博 士 論 文

論文間の参照情報の抽出と利用に関する研究

指導教官　奥村 学 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

難波 英嗣

2001 年 3 月

要旨

学術論文中には, 当該論文と被参照論文との関係について記述されている個所 (参照個所) がある. 参照個所から得られる情報を, 本研究では参照情報と呼んでいる. 本論文では, 論文間の参照情報を自動的に抽出する手法を提案し, その応用例を示す.

引用分析において, これまで論文間の参照・被参照関係は様々な目的に利用されてきた. 例えば, 論文間の関連度の評価, 研究領域の分析, 論文誌や研究者の評価などが挙げられる. しかし, これらの多くはすべての参照を同等に扱っており, 単純に参照の数を数えるだけでこのような分析を行うことに, 引用分析の研究が始まった当初から批判があった.

そこで, 本研究では論文間の参照情報に着目する. 参照個所からは, 被参照論文の重要点や当該論文と被参照論文との相違点を明示する有用な情報が得られる. また, 参照個所を読めば参照の理由 (参照タイプ) が分かる. こうした参照情報から, 当該論文の関連論文の中での位置づけが明らかになるため, 特定分野の研究動向の概要の把握に有用である. さらに, 論文間の関連度の評価を含む多くの引用分析手法にも利用できると考えられる. そこで, 本研究では, まず論文間の参照情報の抽出を試みた.

参照情報の抽出は, 参照個所の抽出と参照タイプの決定という 2 つのステップに分けられる. 本研究では参照個所の抽出は, 参照のある文と文間のつながりが強いと考えられる文を, 参照の前後の文から抽出する処理であると考え, 手がかり語を用いて参照個所の自動抽出を行った. その結果, 再現率 80%, 精度 76%の抽出精度が得られた. また, 抽出された参照個所を手がかり語を用いて解析し, 参照タイプを明らかにした. 実験の結果, 参照タイプ決定では 83%の解析精度が得られた.

参照情報の応用例の一つとして, サーベイ論文の作成支援を取り上げた. サーベイ論文には 2 つの処理 (1) 特定分野の論文の収集 (2) 論文間の類似点, 相違点の抽出が必要であると考えられる. 本論文では参照情報を用いることで, これらの処理を部分的に実現した. 特定分野の論文の収集は, 参照タイプを考慮して特定の参照・被参照関係を辿ることで可能になった. また, 論文間の類似点・相違点は参照個所中に記述されている. そこで，ユーザに論文間の参照関係を表すグラフ，グラフ中の個々の論文の概要，参照個所を提示するシステムを構築した．このシステムを利用することで特定分野の論文が自動収集され，また収集された論文集合の論文間の相違点が明らかにされるため，参照情報がサーベイ論文作成支援に有用であることが示された．

また, 参照情報の他の応用例として, 関連論文の自動分類を取り上げた. 書誌結合は引用分析の代表的な手法であるが, 書誌結合ではすべての参照を等価に扱っているため, 十分な分類精度が得られていない. そこで, 本研究では参照情報を利用した関連論文の分類手法を提案し, 実験により, 提案手法と書誌結合を含む既存の手法を比較した. その結果, 分類精度において提案手法が最も優れており, また計算速度においても, 他の手法と比較して提案手法が十分高速であることが分かった.

本論文では, 参照情報の応用例としてサーベイ論文の作成支援と関連論文の自動分類を取り上げたが, 参照情報は引用分析研究で行われてきた研究領域の分析, 論文誌や研究者の評価などにも応用できる. また, 学術論文だけでなく, 特許やウェブ文書といった他のジャンルのテキストへの適用なども考えられる.

キーワード: 引用分析, 参照情報の抽出, 複数論文の要約, 文書分類

Extraction of Citation Information and Its Applications

Hidetsugu Nanba

School of Information Science,

Japan Advanced Institute of Science and Technology

Abstract

In a research paper, there are passages where the author of a current paper describes the essence of cited papers and the differences between the current paper and the cited papers (we call these passages "citing areas"). We call the information derived from these passages "citation information". In this thesis, we propose a method for extracting citation information and show its applications.

In the field of citation analysis, citation relationships have been used for classifying papers, evaluating the importance of papers or journals, and analysing the relationships between research fields. However, most analyses treat all citations equally, although there are actually several reasons for citations.

In this work, we make use of citation information. With the information from citing areas, we can know the similarities and differences between the current paper and the cited papers. We can also identify the types of citation relationships that indicate the reasons for citations (we call them citation types). Citation information makes it possible to understand a stance of a paper among several related papers, or to grasp the outline of the domain. It can also contribute to the refinement of several techniques in citation analysis. We therefore attempt to extract citation information.

Extraction of citation information consists of two processes: extraction of citing areas and identification of citation types. Citing areas are defined as a succession of sentences that have a connection with the sentence that includes the citation in the paragraph. As we believed that such a connection between sentences could be indicated by some cue phrases, we used those cue phrases for citing area extraction. As a result, we obtained recall of 80 % and precision of 76 %. We then proposed a method to identify citation types automatically using several cue phrases. As a result, we obtained the accuracy of 83 %.

We use citing areas and citation types for support to write a survey article. To write a survey article, at least two processes are necessary. One is to collect papers from some domain. Another is to make clear the differences between papers. We believe that citation information is useful for both these processes. Making use of citation types, we can collect a set of papers in the same domain. Finally, we build up a system to display the citation graph of the papers. With our system, abstracts and citing areas of papers can be seen. Users of this system can easily collect papers from some specific domain and also can understand the differences between the related papers.

We also use citation types for classification of research papers. It is well known that using citation analysis makes it possible to obtain topical collections of papers. However, most previous research in citation analysis treats all citations equally. We therefore refine citation analysis by taking account of citation types. The results of our experiments showed that our method based on bibliographic coupling ("BCCT-C") is more effective than other methods.

In this thesis, we focus on support for writing a survey article and for classification of research papers. Citation information is also generally applicable to other purposes (e.g., the analysis of research fields and

the evaluation of research papers). It is also applicable to other genres of texts, such as patents and texts on the World Wide Web.

Keywords: citation analysis, extraction of citation information, multi-paper summarization, classification

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の背景

論文中で先行する著作を参照[1]する習慣は, 19 世紀に確立したと言われている [40]. 論文の参照は情報流通の 1 つの形態と考えることができ, また, 参照論文を分析することで文献間の結び付きを見出したり, 論文や研究者の評価を行うことは, 十分意義深いことであると考えられる [63]. このような分析は, 一般に引用分析 (citation analysis) と呼ばれている.

論文間の参照・被参照関係は, これまで様々な目的に利用されている. 例えば, 論文間の関連度を測るための尺度 [21, 54], 研究領域の分析 [43, 17], 論文誌や論文の重要性の評価 [13, 44] などが挙げられる. しかし, これらの多くはすべての参照を同等に扱っており, 引用分析の研究が始まった当初から問題点として指摘されていた.

この問題に対し, 参照・被参照論文の関係を分類する必要性が唱えられ, また, 参照を分類するためのいくつかのカテゴリが提案された [35, 3, 71]. 学術論文の検索や分類は, 参照カテゴリを考慮することで精度の向上が期待できる. しかし, 論文の参照カテゴリを自動的に判別する方法は実用化には至っていない. 一方, 現状の学術情報の爆発的な増加を考えると, 人手ですべての論文の参照カテゴリを判定するのは不可能であり, 参照カテゴリ自動判定の技術の実現が望まれる.

論文中には, その論文が参照している論文 (被参照論文) の重要点や, 当該論文と被参照論文との関係について記述された個所 (以後, 参照個所) が存在する. この個所を読めば, 当該論文の著者がどのような理由で被参照論文を参照したのか (以後, 参照タイプ) が分かる.

[1]「参照」と「引用」という言葉の使い分けについては, [43] 等を参照. 本論文では, 「参照」を用いる.

本研究では, 論文間の参照・被参照関係, 参照個所, 参照タイプをまとめて参照情報と呼ぶ.

参照個所中には, (1) 参照論文の著者から見た被参照論文の重要点, (2) 参照・被参照論文間の類似点, (3) 相違点と言った情報が記述されている. 論文間の参照・被参照関係は, 研究者が関連論文を収集するのに使われる重要な情報源の 1 つであると言われているが [63], どの参照を辿って論文を収集するかは, 参照個所中の (1)–(3) の情報に基づき研究者が判断していると考えられる. さらに, 研究者は集めた論文を対比・分類し, 分析した上で, 論文を書く. 論文間の参照情報は複数の論文を関連付ける上で重要な情報であり, 参照情報を利用すれば, このような作業の支援が可能になると考えられる.

また, ある論文を参照する複数の論文の参照個所を集めて読めば, その論文に関する様々な解釈や評価, 関連論文の中での位置づけなどが分かる. こうした情報は, その分野の初学者がより深く論文を理解する手助けとなる. また, 初学者ばかりでなく, サーベイ論文の作成者が, 論文間の関係を分析するのにも役立つと考えられる.

さらに, 多くの論文から評価されている論文は分野の中心的な役割を担っていると考えることができるが, このような重要論文を発見するのにも, 参照情報が有用であると考えられる.

1.2 研究の目的

本論文では, 参照情報を自動的に抽出する手法を提案し, 抽出した参照情報をサーベイ論文作成支援や複数論文の分類に利用し, その有効性を示す.

1.2.1 参照情報の抽出

論文間の参照情報の抽出は, 参照個所の抽出と参照タイプの決定という 2 つのタスクに分けられる. 本研究では, いずれの処理も手がかり語に基づき, 自動的に行う.

参照個所の抽出は, 論文中で参照が出現する文と文間のつながりが強いと考えられる文を, 参照の前後の文から抜き出すことで実現可能になると考えられる. このような文間のつながりとして, 例えば接続詞や照応詞といったものが考えられる. また, 抽出された参照個所中で, “However” や “can not” といった否定的な表現が出現すれば, その参照個所では既存の研究の問題点を指摘をしていると考えられる. 逆に “We adopt” や “We use” といった表現が出現すれば, 理論的な根拠を示すための参照であると考えることができる. このよ

うな手がかり語に注目して参照個所を解析すれば, 参照タイプの自動推定が可能になると考えられる. 本研究では参照個所抽出や参照タイプ決定の手法を提案し, 実験により提案手法の有効性を確認する.

1.2.2 サーベイ論文作成支援

近年, 学術情報量の増加と共に, サーベイ論文の需要が益々高まりつつある. 一方, サーベイ論文を書くことは人間にとって非常に負荷のかかる作業であり, その負荷を軽減するためにサーベイ論文作成支援の技術が必要とされている. 研究者がサーベイ論文を書く際, 関連論文を収集し, それらを対比する必要があるが, 本研究ではこのような人間の作業を参照情報を利用して支援する.

