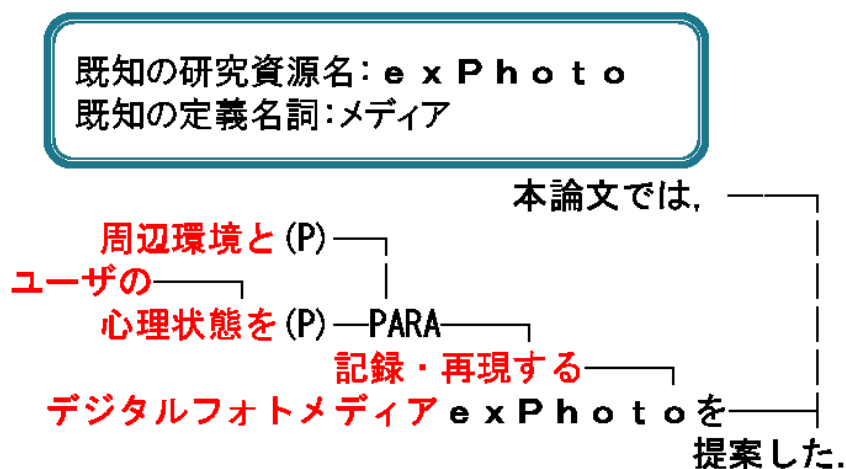


図 3.10: 研究資源名と接続している定義表現の抽出

るので、研究資源名を含む文節を根とした部分木に対応する形態素列から研究資源名を削除した「周辺環境とユーザの心理状態を記録・再現するデジタルフォトメディア」の部分で定義表現として抽出する。

また、定義の手がかり表現を含む文節の係り先に対して、格関係がヲ格となる形態素が研究資源名の一部であった場合は、構文木において、手がかり表現の最初の形態素を含む文節を根とした部分木に対応する形態素列を定義表現として抽出する。ただし、手がかり表現の部分は定義表現に含めない。このケースは、2.2.2 項で述べた定義表現の出現パターン「定義表現-研究資源名」型に対応する。

図 3.11 中の木構造は、「実行不能解を生成しない交叉方式として予備実験の結果に基づいて CX を用いることとした。」という文の構文解析結果である。図 3.11 の場合は、手がかり表現「として」の一部「して」の係り先に対して、研究資源名「CX」がヲ格となるので、手がかり表現の最初の形態素「と」を含む文節を根とした部分木から手がかり表現を削除した「実行不能解を生成しない交叉方式」の部分で定義表現として抽出する。

3.4.2 用途表現の抽出

定義表現の抽出手法と同様に、まず、本文テキストから既知の研究資源名を含む文を選択し、研究資源名の前後に「を用いて」や「するための」のような用途の手がかり表現が接続しているかどうかを判断する。

研究資源名の直前に手がかり表現が接続していたら、構文木において、手がかり表現の最初の形態素を含む文節を根とした部分木に対応する形態素列を用途表現として抽出する。ただし、手がかり表現の部分は用途表現に含めない。このケー

既知の研究資源名: CX
既知の手がかり表現: として

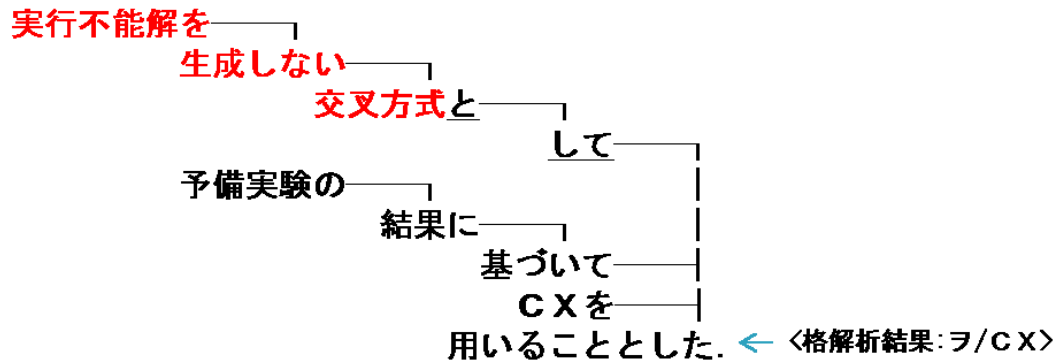


図 3.11: 手がかり表現「として」に基づく定義表現の抽出

スは、2.2.2 項で述べた用途表現の出現パターン〔用途表現-研究資源名〕型に対応する。

図 3.12 中の木構造は、「探索・意思決定時間とポインティング時間を考慮した、階層メニューの項目選択に要する時間の平均値を最小化するための遺伝的アルゴリズムを提案した。」という文の構文解析結果である。図 3.12 の場合は、研究資源名「遺伝的アルゴリズム」の直前に用途の手がかり表現「するための」が接続しているため、「するための」を含む文節を根とした部分木に対応する形態素列「階層メニューの項目選択に要する時間の平均値を最小化するための」から、手がかり表現の部分を削除した「階層メニューの項目選択に要する時間の平均値を最小化」の部分を用途表現として抽出する。

研究資源名の直後に手がかり表現が接続していたら、手がかり表現の最後の形態素を含む文節の係り先を根とした部分木に対応する形態素列を用途表現として抽出する。ただし、手がかり表現を含むノードより前方の形態素は用途表現に含まない。また、部分木の根となる文節の主辞以降の形態素も用途表現には含まない。このケースは、2.2.2 項で述べた用途表現の出現パターン〔研究資源名-用途表現〕型および〔成果物表現-研究資源名-動詞・サ変名詞〕型に対応する。

図 3.13 中の木構造は、「不完備情報ゲームを用いてエージェントの意思決定モデルを構築する。」という文の構文解析結果である。図 3.13 の場合は、研究資源名「不完備情報ゲーム」の直後に用途の手がかり表現「を用いて」が接続しているため、手がかり表現の最後の形態素「用いて」を含む文節の係り先を根とした部分木に対応する形態素列「エージェントの意思決定モデルを構築する」から、手が