サーベイ論文作成には少なくとも 2 つの処理 (1) 特定分野の論文の収集, (2) 論文間の相違点の抽出が必要であると考えられる. 本論文では, 参照情報を用いることで (1)(2) の処理が部分的に実現可能であることを示す.

(1) において, 論文間の参照・被参照関係を辿ることで, ある程度関連論文を集めることが可能になる. しかし, 単純に参照・被参照関係だけで辿ると, 分野外の論文まで収集してしまう可能性がある. 本研究では参照タイプを考慮し重要な参照だけを辿ることにより, 特定分野の論文の自動的な収集を試みる. また, 論文中の参照個所は, 当該論文と被参照論文との関係について記述されているが, 参照個所を抽出しユーザに提示できれば, ユーザは収集された論文間の関係を容易に把握することができる. 本研究ではワールド・ワイド・ウェブ上の論文データベースを対象にし, 提案手法を CGI を用いて計算機上に実現する.

1.2.3 関連論文の分類

サーベイ論文作成支援システムを用いて特定分野の論文を収集し, それらの概要や参照個所を読めば, その分野の研究動向を把握することができる. さらに, 関連論文を論文の内容 (トピック) に応じて自動的に分類しておけば, 効率的にこれらの情報を得ることが出来る.

これまで, 学術論文をトピックの類似度に基づいて分類するいくつかの手法が提案されてきたが, その 1 つに論文間の参照・被参照関係を用いた手法 [21, 54] がある. Kessler[21] が提案する書誌結合は「トピックの似た 2 つの論文は, 多くの他の論文を共通に参照する」という性質を利用した 2 論文間の類似度を評価する尺度である. しかし, Weinstock[71] が示すように, 参照には様々な参照の理由 (参照タイプ) が存在する. 従って, 複数の論文をより

正確に分類するためには, 参照・被参照関係だけでなく, 参照タイプも考慮することが不可欠であると考えられる. 本研究では, 2 論文間で同一論文を共に参照しており, かつ参照タイプが一致している結合のみを数えるという方法で, 2 論文間の類似度を測る. このような手法を用いることで, ノイズとなる結合を削減でき, 従来手法と比べ精度向上が期待できる.

本研究では提案手法と書誌結合, 語の共出現に基づいたいくつかの分類の手法を計算機上に実装する. また, 各手法を精度, フォールアウト, 計算時間により比較し, 提案手法の有効性を確認する.

1.3 論文の構成

2 章では, まず, 論文間の参照・被参照関係の分析や利用に関する諸研究をサーベイし，既存の研究の問題点を明らかにする. 次に, 本研究で提案する論文間の参照情報を説明する. ここでは, 参照個所, 参照タイプの定義を行う. また，参照情報の応用についても述べる.

3 章では, 参照情報の抽出方法を説明する. 参照情報の抽出は (1) 論文間の参照・被参照関係の解析, (2) 参照個所の抽出, (3) 参照タイプの決定, という 3 つのステップに分けられる. 各ステップを実現する手法を提案し, また提案手法の有効性を確認するためにそれぞれ実験を行う.

4 章と 5 章では，参照情報の応用例を示す.

4 章では, 参照情報を考慮したサーベイ論文の作成支援について述べる. ここでは，まず，サーベイ論文作成のポイントを説明し，次に, サーベイ論文を作成する上での参照方法の利用方法について述べる. また, サーベイ論文作成支援システムを計算機上に構築し, 実際のシステムの動作例を示す.

5 章では, 関連論文の自動分類手法を提案する. これまでの分類手法は，(1) 語の共出現を用いる手法と，(2) 論文間の参照・被参照関係を用いる手法の 2 種類に分けることができる．これらの先行研究と，その問題点について述べる. さらに, 問題点を解消する新しい分類手法を提案し，実験によりその有効性を確認する. また，提案手法を用いて, 4 章で述べたサーベイ論文作成支援システムを拡張し, その動作例を示す.

6 章では, 結論と今後の課題について述べる.

第 2 章

論文間の参照情報

本章では, まず 2.1 節で参照に関する諸研究を紹介し, その問題点について述べる. 2.2 節では, 論文間の参照情報を定義する. 2.3 節では, 参照個所から得られる情報を具体例を挙げて説明する. 2.4 節で, 参照情報の応用について述べる.

2.1 参照に関する研究

2.1.1 引用分析に関する研究

文献を様々な側面から計数し分析する研究領域は, 計量書誌学 (bibliometrics) と呼ばれている. この中でも, 特に参照・被参照のある文献に対する分析は引用分析 (citation analysis) と呼ばれている. 論文間の参照・被参照関係は, これまで様々な目的に利用されてきた. これらは大きく (1) 論文誌, 論文等の重要度の評価 [11, 13, 5, 6], (2) 論文誌間・研究領域間の影響度の分析 [43, 44], (3) 論文の検索・分類 [21, 54] の 3 種類に分けることができる. このうち (3) については 5 章で述べる. また, 学術論文とは異なるが, 特許文書における参照やワールド・ワイド・ウェブに代表されるようなハイパーテキスト, 裁判における判例の参照等も, 論文間の参照・被参照関係と類似した文書間の構造であると考えることができる. 本節では, 学術論文以外の文書を用いた関連研究についても述べる.

(1) 論文, 論文誌, 研究者の評価

被参照数を数えることで論文誌の重要度を定量的に評価した最初の試みは, Gross らの調査であろうと言われている [13]. この考え方は, 学術論文だけでなく, 近年ではワールド・ワイド・ウェブにおける文書検索にも利用されている. ウェブ文書の場合も, 学術論文と同様, 多くのページから参照されるページは多くの人の関心を集める重要度 (影響度) の高いページであると考えることができる. サーチエンジン google (http://www.google.com) では,「多くの良質なページからリンクされているページはやはり良質なページである」という考え方に基づき, あらかじめ個々のウェブページの被参照数 (重要度) を数えておく. そして, 検索クエリが与えられた時, そのキーワードを含むページを探し, 重要度順にページを並べて検索結果として出力する [48].

サーチエンジン CLEVER も, google と同様, ウェブ文書間の参照・被参照関係をウェブ文書検索に取りいれている [6]. CLEVER は, ウェブのリンク集に相当するハブと呼ばれるページと多くのハブから参照されるオーソリティと呼ばれる 2 種類のページを考慮している.「多くのハブから参照されるオーソリティは重要である」「多くの重要なオーソリティを参照するハブは重要である」という考え方に基づき, 個々のウェブ文書の重要度を測る.

この他にも, 参照・被参照関係により文献の重要度を測る研究, 調査報告がいくつかある. ここではその中でも特に「インパクト・ファクタ (Impact Factor)」と「引用半減期 (Cited Half Life)」という代表的な 2 つの尺度を紹介する.

インパクト・ファクタ

図書館で購入すべき雑誌の選定する場合, 論文誌の重要度を客観的に測る尺度が必要とされる. これには, 単純に論文誌の被参照数を論文誌の重要度とみなす, という方法がある.

しかし, この方法では, (1) 出版論文数の多い大規模雑誌が小規模雑誌よりも有利になる, (2) 歴史の長い雑誌が創刊間もない雑誌よりも有利になる, という問題点がある. これらの問題点を克服するために Garfield によりインパクト・ファクタという指標が提案された [11, 73, 23]. インパクト・ファクタは以下の式 2.1で定義される.

$$\text{インパクト・ファクタ} = \frac{\left(\begin{array}{c}\text{当該雑誌の, 過去 2 年間に発表された論文が, その年の}\\ \text{1 年間に発行されたすべての雑誌に参照された総件数}\end{array}\right)}{\left(\text{当該雑誌の, 過去 2 年間に発表された論文の総件数}\right)} \quad (2.1)$$

式 2.1の分母は問題点 (1) を考慮している. また, 対象を過去 2 年間に発表された論文に限定することで, 問題点 (2) に対応している.

引用半減期

引用半減期とは, 現在から過去へ遡って, その雑誌の被参照回数が被参照総数の 50%になるまでの年数である. 従って, ここ数年で急激に参照されるようになった論文誌の引用半減期は短く, 逆に過去にはしばしば参照されたが, 近年ほとんど参照されない論文誌は引用半減期が長くなる. このような指標は, 図書館で古い文献を破棄する時, あるいはある論文誌のバックナンバを購入する時, どの年代までさかのぼれば良いかの判断などに用いることができる.

(2) 論文誌間・研究領域間の影響度の分析

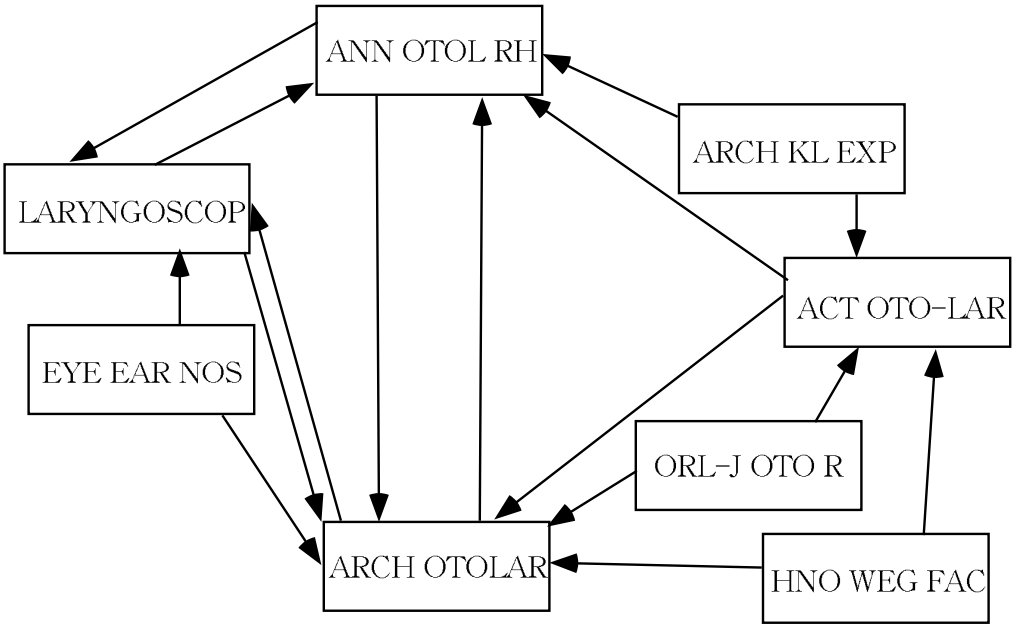
インパクト・ファクタは論文誌を重要度順に順位付けする指標であると言えるが, Narin はこれからさらに一歩進め, 参照・被参照関係を用いて論文誌間 (研究領域間) の関係を明らかにし, 2 ステップ・マップと呼ばれる図 (図 2.1) に表している [43]. 2 ステップ・マップでは, 個々の雑誌を四角で囲んで表し, それぞれの雑誌から自誌への参照を除いて最も頻繁に参照する 2 誌へ向かう 2 本の矢印をひく. こうして作成された図を見ることで, 論文誌間の影響度が容易に把握できる.

また, Narin らは, 科学 (Science) と技術 (Technology) の関係を, 参照関係を用いて調査している [44]. これは, ある特定の期間に学術論文 (科学) がどれだけ特許 (技術) から参照されているか, により測っている. Narin らの調査では, アメリカ, イギリス, 旧西ドイツ, 日本, フランスの 5ヶ国の論文および特許を使用し, 各国内だけでなく国外技術への影響度についても調べている.

2.1.2 参照個所に関する研究

論文中の参照の出現する前後には, 被参照論文に関する記述 (参照個所) があり, 参照個所中の情報に着目した研究がいくつかある [33, 14, 70, 1].

三平らは, 論文中の参照の出現する文の前後から, 手がかり語を用いて著者の主題表現を抽出している [33]. さらに抽出された主題表現中に含まれる語が論文検索をする上で有用であるとの分析結果を示している. 三平らの主題表現は本研究における参照個所に相当すると考えられる.



((Narin, 1976)[43] より抜粋)

図 2.1: 耳鼻咽喉科学の 2 ステップ・マップ

また, ワールド・ワイド・ウェブの文書検索においても, 参照個所の情報が利用されている. 通常の検索エンジンは, 各ウェブページに含まれる語を用いてインデックスを作成する. これに対し, サーチエンジン ODIN(http://odin.ingrid.org/) では, あるページに含まれる単語だけでなく, そのページを参照するページのアンカー文字列[1]をインデックスに含めている [14]. これには 2 つの利点がある. ひとつは, 例えば「NTT」の公式ウェブページは, 「NTT」というアンカー文字列で参照されることが多いため, そのページは「NTT」という語に対して高い適合度を得るようになる. もうひとつは, 「NTT」の公式ウェブページが「エヌティティ」というアンカー文字列で参照されていれば, 被参照ページに一度も「エヌティティ」という語が出現しなくても検索可能になる.

鷲崎らは, ODIN と同様, 参照ページのアンカー文字列に着目している [70]. 情報検索の結果を表示する際, 各ページを参照するページ集合のアンカー文字列を並べて表示すれば, それらがページの適合度を判断するのに役立つと考えている.

Amitay は, あるウェブページに関する複数の参照個所 (annotation) から, そのページの要約として最もふさわしい参照個所を選択する手法を提案している [1]. この参照個所は,

[1]HTML で A タグによって囲まれた文字列をアンカー文字列とここでは呼んでいる.

ウェブ・サーチエンジンにおいて, 検索結果を表示する際, 検索された各ページの要約として表示される. 参照個所は, HTML の段落タグ (<P>) や区切り線を示すタグ (<HR>) や HTML 文書中の空行に着目して抽出する. 次に抽出された複数の参照個所から, 要約として最も適当な個所を一つ選択するが, その際 15 種類の特徴に着目している. すなわち, 参照個所の長さ, 句読点の有無, 人称代名詞の有無, 頭字語 (UNESCO 等, 大文字のみの語) の有無, 意見を表す語句の有無, 内用語の有無, 参照個所中の動詞の位置, 大文字で始まる文数, 句読点の位置, 反復語の頻度等に着目し, 機械学習により 15 種類の特徴を組み合わせた抽出ルールを自動的に獲得している.

2.1.3 引用文脈分析に関する研究

2.1.1 節では引用分析に関する研究を紹介した. 引用分析研究では,「すべての参照・被参照関係が等価である」ことがその前提条件になっている. しかし, 実際にはこのような仮定が成り立たない [69, 35]. この問題点を解消するために, 被参照論文は参照論文の中でどのような文脈で参照されているか, を分析する研究 (引用文脈分析: citation contexts analysis) が行われるようになった.

牛澤は [67] 引用文脈分析研究を, 図 2.2に示す (1) 参照・被参照論文の性質, (2) 参照・被参照論文の関係, (3) 被参照論文の記述のされ方, の 3 種類に分類している. 以下, この 3 分類について簡単に述べる.

(1) 参照・被参照論文の性質

Bonzi は, 引用索引による検索効率の改良を目的に, それまで参照について研究されてきた変数を用いて 13 のカテゴリーリストを設定し, そのうちの何が参照・被参照文献間の関係をよく表すのかについて多角的に分析した [3]. 分析の結果, 被参照文献の種類, 参照文献収載雑誌の種類, 参照論文の種類, 被参照文献についての複数回の参照が明らかにされた.

(2) 参照・被参照論文の関係

これまで, 以前の論文を参照するということについて, 多くの理由が考えられてきた. Weinstock は, 参照の理由を図 2.3に示す 15 種類に分類している [71].

Moravcsik ら [35], Chubin ら [7] は, 2.1.1 節で紹介した, 論文や論文誌の質の, 単純な参照回数での評価に疑問を持ち, 参照回数だけでなく, 参照論文中の文脈を考慮する必要性を述

引用・被引用論文の性質

引用論文の種類

(Lipetz '65)	(Ruff'79)	(Bonzi'82)
レビュー論文	a 包括的レビュー	レビュー
書誌		史的
データ集		調査
		報告
		ガイドライン

被引用論文の種類

(Bonzi'82)
研究論文
マガジン・新聞
単行書
未発表 等

引用論文の長さ

(Moravcsik'75)
big papers
small papers
(引用文献数25を基準として)

(Bonzi'82)
語数

その他

引用・被引用論文の分野 (Bonzi'82)
図書館・情報学
その他

引用・被引用論文の刊行年 (Bonzi'82)

引用者の性別 (Bonzi'82)

引用・被引用論文の関係

(Lipetz'65)
12 自己引用
13 同じテクスト
14 抄録あるいは圧縮
16 続報

(Bonzi'82)
自己引用
自誌引用

(Moravcsik'75)
1 概念的 or 実際的
2 本質的 or 形式的
3 直列的 or 並列的
4 肯定的 or 否定的

(Chubin'75)
肯定的
不可欠 (基礎的 – 補助的)
補足的 (付加的 – 形式的)
否定的
部分的 – 全体的

(Small'78)
概念シンボル

(真弓'84)
1 基礎的引用
2 補助的引用
3 不可的引用
4 儀礼的引用
5 全面否定引用
6 部分否定的引用

被引用論文の記述のされ方

(Lipetz'65)
19 言及のみ
24 言い替え

(Ruff'79)
抜き書き
引用

(Bonzi'82)
特に言及なし
わずかに言及
引用または検討

引用回数

(Voos'76)
op. cit. の数

(Herlach'78)
一回
複数回

(Bonzi'82)
0回, 1回, 2回, 3回以上

引用されている位置

(Voos'76)	(Spiegel-Rosing'77)	(Herlach'78)	(Bonzi'82)	(Peritz'83)
Introduction	Introduction	Introduction	1st quarter	Introduction
Methodology	Discussion	Method	2nd quarter	Method
Discussion		Results	3rd quarter	Results
Conclusion		Discussion	4th quarter	Discussion & Conclusion

他の文献との関係

(Ruff'79)
b 他論文と共に引用

(Bonzi'82)
引用文献数
脚注の引用数

(牛澤, 1992)[67] より抜粋

図 2.2: 参照・被参照関係の分類

(1) 先駆者に敬意を表する.

(2) 関連論文を承認する.

(3) 方法, 器具などを確認する.

(4) 背景となる基礎的文献を紹介する.

(5) 自分自身の書いた論文を訂正する.

(6) 他者の書いた論文を訂正する.

(7) 以前の論文を論評する.

(8) 主張を強固にする.

(9) 研究者にまもなく発表される論文を知らせる.

(10) あまり知られていない論文, 索引誌に掲載されていない論文, または引用されたことのない論文への手懸かりを提供する.

(11) 事実についてのデータや種類を明らかにする - 物理学の定数など.

(12) 特定の考えや概念を最初に述べた論文を紹介する.

(13) 概念や用語の名称となった学者の書いた原文献を確認する.

(14) 他人の書いた論文や考えを否認する.

(15) 優先権の主張で論争する.

(Garvey, 1979)[10] より抜粋

図 2.3: Weinstock の 15 種類の参照の理由

べている. また, 参照カテゴリーを提唱し, 実際に参照の分類を試みている. Moravcsik らは高エネルギー理論物理学の分野の論文 30 本が参照する 575 文献について, 図 2.2(2) に示すカテゴリを用いて調査している. 調査の結果, 全参照のうち 41%が形式的な参照であることを報告している.

(3) 被参照論文の記述のされ方

先に述べた Bonzi ら [3] の他にも, Voos ら [69], Herlach ら [15] の調査の結果, 参照論文中で複数回参照される被参照論文は, 一度しか参照されない論文と比べ, 参照論文と深い関連があると報告している.

また, Voos ら, Herlach らは, 参照論文の章立てに従って参照の位置を分け, 高頻度被参照論文が Introduction でしばしば参照されていることを明らかにしている. これに対し, Bonzi は論文を機械的に 4 等分し同様の調査を行っているが, 統計上有意な差は認められなかったとしている. また, このような考え方に基づき, 参照・被参照関係を自動的に分類する試みがいくつかあるが [19, 61], その具体的な手法については 3 章で紹介する.

判例の分類

アメリカの裁判は「判例主義」であり, 過去の類似した事件の判例に基づいて弁論を組み立てて弁論する必要がある. 個々の弁論は過去の複数の判例を参照することで形成され, それが新たな判例となる. 従って, 判例間には学術論文と同様, 参照・被参照関係が存在する.

弁護士や検察官は判例を参照する際, 過去に判例が破棄されていないか, 却下されていないか, 疑義が生じていないかなどを事前によく調べておく必要がある. シェパード・サイテーション (http://www.bender.com/bender/open/) やキーサイト (http://www.keycite.com/) はこのような要求に応じて作成された判例データベースである.

図 2.4は, シェパード・サイテーションの抜粋である. 上部の “101 MASS.210” はマサチューセッツ州の判例集第 101 巻, 210 頁所載の判例を示している. また, その下は “101 MASS.210” を参照したその後の判例を示している. 図 2.4の左端に付いている小文字のアルファベットはそれぞれ,

101 Mass.210

	101	Mass.	65
a	130	Mass.	89
	165	Mass.	210
q	192	Mass.	69
	205	Mass.	113
o	221	Mass.	310
	281	U.S.	63
	35	H.L.R.	76

(窪田, 1996)[23] より抜粋

図 2.4: シェパード・サイテーションの例

- **a: affirmed** 再確認
- **q: questioned** 疑義
- **o: over-ruled** 却下

を示している. これらは判例における参照・被参照関係の分類指標の一種と考えることができる.