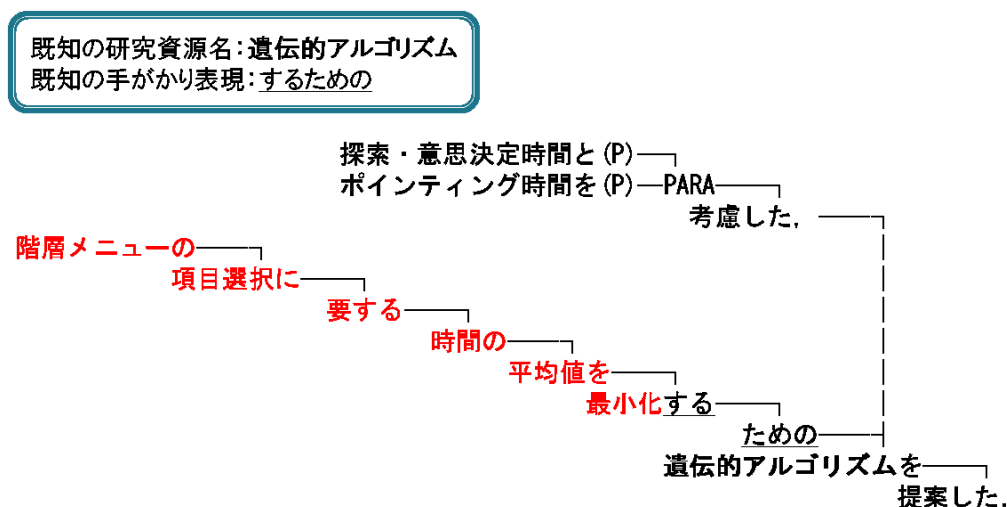


図 3.12: 手がかり表現「するための」に基づく用途表現の抽出

かり表現を含むノードより前方の形態素および部分木の根となる文節の主辞より後方の形態素「する」を取り除いた「エージェントの意思決定モデルを構築」の部分を用途表現として抽出する。

また、用途の手がかり表現を含む文節の係り先に対して、格関係がヲ格となる形態素が研究資源名の一部であった場合は、構文木において、手がかり表現の最初の形態素を含む文節を根とした部分木に対応する形態素列を用途表現として抽出する。ただし、手がかり表現の部分は用途表現に含めない。このケースは、2.2.2項で述べた定義表現の出現パターン〔用途表現-研究資源名〕型に対応する。

図 3.14 中の木構造は、「ユーザは、体験を記録・再現するために exPhoto Camera を用いる。」という文の構文解析結果である。図 3.14 の場合は、手がかり表現「するために」の一部「ために」の係り先に対して、研究資源名「exPhoto Camera」の一部である「Camera」がヲ格となるので、手がかり表現の最初の形態素「する」を含む文節を根とした部分木から手がかり表現を削除した「体験を記録・再現」の部分を用途表現として抽出する。

定義表現の抽出とは異なり、用途表現の抽出では、研究資源名の直前に既知の用途動詞が接続しているかの判断は行わない。

既知の研究資源名: 不完全情報ゲーム
既知の手がかり表現: を用いて

不完備情報ゲームを——
用いて——
エージェントの——
意思決定モデルを——
構築する。

図 3.13: 手がかり表現「を用いて」に基づく用途表現の抽出

既知の研究資源名: exPhoto Camera
既知の手がかり表現: するために

ユーザは、——
体験を——
記録・再現する——
ために——
exPhoto Cameraを——
用いる。 ← 〈格解析結果: ヲ/Camera〉

図 3.14: 手がかり表現「するために」に基づく用途表現の抽出

第4章 評価実験

提案手法の評価を行うために、研究資源名の抽出実験、定義表現および用途表現の抽出実験を行った。

4.1 実験の目的と方法

実験に先立って、学習用データおよび評価用データを作成した。2008年度人工知能学会全国大会の講演論文集に収録されている120論文から人手で抽出した研究資源名・定義表現・用途表現を学習用データとし、同論文集に収録されている30論文から人手で抽出した研究資源名・定義表現・用途表現を評価用データとした。

研究資源名の抽出実験では、前述の学習用データを用いて学習を行い、2008年度人工知能学会全国大会の講演論文集に収録されている30論文に対して研究資源名を抽出した。このとき、2章で分析に用いた50論文から人手で取得した末尾表現および手がかり表現を、既知のものとして利用した。

また、研究資源名候補の取得において定義表現および用途表現を使用せず、SVMによる学習において素性から定義表現および用途表現に関連するものを外して抽出を行った結果をベースラインとした。すなわち、ベースライン手法では、研究資源名候補は未知語と末尾表現に基づいて取得し、SVMの素性は3.3.2項で述べたもののうち9.以降を除いたものを利用した。

さらに、既存の専門用語自動抽出システム termex^{*1}[8] も比較手法として採用した。termexは、接続する語の情報や出現頻度などを利用して、入力として与えたテキストファイルに含まれる名詞・複合名詞の重要度を計算し、重要度の高いものを専門用語として抽出するツールである。研究資源名は専門用語の一種と捉えることができるが、本実験では、既存の専門用語抽出手法が研究資源名の抽出においても有用であるかを評価する。

提案手法、ベースライン手法、termexを用いて研究資源名の抽出を行い、それぞれの適合率・再現率・F値を計算し評価した。また、提案手法とベースライン手法の間の有意差に関してはt検定を行った。t検定には、2章で分析に用いなかった100論文を10分割して、10論文をテスト用データとして用い、残りの90論文