2.1.4 考察

論文間の参照・被参照関係は, 2.1.1 節で紹介したように様々な方面で応用されている. 一方で, 2.1.3 節の冒頭でも述べたように, これらの研究では「すべての参照は等価である」ことが前提になっているが, この仮定には明らかに無理があり, 2.1.3 節で述べた参照・被参照関係を何らかの観点に基づいて分類することは必要不可欠であると考えられる. しかし, これらの研究で提案された参照・被参照関係の分類指標は, 人間が判断することを前提に作られており, 計算機による自動処理には馴染まない. 従って, 参照の理由は, 既存の分類指標とは異なる, 計算機による自動処理を考慮した分類を設定する必要があると考えられる.

以下 2.2 節では, 本研究における重要概念である参照情報を定義する. また, その自動抽出方法については 3 章で述べる.

2.2　参照情報の定義

論文中には, その論文が参照している論文 (被参照論文) の重要点や, 当該論文と被参照論文との違い等について記述された個所 (参照個所) (図 2.5) が存在する. この個所を読めば, 当該論文の著者がどのような理由で被参照論文を参照したのか (参照タイプ) が分かる. 本研究では, 論文間の参照・被参照関係, 参照個所, 参照タイプをまとめて**参照情報**と呼ぶ.

参照情報を例を挙げて説明する. 図 2.6の 5 文は参照論文中 (Bond, 1996)[74] で被参照論文 (Murata, 1993)[78] を参照している文の前後数文を抜粋したものである. (Bond, 1996), (Murata, 1993) は共に, 機械翻訳に関する論文で, 特に数詞表現について取り扱っている. 文 (2) は, (Murata, 1993) について, どのような問題を取り扱った論文であるかについて述べている. 文 (3) は, (Murata, 1993) の問題点を指摘している. そして文 (4) は, (Bond, 1996) がその問題点を考慮した論文であると述べている.

ここで, 参照論文 (Bond, 1996) と被参照論文 (Murata, 1993) の関係は文 (2)–(4) を読めばわかる. このような参照・被参照論文の関係が明示されている個所を**参照個所**と呼んでいる. また, 参照個所を読めば被参照論文の参照の理由 (既存研究の問題点の指摘) が容易に理解できる. 参照の理由を計算機で自動分類するには, 先にも述べたとおり, 計算機処理の可能性と参照の理由の重要性を考慮して新たな指標を設定する必要がある.

(Bond, 1996) における (Murata, 1993) のような, 既存研究の問題点を指摘する参照は重要な参照の理由の一つであると思われる. 何故ならば, このような参照は, 参照論文の問題提起や研究動機を示していると考えられるからである.

また, 問題点指摘型の参照の自動判定は, 否定表現に着目することで可能になると考えられる. 例えば, (Bond, 1996) の場合, 文 (3) に出現する “less studied” という否定的な表現が, 参照が否定的であるか否かを判断する上での手がかりになる.

この他に重要と考えられる参照の理由の一つに, 論説根拠型がある. 論説根拠型の参照とは, 新しい理論・手法を提唱するのに用いた既存の理論, ツール, データ等に関する参照を指す. これは, 参照論文の提案手法を形成する上で基盤となる既存研究の参照であり, 参照論文を特徴づける重要な参照であると考えられる. また, 論説根拠型の参照を自動判定は, “use” や “base” や “adopt” のような語に着目すれば可能になると考えられる.

この他にも参照の理由は存在しうるが, 特徴的な手がかり表現があるとは考えにくく, この他の参照の理由の自動判定は困難であると考えた. 従って, 本研究では参照の理由を次に示す 3 種類の参照タイプに分類する.

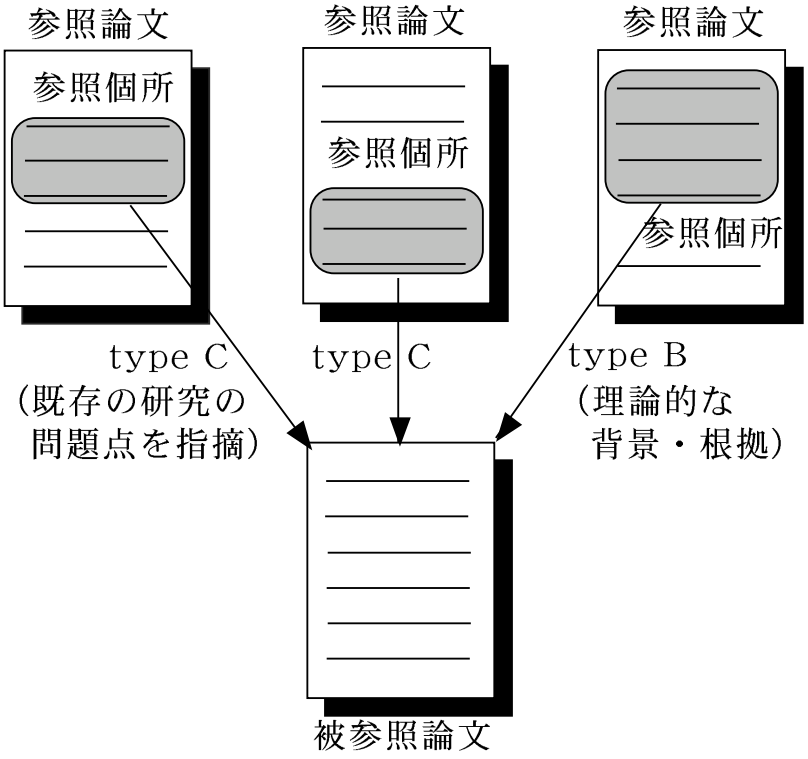


図 2.5: 論文間の参照・被参照関係

(Bond, 1996)[74] より抜粋

(1)In addition, when Japanese in translated into English, the selection of appropriate determiners is problematic.
(2)Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed (Murata, 1993).
(3)The differences between the way numerical expressions are realized in Japanese and English has been less studied.
(4)In this paper we propose an analysis of classifiers based on properties of both Japanese and English.
(5)Our category of classifier includes both Japanese josūshi 'numeral classifiers' and English partitive nouns.

参照個所 文 (2)–(4)

図 2.6: 参照個所の例

- **type C (問題点指摘型)**
  新しく提案した理論や, 構築したシステムの新規性について述べる場合, 関連研究との比較, あるいは既存研究の問題点の指摘を行う場合がある. このような目的の参照タイプを **type C(問題点指摘型)** と呼ぶ.

- **type B (論説根拠型)**
  新しい理論を提唱したり, システムを構築する場合, 他の研究者の研究の成果を利用する場合がある. 例えば, 他の研究者が提唱する理論や手法を用いて新しい理論を提唱する場合などである. このような参照タイプを **type B(論説根拠型)** と呼ぶ.

- **type O (その他型)**
  type B にも type C にも当てはまらない参照を **type O(その他型)** と呼ぶ.

以下, 例を挙げて type B, C, O をそれぞれ説明する [41].

type B(論説根拠型)

> The analysis introduced in this paper has been implemented in NTT Communication Science Laboratories' Japanese-to-English machine translation system ALT-J/E [Ikehara, 1995].

これは, ある論文の著者が過去の自分の論文を参照している例である. この例では, 過去の研究で開発したシステムを用いて分析を行っている.

> There are various definitions for TFS unification, and we base our unification algorithm on the definition given in [Carp, 1992].

論文の著者が他の研究者の提案を基に, 新たな提案を行う場合がある. この例では [Carp, 1992] の単一化アルゴリズムを基に, この論文の著者が新たに TFS 単一化アルゴリズムの定義を行っている.

> The corpus we have used is the 1988, 1989 Wall Street Journal[Liberman, 1991].

実験用コーパスとして Wall Street Journal を用いている.

type C(問題点指摘型)

> Recently, rule-based approaches are re-studied to cope with the limitations of statistical approac hes by learning the tagging rules automatically from the corpus [Brill, 1994]. Some systems even perform the POS tagging as part of syntactic analysis process [voutilainen, 1995]. However, the rule-based approaches alone are in general not robust to handle the unknown words.

この例では, ルールベースの品詞タグ付け手法は, 未知語の取り扱いの点で頑健さに欠けるという既存研究の問題点を指摘している.

> Previous work on extraction of collocation for use in generation [Smadja, 1991] is ... However, extracted collocations were used only to determine realization of an input concept. In our work, stored phrases would be used to provide content...

既存研究 [Smadja, 1991] の問題点を指摘した後, その問題点に対する著者らの提案を述べている.

type O(その他型)

> The first experiment on automatic abstracting was reported in a paer by H.P.Luhn published in 1958 [Luhn, 1958].

この例は, テキスト自動要約の原点と言える Luhn の論文の参照である. このような古典的な論文はこの分野の多くの論文から参照されるが, 形式的な参照が多く, 論文の細部について言及されることはあまりない.

以上は, type B, C, O の典型的な事例であったが, 以下のような type B とも type C とも捉えることのできる参照個所も存在する.

> In a previous paper **[Schütze, 1993]**, we trained a neural network to disambiguate part-of-speech using context; however, no information about the word that is to be categorized was used. This scheme fails for cases like "The soldiers *rarely* come home." vs. "The soldiers *will* come home." where the context is identical and information about the lexical item in question ("rarely" vs. "will") is needed in combination with context for correct classification. In this paper, we will compare two tagging algorithms, one based on classifying word types, and one based on classifying words-plus-context.

参照論文の著者は, 過去の自分の研究 ([Schütze, 1993]) について述べた後, その問題点を指摘している. 従って, この参照タイプは type C であると考えることができる. しかし, 3 文目で述べる提案手法の一つは過去の研究に基づいた手法 ("words-plus-context") であり, 従って, type B であるとも考えられる.

このように, 参照個所の中には複数の参照タイプを割り振ることが可能なものもある.

関連研究

Teufel は [61], 手がかり語を用いて学術論文の構造解析を行っている. Teufel は, 学術論文は 7 種類の要素 ("BACKGROUND (背景)", "OWN (提案手法, 結果)", "AIM (目的)", "TEXTUAL (論文の構成)", "CONTRAST (関連研究との対比, 問題点の指摘)", "BASIS (基礎, 根拠となる参照)", "OTHER (その他の参照)") から構成されると考えている. この中で論文の参照の理由に関するものは "CONTRAST", "BASIS", "OTHER" の 3 つである. これらの分類は, それぞれ本研究の参照タイプ C, B, O に対応する. また, 参照の理由を 3 つに分類している理由を次のように述べている. 一つは, これまで提案されてきた参照の理由の分類指標の多くは, 参照が肯定的か否定的 (対比的) かを区別しており, "CONTRAST" は重要な参照の理由であると考えられること. 一つは, 肯定的な参照の中でも, "BASIS" は既存研究に全面的に同意しているという点で, 他の肯定的な参照とは異なること. そしても

$$\begin{cases} (\alpha) & \text{既存研究の紹介} \\ (\beta) & \text{既存研究の問題点} \\ (\gamma) & \text{参照論文の研究の目的} \end{cases}$$

図 2.7: type C の参照個所中の記述

う一つは, 多くの事例において, この他の参照の理由は, 言語的な手がかりが得られないため自動判定が困難であることである. 本研究でも, Teufel のこの考えに同意する.

2.3 参照個所から得られる情報

本研究では, 3 つの参照タイプの中で type C が最も重要であると考えている. 何故ならば, type C の参照個所からは, 参照・被参照論文間の相違点に関する情報が得られるからである. type C の参照個所から得られる情報を図 2.7に示す. 図 2.6の例の場合, 文 (2) が (α) に, 文 (3) が (β) に, 文 (4) が (γ) にそれぞれ対応する. ここで, (α) は参照論文の著者の観点から見た被参照論文の一種の要約であると考えられ, 同時に参照・被参照論文がどのような観点で共通点があるのかを示している個所であると捉えることもできる. 文 (2) では, 参照・被参照論文の両方が, 冠詞, 所有代名詞, 可算・不可算, 数詞等の生成を問題にしている論文であると述べている. 一方, 既存研究の問題点と著者の研究の目的が文 (3), (4) に書かれており, これが論文間の相違点と考えられる. このような情報は, 特定分野の研究動向を効率的に知る上で有用であると思われる.

次に, 図 2.7に示す (α), (β), (γ) が type C の参照個所に一般的にどの程度含まれているのか, 実際の論文データを用いて調査した. 調査には, E-Print archive という論文データベースの "The Computation and Language" の分野の TEX 形式の論文 (http://xxx.lanl.gov/cmp-lg) を用いた. まず, この論文データベースから, 人手に基づいて type C の参照個所を 51 個所抽出した. これらの参照個所のうち, 図 2.7に示す要素 (α), (β), (γ) の含まれる割合を調べた. 結果を表 2.1に示す.

表 2.1から分かるように, すべての type C の参照個所には必ず「(α) 既存研究を紹介」する記述と「(β) その問題点を指摘」する記述が存在していた. また, 20 個の参照個所 (39 %) において, 「(γ) 研究の目的」が書かれていた. 残りの 31 個で (γ) が記述されていないのに

は, 2 つの理由がある. 一つは, Introduction 等の章において, まずいくつかの既存の研究にまとめて言及した後, 最後に総括して提案手法等について述べる場合である. この時, 論文中の参照から離れた個所に (γ) が出現する.

もう一つは, 関連研究を “Discussion(議論)” や “Future work (今後の課題)” といった章に参照がある場合である. Discussion や Future work では, それより前の章で提案手法や結果について述べているため, 改めて提案手法について説明されることはほとんどない.

一方, 「(γ) 研究の目的」は参照個所に書かれていなくても, 参照論文の概要を読めばわかる. 従って, 参照個所中に (γ) の記述がないこと自体はそれほど重要な問題ではない. もし, (γ) を本文中から補うのであれば, 参照個所の前後から, “In this paper” や “we propose” といった語句を含む文を提示するといった方法も考えられる

2.4　参照情報の応用

参照タイプの利用は以下のような点で有用である.

2.1.1 節で紹介したインパクト・ファクタ [11] や引用半減期などは, 参照タイプを考慮することで, 精度の向上が期待できる. 例えば, ある論文がより多くの論文から type B で参照されていれば, その論文は特定分野における重要 (あるいは基礎的) な理論やツールを提案していると考えられる.

この他に, 研究領域間の影響度を分析する上でも [44], どのように影響を及ぼしたかを把握するために, 参照タイプは重要な役割を果たすと考えられる.

また, これまでに論文間の参照・被参照関係を利用した関連論文を分類する手法がいくつか提案されているが [21, 54], これらの手法においても, 参照情報を利用することで, 精度の向上が期待できる. 5 章では, 参照情報を用いた関連論文の分類手法を提案し, またその有

表 2.1: type C の参照個所に含まれる要素

type C の参照個所に含まれる要素	割合
(α), (β), (γ)	39 % (20/51)
(α), (β)	61 % (31/51)

効性について述べる.

一方, 参照個所の利用は以下の点で有用である.

2.2.2 節で述べた三平の研究において [33], 関連論文の検索に参照個所中の語を利用する場合においても, 参照情報が利用可能である. 例えば, type C の参照個所中に含まれる語を集めてそれらを検索クエリと考えれば, 問題提起の似たような関連論文をより多く集められる可能性がある.

さらに, 参照個所には, 参照・被参照論文間の関係に関する記述があるが, 特定分野の論文のこのような記述を集めて提示すれば, その分野の研究動向を知るのに役立つと考えられる. このような参照情報の利用方法については 4 章で述べる.

2.5 まとめ

本章では, まず, 論文間の参照・被参照関係を用いた諸研究を紹介した. 次に既存研究の問題点を指摘し, 本研究における重要概念である論文間の参照情報 (参照個所, 参照タイプ) について定義した. 参照個所とは参照論文中で, 参照論文と被参照論文について記述された個所である. また, 参照個所を読むことで, 被参照論文の参照の理由 (参照タイプ) が分かる. 参照情報は, 論文の重要度や研究領域間の影響度を知るために有用であると考えられる. また関連論文の検索や分類, サーベイ論文の作成支援といった目的にも利用できる. さらに, 特許やハイパーテキストといった学術論文以外の文書への参照情報の応用も考えられる. なお, 本章で説明した参照情報の応用例は, 4 章と 5 章で述べる.

第 3 章

参照情報の抽出

本章では, 参照情報の抽出方法について述べる. 参照情報を抽出するには, まず, 論文データベース中の論文間の参照・被参照関係を解析する必要があるが, これに関しては 3.1 節で述べる. さらに, 本研究では論文中の参照個所を抽出し, 抽出された参照個所から参照タイプを決定する. 参照個所の抽出手法は 3.2 節で, 参照タイプの決定手法は 3.3 節でそれぞれ説明する. また, 提案手法の有効性を調べるために実験を行った. 実験方法および結果については 3.4 節で述べる. また 3.5 節では関連研究を紹介する.

3.1 論文間の参照・被参照関係の解析

研究対象として, 前章で述べた論文データベース E-Print archive の "The Computation and Language" の分野の論文の TEX ソース 395 本を用いる. この論文データベースでは, 1 論文あたり平均 18.1 本 (7167/395) の論文を参照している. 論文間の参照情報を抽出するには, まず E-Print archive 中の論文間の参照・被参照の関係を解析する必要がある. TEX には参考文献を記述するためのコマンド bibliography があり, これを解析することで自動的に 395 本の TEX ソース間の参照・被参照関係が明らかにできる.

図 3.1は, TEX ファイル [74](cmp-lg/9608014) の参考文献の記述の一部を抜粋したものである. [74] は論文中で [78] を参照している.

一方, E-Print archiveの論文リストファイルを ftp サイトより入手することができる. 図 3.2 はそのリストの一部を抜粋したものである. [74](cmp-lg/9608014) が [78](cmp-lg/9405019) を参照しているという情報を得るには, 図 3.1と図 3.2 の論文が同一であることを判断する

必要がある. そこで, bibliography 中の論文のタイトルや著者名の記述のありそうな個所から単語 (キーワード) を切り出し, 切り出された全ての単語を含むような書誌情報を持つものを論文リストから検索する, という手法で論文間の参照・被参照関係の解析を行う. どのようにして bibliography から検索に有用なキーワードを切り出すかが問題となるが, 参考文献の記述形式に着目する. 齊藤ら [52] によれば, 参考文献の記述形式は多くの場合, 最初に著者名, 次に文献名が記述される. 場合によっては著者名の後に発行日が記述されるケースもある. そこで, 図 3.1のような個々の bibitem の先頭 3 行以内に含まれる単語からアルファベット以外のデータはすべて除去し, 残ったものをキーワードとして利用する. 図 3.1 の場合以下の語がキーワードとなる.

> “Murata”, “Masaki”, “Makoto”, “Nagao”, “Determination”, “of”, “referential”, “property”, “and”, “number”, “of”, “nouns”, “in”

そして, これらのキーワードを用いて E-Print archive の論文リストに対して and 検索をかけ, 論文間の参照・被参照関係の解析を試みた結果, 94 % (94/100) の解析精度が得られた.

```
\bibitem[\protect\citename{Murata and Nagao}1993]{Murata:1993a}
Murata, Masaki and Makoto Nagao.
\newblock 1993.
\newblock Determination of referential property and number of nouns in
Japanese sentences for machine translation into English.
\newblock In {\em Proceedings of the Fifth International Conference on
Theoretical and Methodological Issues in Machine Translation (TMI~'93)},
 pages 218--25, July.
```

図 3.1: TEX ファイル中の bibliography コマンドの使用例

```
\\
Paper: cmp-lg/9405019
Title: Determination of referential property and number of nouns in Japanese
sentences for machine translation into English
Author: Masaki Murata, Makoto Nagao
Comments: 8 pages, TMI-93
\\
```

図 3.2: E-Print archive 論文リスト中の書誌情報の一例

解析に失敗したもの (6 件) の原因として以下のものが挙げられる.

- 同著者, 同タイトルで, 異なる会議で発表された論文の識別が出来なかった (3 件).
- 論文の著者の bibliography 中の書き方に癖があり, キーワードがあまり抽出できず, 少ないキーワードで別の論文が検索されてしまった (3 件).

前者に関しては, キーワードとは別途に, 論文の掲載ページなどを抽出しておき, 同著者, 同タイトルでもページが異なれば別の論文と見なす, といった方法である程度対応できると考えられる. しかし, 書誌情報にページの情報が含まれていない場合 (掲載予定の論文等を参照) にはこの手法では解決できない.

3.2 参照個所の抽出

参照個所の抽出とは, 参照の出現する段落において, 参照のある文と文間のつながりが強いと考えられる文を, 参照の前後の文から抽出する処理と考えることができる. このような文間のつながりは大まかに (1) 照応詞, (2) 否定表現, (3) 一人称代名詞, (4) 三人称代名詞, (5) 副詞, (6) その他の 6 つの種類に分類される語により示されていると考え, これらの 6 つの分類を考慮し, 手がかり語を用いて参照個所の抽出を試みた.

手がかり語は以下の手順で得た. まず, 論文データから人手で参照個所を 200 個所抽出した. 次に抽出した参照個所の n-word gram 統計をとり, 結果を人手で分類・整理した. その結果, 文間のつながりには先に示した 6 種類あることがわかった. これらの 6 種類のつながりを考慮し, 最終的に人手で 86 個の手がかり語を選んだ. なお, n-word gram 統計をとる際, 大文字, 小文字の区別をしている. 表 3.1に手がかり語を示す.