^{*1}<http://gensen.dl.itc.u-tokyo.ac.jp/win.html>

および2章で分析に用いた50論文を学習用データとして用いて、研究資源名の抽出を10通り行った結果を利用した。

ここで、研究資源名抽出の適合率・再現率・F値を以下のように定義した。

$$\text{適合率 (研究資源名)} = \frac{\text{正解数}}{\text{取得数}} \quad (4.1)$$

$$\text{再現率 (研究資源名)} = \frac{\text{正解数}}{\text{研究資源名数}} \quad (4.2)$$

$$F \text{ 値 (研究資源名)} = \frac{2 * \text{適合率 (研究資源名)} * \text{再現率 (研究資源名)}}{\text{適合率 (研究資源名)} + \text{再現率 (研究資源名)}} \quad (4.3)$$

ただし、

- 正解数：正しく研究資源名と判定した研究資源名候補の数
- 取得数：取得した研究資源名候補の数
- 研究資源名数：取得した研究資源名候補の集合に含まれる研究資源名の数

である。

研究資源名の抽出実験結果とは独立して、定義表現および用途表現の抽出実験を行った。研究資源名の抽出実験に用いた150論文から人手で抽出した研究資源名を既知の研究資源名として利用し、同じ150論文から提案手法によって定義表現および用途表現を抽出し、適合率・再現率・F値を計算した。

ここで、定義表現抽出の適合率・再現率・F値を以下のように定義した。

$$\text{適合率 (定義表現)} = \frac{\text{正しく抽出できた定義表現の数}}{\text{定義表現として抽出した文字列の数}} \quad (4.4)$$

$$\text{再現率 (定義表現)} = \frac{\text{正しく抽出できた定義表現の数}}{150 \text{ 論文中的定義表現の数}} \quad (4.5)$$

$$F \text{ 値 (定義表現)} = \frac{2 * \text{適合率 (定義表現)} * \text{再現率 (定義表現)}}{\text{適合率 (定義表現)} + \text{再現率 (定義表現)}} \quad (4.6)$$

同様に、用途表現抽出の適合率・再現率・F値を以下のように定義した。

$$\text{適合率 (用途表現)} = \frac{\text{正しく抽出できた用途表現の数}}{\text{用途表現として抽出した文字列の数}} \quad (4.7)$$

$$\text{再現率 (用途表現)} = \frac{\text{正しく抽出できた用途表現の数}}{150 \text{ 論文中的用途表現の数}} \quad (4.8)$$

$$F \text{ 値 (用途表現)} = \frac{2 * \text{適合率 (用途表現)} * \text{再現率 (用途表現)}}{\text{適合率 (用途表現)} + \text{再現率 (用途表現)}} \quad (4.9)$$

定義表現および用途表現の抽出では、抽出した文字列の中に定義表現あるいは用途表現が含まれていれば、正しく抽出できたとみなした。たとえば、研究資源名「jPop-E」に対して「フレージング表現に重点を置いたルールベース型の演奏表情付けシステム」という定義表現を手手で抽出していた場合、提案手法によって定義表現として「前節で述べたフレージング表現に重点を置いたルールベース型の演奏表情付けシステム」という文字列を抽出したら、「前節で述べた」の部分は定義表現ではないが、正しく定義表現を抽出できたとみなす。また、研究資源名「底面 RFID センサ」に対して、成果物表現「人の位置と ID」とサ変名詞「取得」の組を用途表現として手手で抽出していた場合、提案手法によって用途表現として「人の位置と ID は」「取得」という文字列の組を抽出したら、「は」の部分は用途表現ではないが、正しく用途表現を抽出できたとみなす。

定義表現および用途表現の抽出におけるベースラインとして、研究資源名を含む文をすべて取得するという手法を採用した。これは、定義表現および用途表現は研究資源名を含む文に出現しやすいという特徴的な傾向を単純に利用したものである。本研究では、研究資源名を含む文中に現れる定義表現および用途表現を取得対象としているため、ベースライン手法で取得した場合の再現率は 100 %になると考えられる。

4.2 実験結果

4.2.1 研究資源名の抽出実験

4.1 節で述べたように、提案手法およびベースライン手法によって研究資源名の自動抽出実験を行った。表 4.1 に研究資源名の抽出結果を示す。

表 4.1: 研究資源名の抽出結果

	適合率 (%)	再現率 (%)	F 値 (%)
提案手法	56.3(54/96)	33.3(54/162)	41.9
ベースライン	48.1(38/79)	24.1(38/158)	32.1

提案手法での研究資源名の抽出結果は、適合率が 56.3%、再現率が 33.3%であり、F 値は 41.9%であった。一方、ベースラインでの抽出結果は、適合率が 48.1%、再現率が 24.1%であり、F 値は 32.1%であった。

正しく抽出できた研究資源名の数を比較すると、提案手法では 54 個、ベースラインでは 38 個と提案手法の方が 1.4 倍ほど多い結果となった。適合率、再現率ともにベースラインよりも提案手法の方が高かったため、提案手法の方がより有効であると考えられる。

また、研究資源名「C4.5」には未知語も既知の末尾表現も含まれていない（「C4.5」の形態素解析結果は、「品詞：特殊」である記号「C」と「品詞：名詞」である数詞「4.5」である）ために、ベースライン手法では研究資源名候補として取得できなかったが、定義表現および用途表現に基づいた研究資源名候補の抽出を行っている提案手法では研究資源として正しく抽出することができた。

提案手法によって誤って研究資源名として取得してしまった語としては、「ファシリテータ」や「非言語インタラクション」のような専門用語、「id」などの文字式、「従来法」や「最適化アルゴリズム」のような末尾表現を含む一般的な表現などがあった。

また、提案手法およびベースライン手法のF値の平均についてt検定を行ったところ、 $t = 3.25 (p < 0.05)$ であった。したがって、提案手法とベースライン手法の間には有意な差が認められた。t検定の結果を表4.2に示す。

表 4.2: t 検定の結果

	F 値の平均	標準偏差
提案手法	0.368	0.090
ベースライン	0.268	0.085

図4.1に、専門用語自動抽出システム termex による研究資源名の抽出結果を示す。termex は、取得した研究資源名候補（専門用語候補）を重要度が高い順に出力するため、上位のものから n 個取り出した中に m 個の研究資源名が含まれていた場合、 n を取得数、 m を正解数として、 n の値を変化させてそれぞれのF値を計算した。