次に, 手がかり語を用いた参照個所抽出の手順を図 3.3に示す. 入力は, 予め参照の含まれる段落を 1 行 1 文の形に直し, 配列 (paragraph) に入れておき, ルールを用いて参照個所抽出を行う. 参照個所抽出ルールとは,「参照個所候補となる文の前後の文に手がかり語が出現すれば, その文も参照個所候補に含める」といったものである. 参照個所抽出ルールを図 3.4に示す.

図 3.4において, “FIRST SENTENCE” とは参照個所候補の最初の一文の文番号, “LAST SENTENCE” は最後の一文の文番号を意味する. また, “this.cue”, “but.cue”, “we.cue”, “they.cue”, “and.cue” はそれぞれ, 図 3.3に示す手がかり語の “(1) 照応詞”, “(2) 否定表現”,

表 3.1: 参照個所抽出用手がかり語

(1) 照応詞	For this, For these, On this, On these, In this, in this in these, In these, This, These, Therefore
(2) 否定表現	yet, less, but, in spite of, unlike, rarely in contrast, although, Still, Nevertheless, instead, despite, irrelevant, has not been, not attempt to not possible to, this is not, but is not, less, has not, have not
(3) 一人称	I, in our example, our analysis was, our analysis of by using our, in our work, our analysis is, to our concept, our analysis, our work our example, using our, we
(4) 三人称	they, their, them, he, his, him, she her, hers
(5) 副詞	And, Furthermore, Because, Again, Additionally, Such, In such, So
(6) その他	difference between, defect, drawback, impossible, Using, we incorporate, in the implementation, is implemented, first, second, theory, theoretical, origin, based, base, basis, adopt, apply, applied, foundation, fundamental, radical, element, underlie, underlay, underlain, underlying, In particular, follow

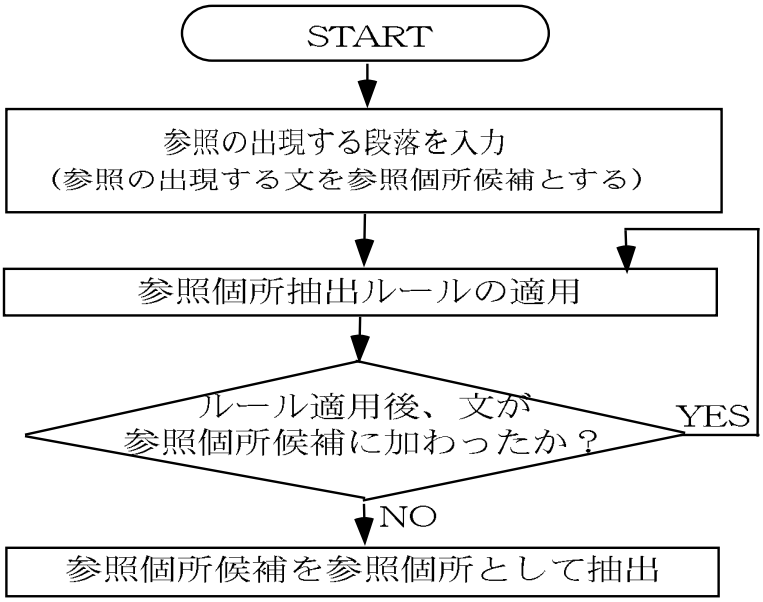


図 3.3: 参照個所抽出の手順

1 FIRST SENTENCE が this.cue で始まる場合, 前の文も抽出する.

2 FIRST SENTENCE が but.cue で始まる場合, 前の文も抽出する.

3 FIRST SENTENCE が and.cue で始まる場合, 前の文も抽出する.

4 LAST SENTENCE の次の文が but.cue で始まる場合、次の文も抽出する.

5 LAST SENTENCE の次の次の文が but.cue で始まる場合, 次の次の文まで抽出する.

6 LAST SENTENCE に we.cue が含まれなくて, 次の文に we.cue が含まれる場合, 次の文も抽出する.

7 LAST SENTENCE に we.cue が含まれなくて, 次の次の文に we が含まれる場合, 次の次の文まで抽出する.

8 LAST SENTENCE に we.cue が含まれなくて, 次の文に大文字のみシステム名が含まれる場合次の文も抽出する.

9 LAST SENTENCE の次の文が and.cue で始まる場合, 次の文も抽出する.

10 LAST SENTENCE の次の文に they.cue が含まれる時, 次の文も抽出する.

11 LAST SENTENCE の次の文に this.cue が含まれる場合, 次の文も抽出する.

図 3.4: 参照個所抽出ルール

“(3) 一人称”, “(4) 三人称”, “(5) 副詞と (6) その他” の項目に対応している. これら 11 種類のルールを用いて, 参照個所の抽出を試みた. 一方, これらの 11 種類のルールの中には参照個所抽出精度低下の原因となるルールも含まれる可能性が考えられる. 従って, 11 種類のルールの組み合わせ 2^{11} 通りの中で最も精度が高くなる場合が, ルールの最適な組み合わせであると考えられる. ルールの最適な組み合わせの学習方法およびその結果については 3.4.1 節で述べる

3.3 参照タイプの決定

参照個所中で, 例えば参照の後の文が “However” で始まるような場合, 参照論文の著者は被参照論文の何らかの問題点を指摘している (type C) と考えられる. また, 参照の前に “We use” や “We adopt” といった語が出現する場合, 参照論文は被参照論文の理論や手法等をベースにしている (type B) と思われる. 従って, 参照タイプ決定には, まず “However” や “We adopt” といった, 参照タイプ決定のための手がかり語リストを作成し, 次に手がかり語と参照の出現順序を考慮したルールを作成することが必要であると考えられる.

まず, 手がかり語の抽出方法について述べる. 学術論文には, 論文特有の構造がある. Biber らは, 医学論文において “Introduction”, “Methods”, “Discussion”, “Results” の 4 つの section で使われる言語の特徴を調査し, 4 つの section 間の言語的な特徴の違いを明らかにしている [2]. 本研究では参照タイプ毎にこのような section に注目した. type C の場合, 論文中の “Introduction”, “Related Work”, “Discussion” に注目した. また, type B については, “Introduction”, “Experiment” の section に注目した. E-Print archive の論文 395 本から section 毎に n-word gram をとり, 次に cost criteria [22] を利用することで手がかり語の候補のリストを自動的に作成した. n-word gram 統計をとる際, 大文字と小文字の区別を行った. また, カンマやピリオドも一語として取り扱った. こうして得られたリストから, 参照タイプ決定に有用であると思われるものを, type C 用に 76 個, type B 用に 84 個を, 手がかり語 として選びだした. 手がかり語を表 3.2, 表 3.3に示す.

表 3.2: type C 決定用手がかり語

although the	, although	Though,	however, ... our
however, ... they	recently ... however	, however	however, ... not
However,	however, the	however, that	the only
But	but a	but the	but it
but is	but are	but rather	but no
but they	but their	but he	but his
but she	but her	but it	but instead
Instead,	In spite of	does not	did not
that is not	not be	it is not	this is not
was not	were not	it does not	may not
might not	will not	would not	wouldn't
should not	can not be	could not	(citation) ... can not
need not	not always	not have	have not
is too	has not	little influence	that do not
they do not	he does not	she does not	not require
not provide	not cover	not in effect	more efficient than ... (citation)
not enough	less studied	difference from	different from
more difficult	a difficult	difference between	

表 3.3: type B 決定用手がかり語

based mainly on	based on ... in	is based on	are based on
, based on	this ... based on	employ	invoke
Based on	the basic	underlie	underlain
assume	widely used	has used	used by
use ... to	can use	used as a	by using
Using the	is checked	we use	We will use
we used	result	make use of	made use of
Making use of	advantage of	we describe	is described in
are described in	accord with	accorded with	correspond to
mentioned	benefit	beneficial	we introduce
And	We argue	In such	we present
support	is given in	are given in	offer
For ... reason	we ... influence	assume	is needed to
are needed to	been given	a given	given the
follows	following	we believe	implementation
we investigate	to consider	which can be	the possible
available for	apply	applied to	application to
adopt	We adopted	extend the	we extended
we consider	extended to	expands	expanded
refer ... to	referred to	This ... importance	This ... important
implement			

```
sub reference_type_decision($@){ # 参照タイプ決定ルーチン
  ($citeline,@ra)=@_; # $citeline : 参照の位置
                      # @ra       : 参照個所, 1 行 1 文のリスト
# type C 決定用ルール
  for($i=1;$i<=3;$i++){if($ra[$citeline+$i]=~/However/]){return(C)}}
  for($i=0;$i<=2;$i++){if($ra[$citeline+$i]=~/ less studied/]){return(C)}}
  for($i=0;$i<=2;$i++){if($ra[$citeline+$i]=~/In spite of/]){return(C)}}
  …
# type B 決定用ルール
  for($i= -2;$i<=0;$i++){if($ra[$citeline+$i]=~/ based mainly on/])
                                                        {return(B)}}
  for($i= -3;$i<=0;$i++){if($ra[$citeline+$i]=~/ apply to /]){return(B)}}
  for($i= -2;$i<=0;$i++){if($ra[$citeline+$i]=~/Using the/]){return(B)}}
  …
# B, C に割り振られなかったものは type O
  return(O);
}
```

図 3.5: 参照タイプ決定ルーチンの一部

次に参照タイプ決定ルールについて説明する. 参照タイプの決定は, 表 3.3, 表 3.2に示す手がかり語を用いてルールを作成した. 参照タイプ決定には, 本節の冒頭でも述べたような参照と手がかり語の出現順序を考慮することが有用であると考えられ, この情報を用いたルールを作成した.

また, 2.2 節で述べたように一つの参照個所に複数の参照タイプを割り振ることが可能な場合もあるが, 今回は一つの参照個所に一つの参照タイプを割り振るルールを作成した.

ルールは大きく 2 種類に分けることができる. ひとつは type C に決定するためのルール, もうひとつは type B に決定するためのルールである. そして, B, C どちらのタイプも割り振られなかった参照個所を type O とする. ルールは各手がかり語毎に作成されているため, type C 決定用ルールは 76 個, type B 決定用ルールは 84 個ある. これらのルールの適用順序について説明する. type C 決定用ルールは 76 個の順序を入れ換えても参照タイプ決定精度には影響がない. type B 用ルール 84 個についても同様である. そこで, type C 用ルール, type B 用ルールの順に適用した後に type O を割り振った場合と, type B 用

ルール, type C 用ルールの順に適用した後に type O を割り振った場合について調べた. その結果, 先に type C 用ルールを用いた方が解析精度が高くなったので, type C 用, type B 用ルールの順に適用した後, 参照タイプがどちらにも割り振られなかったものを type O とした. 参照タイプ決定ルーチンの一部を図 3.5に示す. 参照タイプ決定ルーチンでは, 1 行 1 文に整形された参照個所を配列として, また配列中の参照の位置を入力値として受け取り, 参照タイプ C, B, O を値として返す.

3.4 実験

3.4.1 参照個所の抽出

前章で述べた手法の有効性を評価するため, 参照個所抽出実験を行った. 評価は式 (3.1)(b=1) に示す F-measure[68] を用いて行う.

$$F(F\text{-}measure) = \frac{(1+b^2)PR}{b^2P+R} \tag{3.1}$$

ここで, P, R は以下により計算される.

$$R(\text{再現率}) = \frac{\text{抽出された文のうち正解のものの数}}{\text{参照個所コーパスの抽出すべき文の総数}} \tag{3.2}$$

$$P(\text{精度}) = \frac{\text{抽出された文のうち正解のものの数}}{\left(\begin{array}{l}\text{参照個所抽出ルールにより}\\ \text{抽出された文の総数}\end{array}\right)} \tag{3.3}$$

実験用データとして, 参照の含まれる段落を 1 行 1 文に整形したものと, 段落中の何文目から何文目までが参照個所かを記したものを 150 個用意した. 段落の切れ目は話題の切れ目と考え, 参照個所は最大でも参照の含まれる段落全体までとした. 150 データのうち 100 個を訓練用, 50 個を評価用として, 3 回繰り返し (3 分割 Cross Validation), 参照個所抽出用の 11 種類のルールの最適な組み合わせを得た. また, ルール作成用データを用いて, この組み合わせで評価用データに対して実験を行った.

また, 本手法の有効性を示すために, 3 つのベースラインを考慮した. 参照の含まれる文のみを参照個所として抽出した場合, その文は必ず参照個所である (ベースライン 1). 一方, 参照のある段落全体を参照個所として抽出した場合, 参照個所として抽出されうる文はす

表 3.4: 参照個所抽出精度 (3 分割 Cross Validation)

		再現率 (%)	精度 (%)	F-measure
セ	本手法 (訓練用)	78.0 (174/223)	87.0 (174/200)	0.823
ッ	本手法 (評価用)	75.5 (80/106)	83.3 (80/96)	0.792
ト	ベースライン 1 (参照を含む文)	54.7 (58/106)	100.0 (58/58)	0.707
1	ベースライン 2 (段落全体)	100.0 (106/106)	13.1 (106/807)	0.232
	ベースライン 3 (参照の文＋前後)	88.7 (94/106)	59.1(94/159)	0.709
セ	本手法 (訓練用)	81.0 (171/211)	87.7 (171/195)	0.842
ッ	本手法 (評価用)	66.9 (79/118)	88.7 (79/89)	0.763
ト	ベースライン 1 (参照を含む文)	44.1 (52/118)	100.0 (52/52)	0.612
2	ベースライン 2 (段落全体)	100.0 (118/118)	11.2 (118/1049)	0.202
	ベースライン 3 (参照の文＋前後)	79.7 (94/118)	68.6 (94/137)	0.737
セ	本手法 (訓練用)	78.1 (175/224)	86.6 (175/202)	0.822
ッ	本手法 (評価用)	81.0 (85/105)	79.4 (85/107)	0.802
ト	ベースライン 1 (参照を含む文)	53.3 (56/105)	100.0 (56/56)	0.696
3	ベースライン 2 (段落全体)	100.0 (105/105)	11.1 (105/948)	0.199
	ベースライン 3 (参照の文＋前後)	86.7 (91/105)	63.2 (91/144)	0.731

表 3.5: 訓練データで選択された参照個所抽出ルール

	1	2	3	4	5	6	7	8	9	10	11
セット 1	*	*	*	*	*	*	*		*	*	
セット 2	*	*	*	*	*	*	*		*		*
セット 3	*	*	*	*			*		*	*	*

(各番号に対応する参照個所抽出ルールは図 3.4参照)

べて含まれてしまう (ベースライン 2). この他のベースラインとして, 参照の含まれる文とその前後の 1 文を参照個所として抽出した (ベースライン 3).

結果を表 3.4に示す. 表 3.4において, 本手法の F-measure 値はどのセットにおいても 3 つのベースラインの値を上回っており, 従って参照個所抽出手法の有用性が示されたと言える.

表 3.5は, セット毎に学習された参照個所抽出ルールの組み合わせである. ルール 8「LAST SENTENCE に we.cue が含まれなくて, 次の文に大文字のみのシステム名が含まれる場合次の文も抽出する.」はどのセットにおいても選択されていない. 従って, 4 章以降で述べる研究では, ルール 8 以外の 10 ルールを用いて参照個所の抽出を行う.

参照個所の抽出に失敗した理由は大きく以下の 2 種類に分類できる.

原因 1 照応・省略解析に関連した失敗

原因 2 語彙的連鎖 (同語反復) に関連した失敗

参照個所抽出の失敗例を図 3.6に示す. (例 1) は, 原因 1 に関する参照個所抽出の失敗例である. (例 1) では, 文 2 で [Macgregor, 1998] を参照しており, また文 1, 2 が [Macgregor, 1998] に関する参照個所であるが, 参照個所抽出ルールは文 2 しか抽出できていない. この例では文 2 の照応詞 “It” の先行詞は文 1 の “the task model” である. 一方, 参照個所の抽出には, “It” は考慮されていない. 何故ならば, “It” には, 照応詞の他にも形式主語, 形式目的語 (強調構文) 等, 様々な用法があり, 単純に “It” で始まる文の前文を参照個所として抽出すると, 不必要に多くの文を抽出してしまう可能性がある. 従って, “It” が照応詞であるかどうかの判定は必要不可欠である.

(例 2) においても同様の失敗例が見られる. (例 2) では, 文 1 で [Melamed, 1996] を参照しており, 文 1, 2 が [Melamed, 1996] に関する参照個所であるが, 参照個所抽出ルールでは文 1 しか抽出されていない. 文 2 の “The translation lexicon” の先行詞は, 文 1 の “plateus” であると考えられるが, 定冠詞 “the” は, 既出の名詞を指す用法の他にも, 慣用的に用いられる場合, 抽象名詞化する場合など様々な用法があり, “It” と同様, 照応的な用法であるかどうか判定する必要がある.

これらの 2 例から, 文間のつながりを計る手がかりとして, ある程度の照応解析 (“it” や “the” 等の語が照応詞であるか否かの判定) は必要である.

(例 3) は, 原因 2 に関する参照個所抽出の失敗例である. (例 3) は, 文 1 から文 4 が [Guth, 1992](文 2) に関する参照個所であるが, 参照個所抽出ルールでは文 1 と文 2 しか抽出され

(例 1) (原因 1: 照応・省略解析に関連した失敗)

1. The knowledge base supports the construction of the task model discussed above.
2. It is an hierarchical structure implemented in loom [**Macgregor, 1988**].

(例 2) (原因 1: 照応・省略解析に関連した失敗)

1. As Melamed [**Melamed, 1996**] observes, SABLE's output groups naturally according to "plateaus" of likelihood (see Figure 1).
2. The translation lexicon obtained by running SABLE on the Answerbooks contained 6663 French-English content-word entries on the 2nd plateau or higher, including 5464 on the 3rd plateau or higher.

(例 3) (原因 2: 語彙的連鎖 (同語反復) に関連した失敗)

1. Method has recently been optimised by Cowie and Guthrie [**Guth, 1992**] using simulated annealing and they report results of 72% correct assignment at the homographic level in LDOCE and a much lower level for individual sense assignment.
2. This result must be seen against a background figure of 62% correct sense assignment in LDOCE achieved by assigning the first LDOCE sense in an entry.
3. However, we suspect that the wrong optimisation function was used in the annealing, one that tended to assign senses with long definitions in LDOCE, and so the figure could have been much better, a matter we intend to remedy later.
4. The importance of this method is that it disambiguates all the content words in a sentence, even though it involved a vast computation for a sentence if all the LDOCE senses were considered, often optimising more than 10^9 sense combinations for a 12 word sentence.

図 3.6: 参照個所抽出の失敗例

ていない. この例において, “LDOCE” という語の反復が文間のつながりを示していると考えられるが, 参照個所抽出ルールでは同語反復 (語彙的連鎖) による文間のつながりは考慮していないため, 文 3 と文 4 が抽出されていない. 従って, 参照個所を抽出する際, 今後は同語反復 (語彙的連鎖) も考慮する必要がある. しかし, 論文の主題 (テーマ) を表す語は, 論文全体にわたって高頻度で出現するため, このような語の反復を考慮すると, 参照個所が過剰に抽出される可能性がある. 従って, まず参照の前後で, 局所的な話題 (焦点) を示す語を見つけ, 次にこのような語の反復を参照個所の抽出に用いるといった戦略が必要であると思われる.

3.4.2 参照タイプの決定

参照タイプ決定実験の評価方法も参照個所抽出と同様, 再現率, 精度を用いた. 式 (3.4)(3.5) は type C のタイプ決定精度の評価方法である.

$$\text{再現率} = \frac{\left(\begin{array}{l}\text{ルールを用いて } type\ C \text{ に決定された}\\ \text{参照個所のうち正解の数}\end{array}\right)}{\text{参照個所コーパス中の } type\ C \text{ 参照の数}} \tag{3.4}$$

$$\text{精度} = \frac{\left(\begin{array}{l}\text{ルールを用いて } type\ C \text{ に決定された}\\ \text{参照個所のうち正解の数}\end{array}\right)}{\text{ルールを用いて } type\ C \text{ に決定された参照個所の数}} \tag{3.5}$$

実験用データとして, 参照個所とそのタイプを人手で決定したものを 382 個用意し, そのうち 282 個をルール作成用, 残り 100 個を評価用とした. また, 正解データ作成の際, 一つの参照個所には一つの参照タイプを割り振った. 2.2 節で挙げた例のように, type B とも type C とも考えることのできる参照個所については, 「既存の理論を全く変更, 修正することなく利用するのであれば type B であるが, 既存の理論に何らかの問題点を見いだし, その一部を修正して用いるのであれば type C」 という基準に基づいて判断した.

ルール作成用データにおけるタイプ決定精度を表 3.6に, 評価用データにおけるタイプ決定精度を表 3.7に示す.