抽出実験では合計4386個の研究資源名候補を抽出し、うち138個が研究資源名であった。研究資源名候補を上位からすべて取得した場合、すなわち取得数 $n = 4386$ のときにF値は最大(6.1%)となったが、提案手法よりも大幅に低い結果となった。このため、既存の専門用語抽出手法は、より限定的な概念である研究資源名の抽出には有効ではないと考えられる。

4.2.2 定義表現および用途表現の抽出実験

提案手法およびベースライン手法によって定義表現および用途表現の自動抽出実験を行った。表4.3に定義表現の抽出結果、表4.4に用途表現の抽出結果を示す。

提案手法での定義表現の抽出結果は、適合率が52.8%、再現率が37.4%であり、F値は43.8%であった。一方、ベースラインでの抽出結果は、適合率が12.5%、再現率が100%であり、F値は22.3%であった。

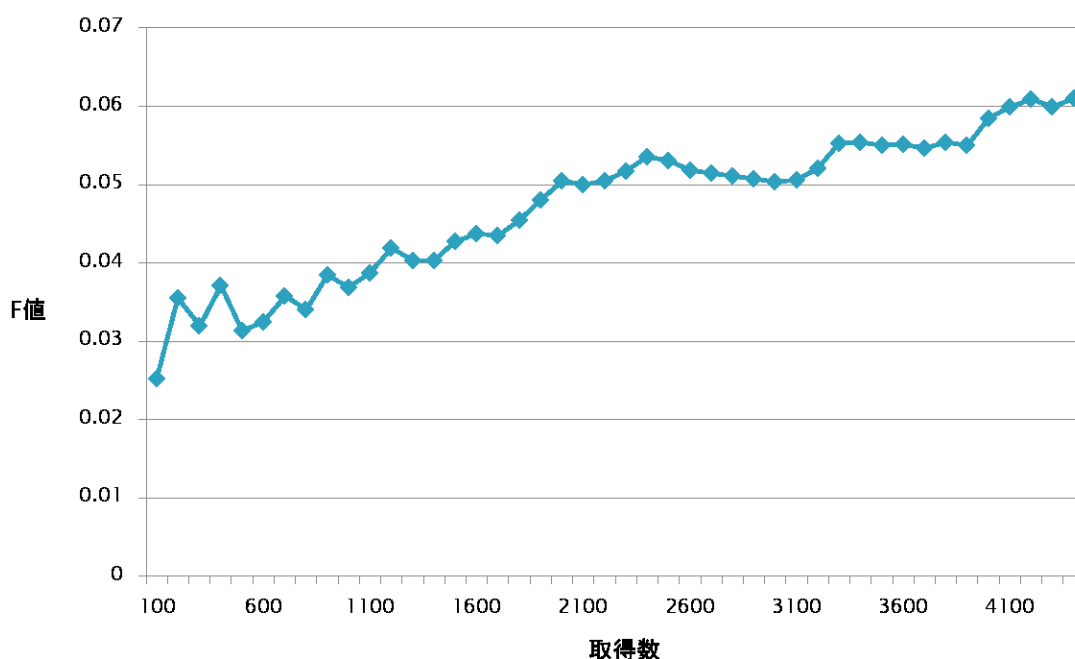


図 4.1: termex による研究資源名の抽出結果

表 4.3: 定義表現の抽出結果

	適合率 (%)	再現率 (%)	F 値 (%)
提案手法	52.8(134/254)	37.4(134/358)	43.8
ベースライン	12.5(358/2856)	100(358/358)	22.3

提案手法での用途表現の抽出結果は、適合率が 43.8%、再現率が 54.6%であり、F 値は 48.6%であった。一方、ベースラインでの抽出結果は、適合率が 11.4%、再現率が 100%であり、F 値は 20.5%であった。

4.1 節で述べたように、本研究では、研究資源名を含む文中に現れる定義表現および用途表現を取得対象としているため、研究資源名を含む文をすべて取得するというベースライン手法での再現率は 100 %となった。しかし、定義表現も用途表現も含まない文も多く取得しているため、提案手法に比べてベースラインの適合率は大幅に低くなった。

また、ベースラインでは文全体を取得しているため、取得した表現には無駄な部分が含まれていることが多い。例えば、研究資源「CHILD」に対して、提案手法では「Schank による概念依存構造を意味表現として用いた言語獲得システム」という定義表現を取得しているが、ベースライン手法では、「視覚情報や音声言語情

表 4.4: 用途表現の抽出結果

	適合率 (%)	再現率 (%)	F 値 (%)
提案手法	43.8(178/406)	54.6(178/326)	48.6
ベースライン	11.4(326/2856)	100(326/326)	20.5

報の認識が終わっていることを前提として、語彙と文法の学習をモデル化している研究としては、古くは、言語表現とその意味を表す意味ネットワークを入力として ATN の文法を学習するシステム LAS, Schank による概念依存構造を意味表現として用いた言語獲得システム CHILD, 入力文と意味役割記述から決定的パーザの構文解析規則を逐次的に学習するモデル, などがよく知られている。」というような冗長な情報を取得している。また、この例では、研究資源「CHILD」のみならず、別の研究資源「LAS」についての定義表現まで含まれてしまっている。

2章で述べたように、本研究では、研究資源の広範な利用を促すために定義表現や用途表現などの研究資源情報を機械的に取得することを目的としており、研究資源情報の利用例として、キーワードを検索クエリとして、データベース化された研究資源情報と文字列照合することによって、目的に合った研究資源を検索する、というものを挙げている。ベースライン手法のように研究資源名を含む文全体を取得するのでは、研究資源情報ではない無駄な部分も多く取得されてしまう、また、同じ文に含まれる別の研究資源に関する研究資源情報まで取得されてしまう、などの問題が生じるため、取得する文字列がより適切に絞られている提案手法の方がより有効であると考えられる。

2章で分析に用いていない100論文から無作為に50論文を選び、提案手法によってその50論文から抽出できなかった定義表現および用途表現について分析を行った。50論文中に87個の定義表現があったが、そのうち42個は取得に失敗していた。また、50論文中に99個の用途表現があったが、そのうち31個は取得に失敗していた。提案手法による定義表現および用途表現の取得失敗例としては、主に3種類のケースがあった。1つ目は研究資源名と定義表現または用途表現の間に手がかり表現が存在しなかった場合、2つ目は手がかり表現となりうる表現が研究資源名と定義表現また用途表現の間に存在するものの、既知の手がかり表現ではなかった場合、3つ目は構文解析に誤りがあった場合である。

1つ目の場合は、定義表現または用途表現は、2章で述べた定義表現または用途表現の出現パターンのうち「その他」に当てはまるような出現をしている。たとえば、「治験薬構造データベース中に収載されている114,501件の薬物構造データを対象に、そのNTGおよび関連する薬理活性情報を抽出し、相互の関係辞書ファイルを作成した。」という文では、研究資源「治験薬構造データベース」は「収載されている114,501件の薬物構造データを対象に、そのNTGおよび関連する薬理

活性情報を抽出し、相互の関係辞書ファイルを作成」することだと考えられるが、提案手法では正しく用途表現を抽出することができない。50 論文中、12 個の定義表現と、13 個の用途表現がこのケースに該当し、正しく抽出されなかった。

2 つ目の場合は、2 章で分析に用いた 50 論文から人手で取得した定義の手がかり表現および用途の手がかり表現のみでは網羅しきれなかったため、取得し損ねてしまったということである。50 論文中、9 個の定義表現と、8 個の用途表現がこのケースに該当し、正しく抽出されなかった。人手によって手がかり表現を大量に取得することは非常に困難であるため、研究資源名・定義表現・用途表現のみならず、手がかり表現を自動抽出する手法を実現することも今後の課題の一つとして挙げられる。

3 つ目の場合は、さらに 2 つのケースに分けられる。1 つは、3.2 節で述べた独自プログラムによって正しく文が取得できず、正しく構文解析できなかった場合、もう 1 つは正しく文は取得できたのにも関わらず構文解析が誤っていた場合である。50 論文中、6 個の定義表現と、1 個の用途表現が、本文の抽出ミスによって正しく抽出されなかった。また、15 個の定義表現と、9 個の用途表現が、構文解析の失敗によって正しく抽出されなかった。3.2 節でも述べたが、学術論文の PDF から独自プログラムを用いて本文テキストを抽出する適合率は 93.1%、再現率は 90.9%であった。また、構文解析器 KNP の精度は 80%~90%である。

第5章 関連研究

本論文では、研究資源を検索・選択する際に有用となる情報を機械的に取得することを目的として、学術論文からの研究資源名・定義表現・用途表現の自動抽出手法を提案した。関連研究として、5.1 節では固有表現抽出に関する研究について、5.2 節では知識獲得に関する研究について紹介する。

5.1 固有表現抽出

人名や地名などの固有名詞や日付、時間表現などの固有表現を自動抽出することは、情報抽出において重要なタスクの一つである。例えば、MUC(Message Understanding Conference) プロジェクト^{*1}では、人名、組織名、地名、日付表現、時間表現、金額表現、割合表現の7種類の固有表現を扱っていた。また、IREX(Information Retrieval and Extraction Exercise) プロジェクト^{*2}では、MUC の分類に加えて固有物名を含めた8種類の固有表現を定義した。また、関根らは、MUC や IREX、ACE (Automatic Content Extraction) プロジェクト^{*3}にて定義された固有表現のカテゴリを基に、カテゴリ数を200種類にまで拡張した拡張固有表現を提案している^{*4}[2]。

本研究の抽出対象である研究資源名も、固有表現の一種だと考えられる。しかし、汎用的な手法や評価法といった無形のものも広義の研究資源として捉えているため、関根らの拡張固有表現では、研究資源そのものを示すカテゴリは存在せず、該当するカテゴリは複数に及ぶことになる。

笹野らは、文の構造的な解析などから得られる大域的な情報から日本語固有表現を抽出する手法を提案している [15]。笹野らの手法では、日本語文の構文解析、格解析、共参照解析を行い、解析結果から得られるキャッシュ素性（解析対象である文に先行する文の解析結果から得られる素性）・共参照関係・係り先素性・格フレーム素性などの大域的情報を利用してSVMによって固有表現を抽出する。

また、塩入らは、日本語新聞記事からの拡張固有表現の抽出手法を提案している [4]。塩入らの手法では、まず、既存の固有表現辞書を用いて日本語新聞記事から固

^{*1}http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc.7-toc.html

^{*2}<http://nlp.cs.nyu.edu/irex/index-j.html>

^{*3}<http://www.itl.nist.gov/iaui/894.01/tests/ace/>

^{*4}<https://sites.google.com/site/extendednamedentityhierarchy/>

有表現の前後文脈を大量に取得する。取得した文脈から各カテゴリにとって特徴的と考えられるものをカテゴリごとの共起頻度を用いて計算し、得られた特徴的な文脈を利用して新たな固有表現を抽出する。例えば、「アカデミー賞」や「ノーベル賞」などの、賞の名称を示す AWARD カテゴリの固有表現から、「で#AWARD#を受賞」「回#AWARD#を受賞」などの特徴的な文脈を獲得し、獲得した文脈を利用して「ノーベル平和賞」「毎日芸術賞」などの新たな固有表現を抽出する。

本研究では、形態素解析および構文解析の結果を利用して研究資源名の抽出を行う。単純に研究資源名の前後の文脈を利用するのではなく、研究資源名と研究資源名を含む文中に出現する定義表現および用途表現との間にある「を用いて」「である」などの手がかり表現を利用する。研究資源名と接続関係にある手がかり表現のみならず、係り受け関係にあるものも研究資源名の抽出に利用しているため、より深く文章の構造を捉えた抽出手法となっている。また、研究資源名を集積した辞書が存在しないため、学術論文から人手によって抽出した研究資源名を用いて SVM の学習を行っている。

中川らは、専門分野のコーパスから専門用語を自動抽出する手法を提案している [14]。中川らの手法では、単名詞と複合名詞を抽出対象として、単名詞のバイグラムから得られる単名詞の出現頻度と、複合名詞を形成するために接続する名詞の頻度を用いて、単名詞および複合名詞のスコア付けを行い、専門用語を抽出する。

研究資源名も専門用語の一つであると考えられるが、研究活動において何らかの用途に使われるものの名前という点で、専門用語に比べて研究資源名はより限定的な概念であり、文内での出現の仕方にも専門用語という括りでは発見しにくい特徴がある。例えば、本研究の抽出手法では、研究資源名を構成する語の形態的特徴や統計量のみならず、研究資源名と文内で共起する定義表現および用途表現の特徴も利用している。

萩原らは、分かち書きされていない文章から、形態素解析および構文解析を用いずに、国名などの意味的に関連のある語彙カテゴリを抽出するアルゴリズム g-Monaka を提案している [10]。萩原らの手法では、ある与えられた文に出現するすべての文字 n グラムの隣接関係を有向グラフによって表現し、文字 n グラムを文脈パターンとインスタンスの両方として用いる。

本研究では、研究資源名の形態的特徴に着目し、研究資源名の抽出に形態素解析を用いる。例えば、研究資源名には未知語が含まれていることが多いという特徴的な傾向があり、研究資源名の抽出において非常に有用である。また、前述のように定義表現および用途表現の持つ特徴も研究資源名の抽出に有用であると考えているため、構文解析の結果も利用して研究資源名の抽出を行う。

5.2 知識獲得

鳥澤は、ある名詞あるいは名詞句に対する準備を表す表現（準備表現）と用途を表す表現（用途表現）を自動抽出する手法を提案している [5]。例えば、「本」という名詞に対しては、「書く」が準備表現、「読む」が用途表現となる。鳥澤の手法では、語の共起頻度を利用して用途表現を抽出し、抽出した用途表現を手がかりにして準備表現の出現パターンから準備表現の候補を選択し、SVMによって最終的な準備表現と用途表現の対を抽出する。

本研究では、一般的な名詞あるいは名詞句に対する用途表現ではなく、研究資源名という、より専門的な分野のものの名称に対する用途表現を抽出する。また、「書く」や「読む」などのような単純な動詞のみの表現ではなく、「単語への分割と品詞の推定」のような具体的な用途例を示すような用途表現を抽出対象とする。

乾らは、ある出来事と、その出来事の原因・前提条件・効果・手段となる出来事との関係を表す知識（因果関係知識）を、電子化された文書の集合から自動抽出する手法を提案している [6]。乾らの手法では、接続表現「ため」を含む文には高い必然性を持つ因果関係が含まれていることと、「ため」を含む文内に出現する2つの出来事の意志性（行為を表しているか、事態を表しているか）を推定することは因果関係を分類する際に有効であることに着目している。まず、接続表現「ため」を含む複文形式の文から、形態素解析と単純なルールを用いて出来事の命題要素を抽出する。抽出した2つの命題要素の意志性をSVMを用いて推定し、推定値によって2つの出来事の因果関係を原因・前提条件・効果・手段の4種類に分類する。

本研究では、一般的な出来事に関する手段（用途）ではなく、研究活動という、より専門的な分野における研究資源の定義・用途を表す表現を抽出する。また、抽出の際には「ため」だけではなく、「を用いて」や「である」のような様々な手がかり表現を利用する。

酒井らは、新聞記事のテキストデータに含まれる交通事故を扱った記事から、事故原因を表す表現（事故原因表現）を自動抽出する手法を提案している [7]。酒井らの手法では、まず前処理として、SVMを用いて新聞記事コーパスから交通事故を扱った記事を抽出する。事故原因表現と接続および係り受け関係にある「のが原因」「とみて」などの表現を「種表現」と定義し、SVMで抽出した記事から種表現を手がかりにして「前方不注意」「スピードの出し過ぎ」などの事故原因表現を抽出する。取得したいくつかの事故原因を表す表現を手がかりにして、新たな種表現を自動的に抽出する。このプロセスを繰り返すことによって、より多くの事故原因表現を取得する。

また、酒井らは、特許文書から、利用者にとっての直接的な利益を示す表現（効果表現）と、それを実現するために行った技術的な課題解決方法を示す表現（課題表現）を自動抽出する手法を提案している [11]。特許文書において効果表現と課

題表現が一文中に両方とも出現する傾向が強いことに着目し、異なる2種類の手がかり表現と、効果表現および課題表現中に現れる特徴的な動詞や名詞を利用して、ブートストラップ的に効果表現および課題表現を抽出する。

本研究では、学術論文において研究資源名と定義表現および用途表現が一文中に出現する傾向が強いことに着目している。研究資源名、定義表現および用途表現と、接続関係あるいは係り受け関係にある「を用いて」「である」などの表現を手がかりにして抽出を行う。

小澤らは、言語資源の効率的利用を促すために、言語資源の用途表現を、その言語資源の利用者によって記されたテキストから自動抽出する手法を提案している [9]。小澤らの手法では、構文構造を考慮した抽出ルールを人手で作成し、言語資源の名称を既知の情報として利用することで学術論文からその言語資源の用途表現を抽出する。

本研究では、言語資源以外の研究資源の情報も抽出対象とし、研究資源名は既知でないことを前提とする。すなわち、定義表現および用途表現のみならず、研究資源名も学術論文から機械的に抽出している。

第6章 おわりに

6.1 まとめ

本論文では、研究資源を検索・選択する際に有用となる情報である、研究資源名・定義表現・用途表現を網羅的に取得し提供することを目的とし、学術論文からの研究資源情報の自動抽出手法を提案した。

研究資源名の抽出には、学術論文の分析によって得られた、(1) 研究資源名の形態素的特徴、(2) 定義表現および用途表現の持つ特徴、(3) 手がかり表現を利用した。具体的には、まず、学術論文の本文テキストから研究資源名の候補となる形態素列を取得した。研究資源名候補の抽出には、未知語に基づく抽出、既知の末尾表現に基づく抽出、既知の定義表現および用途表現に基づく抽出の3種類の方法があり、それぞれの方法によって得られた研究資源名候補の集合に対して、SVMを用いて研究資源名を選別し、それぞれの集合ごとのSVMによる選別結果の和集合（重複を除く）を、研究資源名として取得した。SVMの素性にも、研究資源名の形態素的特徴に基づく素性だけではなく、既知の定義表現および用途表現の持つ特徴や手がかり表現に基づく素性も用いた。

定義表現および用途表現の抽出には、手がかり表現および研究資源名を利用した。具体的には、まず、本文テキストから研究資源名を含む文を選択した。研究資源名の前後に接続している手がかり表現を参考にして、接続関係・係り受け関係・格解析結果を利用して定義表現や用途表現を抽出した。文中に手がかり表現が存在せず、定義表現と研究資源名と接続している場合があるため、既知の定義表現の末尾の名詞（定義名詞）を利用した抽出も行った。

また、提案手法を評価するために、研究資源名の抽出実験と、定義表現および用途表現の抽出実験を個別に実施した。実験には、2008年度人工知能学会全国大会の講演論文集に収録されている150論文を使用した。研究資源名の抽出実験の結果から、研究資源名の抽出において、既知の定義表現および用途表現の持つ特徴を活用することの有効性が確認できた。また、定義表現および用途表現の抽出実験の結果から、定義表現および用途表現の抽出において、手がかり表現を活用することの有効性が確認できた。

6.2 今後の課題

今後の課題として以下が挙げられる。

研究資源名の自動抽出手法の改善

提案手法では、研究資源名の形態素的特徴や、定義表現や用途表現の持つ特徴を利用して研究資源名の候補となる形態素列を抽出し、SVMを用いて研究資源名を選別している。しかし、適合率に比べて再現率が低く、研究資源名であるのにも関わらず取得できていないものが存在する。取得できていない研究資源名の多くは、その出現回数、あるいは定義表現や用途表現と文内で共起する回数が少なかった。

このような研究資源名にも対応できるよう、SVMの素性の追加やさらなる素性の洗練を行うことが解決策として考えられる。

定義表現および用途表現の自動抽出手法の改善

提案手法では、構文解析の誤りなどが原因で、定義表現や用途表現ではない無関係の形態素列まで抽出してしまうことがあった。抽出した形態素列をすべて定義表現あるいは用途表現として取得するのではなく、その形態素列の長さやその他の形態素的特徴から、どの程度定義表現らしいか、あるいはどの程度用途表現らしいかを判定して選別することにより、さらに抽出精度を向上できると考えられる。

また、本研究では、「である」や「を用いて」などの、研究資源名・定義表現・用途表現と接続関係あるいは係り受け関係にある手がかり表現を利用して、定義表現および用途表現の抽出を試みた。しかし、研究資源情報を含む文の中には、このような手がかり表現を含まない文も存在する。

さらに、提案手法は、研究資源名を含む文から定義表現や用途表現を抽出するというものであり、研究資源名を含まない文中の定義表現や用途表現については、抽出の対象外としていた。しかし、複数の文に渡って研究資源の用途が説明されることや、研究資源名の代わりに「このツール」などと言い換えられることなどによって、研究資源名を含まない文にも定義表現や用途表現が含まれている可能性がある。

これらの場合に対する解決策として、文章の段落情報や照応解析結果など、文単位の解析では得られないような特徴も抽出に利用し、手がかり表現や研究資源名を含まない文からも定義表現や用途表現が抽出できるような手法を考案することが考えられる。

手がかり表現の自動抽出

本研究では、2008年度人工知能学会全国大会の講演論文集に収録されている50論文から人手によって抽出した手がかり表現を用いて、研究資源名・定義表現・用途表現の抽出を行った。しかし、人手で抽出できる手がかり表現の数には限度がある。4章の実験結果でも、既知の手がかり表現が文中に存在しなかったために定義表現や用途表現が正しく抽出できないケースが見られた。

手がかり表現が研究資源名・定義表現・用途表現と接続関係あるいは係り受け関係にあることを利用し、既知の研究資源名と既知の定義表現あるいは用途表現を用いて、学術論文から新たな手がかり表現を自動抽出するという手法が、手がかり表現が不十分である問題の解決策として考えられる。

その他の研究資源情報の抽出

本研究では、提案手法の適用対象となる研究資源情報を、研究資源名、定義表現、用途表現の3種類に限定した。これらの研究資源情報の他にも、研究資源の機能や特徴、研究資源を使用することで得られる成果物の内容、研究資源を適用する対象、研究資源の規模など、様々な研究資源情報が学術論文中には存在する。しかし、研究資源名と同じ文に出現することが少ない、出現回数自体が少ない、などの理由で、提案手法の適用の対象外となっていた。

より多くの学術論文から、研究資源情報が出現している文の特徴を分析し、提案手法の適用対象外となっている研究資源情報を抽出する手法を考案する必要があると考えられる。

同じ研究資源を指す複数の研究資源名の対応付け

研究資源の中には、「ChaSen」と「茶筌」や、「自己組織化マップ」と「SOM」のように、複数の表記・名称を持つものが存在する。しかし、提案手法では各名称を別個の研究資源として扱っており、同じ研究資源であるにも関わらず、「ChaSen」の定義表現や用途表現を「茶筌」の定義表現や用途表現として見なすことができない。同じ研究資源を指す複数の研究資源名の対応付けがなされていないことは、研究資源情報を研究資源の検索・比較に用いる際に問題になる。

「自己組織化マップ (SOM)」などのような、論文中での括弧表記による言い換えや、研究資源名を含む文で参照されている論文の情報を利用して、研究資源名の対応付けを行うという手法がその解決策として考えられる。

研究資源の検索・選択における研究資源情報の有用性の検証

本研究の目的は、研究資源を検索・選択する際に有用となる情報である、研究資源名・定義表現・用途表現を網羅的に取得し提供することであった。学術論文で言及される研究資源には、利用者の研究活動によって新たに創造された定義や用途も多く存在する。そのため、研究資源名とともに定義表現および用途表現を提供することは、研究資源の広範な利用に大きく貢献すると考えられる。今後の課題の一つとして、提案手法によって取得した定義表現および用途表現が、研究資源の検索・選択において確かに有用であるか検証することが挙げられる。

その検証方法としては、自動抽出した研究資源名・定義表現・用途表現をデータベース化し、そのデータベースを用いた研究資源検索システムを構築し、そのシステムの検索の性能を評価するというものが考えられる。

謝辞

本研究を進めるにあたって、指導教員である長尾確教授には、研究に関する貴重な御意見や助言をいただくとともに、研究に対する姿勢や心構え、論文執筆におけるテクニックなど、様々な面で多くの御指導をいただき、大変御世話になりました。心より御礼申し上げます。

松原茂樹准教授には、プロジェクトのミーティングやゼミなどにおいて、手法の提案から論文執筆に至るまで、多くの御意見や助言をいただき、大変御世話になりました。また、研究活動における様々な面で支援していただき、感謝しております。心より御礼申し上げます。

大平茂輝助教には、ゼミなどで、本研究の本質的な部分や細部の仕様について、多くの御意見をいただき、大変御世話になりました。心より御礼申し上げます。

石戸谷顕太郎さんには、ゼミなどで、本研究に関して多くの有益な御意見や助言をいただきました。また、研究活動において多くの御指導をいただき、大変御世話になりました。ありがとうございました。

竹島亮さんには、ゼミなどで多くの有益な御意見をいただいたことに加え、実装に関して多くの的確な助言をいただき、大変御世話になりました。ありがとうございました。

渡邊賢さんには、ゼミなどで多くの有益な御意見をいただいたことに加え、研究室での様々な活動の中で大変御世話になりました。ありがとうございました。

同じプロジェクトのメンバーである馮思萌さんには、実装や論文執筆などについて数多くの助言をいただいたことに加え、学術論文のコーパス作成に協力していただき、大変御世話になりました。ありがとうございました。

棚瀬達央さん、尾崎宏樹さん、川西康介さん、矢田幸大さんには、スライド作成技術などの研究活動に関する基礎的なことから、プログラミングに関する専門的な内容まで、多くの御指導をいただきました。また、ゼミなどで多くの有益な御意見をいただき、大変御世話になりました。ありがとうございました。

久保田芙衣さん、高雨蘇さん、小林尚弥さん、西脇雅幸さんには、ゼミなどで多くの有益な御意見をいただきました。また、研究室での様々な活動の中で大変御世話になりました。ありがとうございました。

長尾研究室秘書の鈴木美苗さん、土井ひとみさんには、研究室での活動全般において、大変御世話になりました。ありがとうございました。

最後に、日々の生活を支えてくれた家族に心より感謝します。

参考文献

- [1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, Vol. 20, No. 3, pp. 273–297, Sep. 1995.
- [2] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC2002)*, pp. 1818–1824, May. 2002.
- [3] Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, and Hitoshi Isahara. Construction of an infrastructure for providing users with suitable language resources. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING-2008)*, pp. 119–122, Aug. 2008.
- [4] 塩入寛之, 関根聡, 梅村恭司. 拡張固有表現獲得の精度向上. 情報処理学会研究報告. 自然言語処理研究会, Vol. 2007, No. 76, pp. 67–72, Jul. 2007.
- [5] 鳥澤健太郎. 対象の用途と準備を表す表現の自動獲得. 自然言語処理, Vol. 13, No. 2, pp. 125–144, Apr. 2006.
- [6] 乾孝司, 乾健太郎, 松本裕治. 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得. 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933, Mar. 2004.
- [7] 酒井浩之, 梅村祥之, 増山繁. 交通事故事例に含まれる事故原因表現の新聞記事からの抽出. 自然言語処理, Vol. 13, No. 2, pp. 99–123, Apr. 2006.
- [8] 小島浩之, 前田朗. キーワード（専門用語）自動抽出システムの構想とその展開. 第51回日本図書館情報学会研究発表要綱, pp. 17–20, Oct. 2003.
- [9] 小澤俊介, 遠山仁美, 内元清貴, 松原茂樹. 言語資源の用途情報の獲得と利用. 電子情報通信学会論文誌, Vol. 95, No. 7, pp. 611–622, Jul. 2012.
- [10] 萩原正人, 小川泰弘, 外山勝彦. グラフカーネルを用いた非分かち書き文からの漸次的語彙知識獲得. 人工知能学会論文誌, Vol. 26, No. 3, pp. 440–450, 2011.

- [11] 坂地泰紀, 野中尋史, 酒井浩之, 増山繁. Cross-bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法. 電子情報通信学会論文誌, Vol. 93, No. 6, pp. 742–755, Jun. 2010.
- [12] 黒橋禎夫. 結構やるな, KNP. 情報処理, Vol. 41, No. 11, pp. 1215–1220, Nov. 2000.
- [13] 妙木裕, 松本裕治, 長尾真. 汎用日本語辞書および形態素解析システム. 情報処理学会第42回全国大会講演論文集, 第42巻, pp. 17–18, Feb. 1991.
- [14] 中川裕志, 湯本紘彰, 森辰則. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, Jan. 2003.
- [15] 笹野遼平, 黒橋禎夫. 大域的情報を用いた日本語固有表現認識. 情報処理学会論文誌, Vol. 49, No. 11, pp. 3765–3776, Nov. 2008.