タイプ決定精度について考察する. 手がかり語選定の際, uni-gram は極力排除した. それは uni-gram が参照タイプ決定の精度を低下させる要因になっていたためである. 例えば以

表 3.6: ルール作成用データを用いた参照タイプ決定精度 (282)

		ルールで決定されたタイプ			タイプ毎の精度 (%)
		C	B	O	
正解のタイプ	C	**46**	2	1	93.9 (46/49)
	B	1	**105**	13	88.2 (105/119)
	O	3	8	**103**	90.3 (103/114)
					90.1 (254/282)

表 3.7: 評価用データを用いた参照タイプ決定精度 (100)

		ルールで決定されたタイプ			タイプ毎の精度 (%)
		C	B	O	
正解のタイプ	C	**12**	0	4	75.0 (12/16)
	B	2	**25**	5	78.1 (25/32)
	O	1	5	**46**	88.5 (46/52)
					83.0 (83/100)

(例 1) (原因 1: 手がかり語と論文の参照の順序がルールと異なる場合の失敗)

1. The interpretation of 'ex0b' that we would predict as possible would be the *Ziroo dislikes Taroo* (RETAIN) which native speakers rarely get.
2. **However** Kuno's analysis does not block this reading either; the zero in 'ex0b' could also be a REAL-ZERO-PRONOUN, with *Taroo* as its antecedent.
3. Kuno says that this interpretation is dispreferred because of a preference for parallel interpretation [**Kuno, 1989**].

(例 2) (原因 2: 参照個所中に手がかり語の出現しない場合の失敗)

1. Thirdly, our method is a generalization of the strategy employed by [**McCord, 1993**].
2. Our comparisons are more global and therefore can result in more effective pruning.

図 3.7: 参照タイプ決定の失敗例

前の研究 [41] では "not" や "but" といった語を手がかり語として用いていたが, "not only ... but also" のように "not" や "but" が否定以外の目的で使われているものもある. 今回は例えば "not" に関する手がかり語では, "can not", "could not", "might not" といった bi-gram をタイプ決定に利用している. これにより, 以前の解析精度 (約 66%) を大幅に改善することができた.

参照タイプの決定に失敗したものをいくつか調べた結果, 大きく次の 2 種類の要因があることが分かった.

原因 1 手がかり語と論文の参照の順序がルールと異なる場合の失敗

原因 2 参照個所中に手がかり語の出現しない場合の失敗

図 3.7(例 1) は, 原因 1 の失敗例である. 例 1 において, 参照論文の著者は文 2 で先行研究 [Kuno, 1989] の問題点の指摘をしている. 続く文 3 で, Kuno の研究内容について述べている. 参照タイプ決定ルールは図 3.5に示すように, 参照の後に "However" のような否定表現が出現すれば type C と決定するため, 例 1 のように "However" が参照より前に出現する場合は失敗する. 実際, 評価用データにおけるいくつかの事例において, 参照個所中に手

がかり語が出現しているにもかかわらず, 論文の参照と手がかり語の順序が参照タイプ決定ルールと異なったため失敗した. これらの失敗事例のいくつかは, 被参照論文が参照論文中で複数回参照されていた. [Kuno, 1989] もその一例であるが, (例 1) に示す参照より以前に, 論文中で [Kuno, 1989] が参照されており, その参照個所ですでに [Kuno, 1989] の内容について述べられている. このような場合, 論文の読み手はすでに [Kuno, 1989] の内容を把握しているため, 例 1 のように参照個所中でいきなり問題点を指摘しても, 理解できる. これらの失敗事例を処理するためには, 個々の参照個所だけに着目するのではなく, 論文全体の構造の中で参照個所を捉え, その上で参照タイプを決定する必要がある. より具体的には, 論文が複数回参照されている場合の参照タイプの決定は, 2 回目以降の参照に関しては, 参照個所中に type B, type C 決定用の手がかり語が出現すれば, 参照と手がかり語の順序を考慮せず, type B あるいは type C に決めるといった方法が考えられる.

図 3.7(例 2) は, 原因 2 の失敗例である. 例 2 において, 参照論文の著者は, 「(文 1) 提案手法は [McCord, 1993] の方法よりも一般的である. (文 2) 従って, 提案手法により効果的な枝刈りが可能になる.」と述べている. これは, 「[McCord, 1993] の手法は一般性に欠けている」という点で問題点を指摘していると考えることができ, 評価用データ作成の際には type C と判断した. 一方, 参照タイプ決定ルールが type C の判定に用いる手がかり語は, この個所には含まれていないため, ルールでは type O と判定された. この個所の中では, “generalization” や “more global” が type C と判定するキーワードになっている. しかし, これらは手がかり語と呼べるほど一般的な表現であるとは考えにくい.

このように, 手がかり語だけでは “type B” や “type C” と判別できない例がいくつかあり, これ以上のタイプ決定精度の向上には意味処理等を行う必要がある.

3.5 関連研究

参照個所の抽出に関する研究

三平らは [33] において, 日本語論文を対象に論文中の参照の出現する文の前後から著者の主題表現を抽出している. 主題表現の抽出は, 本研究における参照個所の抽出方法と同様に手がかり語に基づいている. 参照の出現する文の前後の文に手がかり語が出現すれば, その個所を主題表現として出力する. しかし抽出手法の定量的な評価については行っていない.

参照タイプの決定に関する研究

神門は, 手がかり語に基づくいくつかのルールにより, 論文中の各文に構成要素カテゴリの自動付与を行っている [18]. このカテゴリを用いることで, 参照が含まれる文に割り振られたカテゴリ毎に, 参照を分類することが可能になる. 例えば, 「A12. 既存の研究」や「B46. 測定法の妥当性」というカテゴリの文に含まれる参照は, それぞれ「基本的な研究動向を示すための参照」, 「方法の妥当性を示すための参照」と考えることができる.

Teufel も, 神門と同様, 論文中の各文に argumentative zone と呼ぶ一種の構成要素カテゴリの自動付与を行っている [61]. Teufel は論文の構成要素を 7 種類に分類しているが, このうち論文の参照の理由に関しては “CONTRAST”, “BASIS”, “OTHER” の 3 種類に分けている. これらは本研究における参照タイプ C, B, O に対応している.

Teufel は, argumentative zone を自動判定する際, 手がかり語の他に, 論文中の各文の位置情報, 段落内での各文の位置情報, 文中の参照の位置 (Beginning, Middle, End), 文の長さ, 文中の重要語 (tf*idf, 論文表題中の語) の有無等の特徴に着目し, これらの特徴を用いて機械学習により argumentative zone 自動付与のルールを獲得している. その結果, argumentative zone 全体では 70%程度の再現率と精度が得られている. 一方, “CONTRAST”, “BASIS”, “OTHER” に関しては, 再現率, 精度共に 50% をほぼ下回っている.

Teufel も本研究と同様に E-Print archive の論文データベースを用いている. また, 評価用データおよび実験条件に違いはあるものの, 本研究における参照タイプ決定精度 (83%) と比べると, 大きな開きがある. この原因の一つは, “CONTRAST”, “BASIS”, “OTHER” は, 他の argumentative zone と比べ論文中の出現頻度が低く, 学習用コーパスの中でこれらの事例がノイズとして取り扱われてしまっている可能性がある[1]. 従って, 低頻度の事例が不利にならないような機械学習 (例えば [16, 30]) の方法をとれば, “CONTRAST”, “BASIS”, “OTHER” の分類精度は改善されると考えられる.

3.6 まとめ

本章では, 2 章で定義した参照情報を自動的に抽出する手法を提案した. 本研究では, E-Print archive という論文データベースの TEX ファイルを対象に参照情報の抽出を試みた.

[1]逆に, 一番頻度の高い “OWN”(全体の 2/3) では再現率 91%, 精度 81% という高い分類精度が得られており, “OWN” の分類に関するルールが多く獲得されていると推測される.

参照情報の抽出は, (1) 論文間の参照・被参照関係の解析, (2) 参照個所の抽出, (3) 参照タイプの決定という 3 つのステップに分けられる. (1) は, TEX の bibliography というコマンドに着目して自動的に解析を行い, 94% の精度が得られた. (2)(3) は汎用性を考慮し, 手がかり語を用いた手法を提案した. (2) については, 6 種類の文間のつながりを考慮しこれらのつながりを表す手がかり語を考慮した参照個所抽出方法を提案した. その結果, 評価用データにおいて F-measure 約 0.80 の値が得られた. また, (3) については, 手がかり語を用いて参照個所を解析することで, 参照タイプを自動決定する手法を提案した. 実験の結果評価用データで 83% の解析精度が得られた. 表層的な情報を用いた参照個所の抽出や参照タイプの決定の精度としてはほぼ限界に達していると考えられ, 今後, 参照情報の抽出精度を向上させるためには, より深い意味処理が必要不可欠であると考えられる.

3.7 今後の課題

3.4.1 節, 参照タイプ決定実験で述べたとおり, 本研究では, 一つの参照個所には一つの参照タイプのみを割り振った. しかし 2.2 節で示した例では type B と type C の両方の参照タイプを割り振るという考え方の方が, 実際には, より自然であると考えられる.

図 3.5において, 現在の参照タイプ決定ルール サブルーチンでは, まず参照個所が type C の手がかり語を含んでいるか調べ, 含んでいれば “C” という値を返して処理を終了してしまう. ここで処理を終了せず, 続けて type B の手がかり語を含んでいるかについても調べ, もし含んでいれば “B” と “C” の両方の値を返すことで, 上で述べた問題点に対処することができる.

type C の参照個所から得られる情報は, 2 章でも述べたように, 以下のとおりである.

(α) 既存研究の紹介

(β) 既存研究の問題点

(γ) 参照論文の研究の目的

本章で提案した参照個所の抽出方法は, 上記に示すような参照個所中の要素毎の抽出を行ってはいない. しかし, 参照個所中からさらにこのような要素毎の情報を抽出することは, 有用であると考えられる. 例えば, 三平は参照個所中の語を用いて, 参照論文の著者の主題表現と関連のある論文の検索を行っているが [33], 参照個所中の各要素毎に単語の重みを変

えることで,「参照論文の著者と共通の問題意識を持った論文の検索」や「参照論文の著者と問題の解決方法が類似した論文の検索」など, より詳細な目的に応じた検索が可能になると考えられる.

また, 論文の参照は, 2 章の関連研究でも述べたように, ある論文を一論文中で複数回参照する場合もあれば, 一個所で複数の論文を同時に参照する場合もある. また, 一般的には, 前者はより被参照論文の詳細な内容に関する記述があり, 後者はより一般的な記述がなされていると考えられる. 現在は, このような参照の形式的な違いを特に考慮はしていないが, 参照個所抽出の精度をさらに向上させるためには, こうした参照方法の形式的な違いと実際の参照個所中の記述との関係の分析が有用であると考えられる. さらにこのような分析は, 3.4.2 節でも述べたとおり, 参照タイプ決定ルールの作成にも関連する可能性がある.

第 4 章

参照情報を考慮したサーベイ論文の作成支援

本章では, 参照情報を利用したサーベイ論文作成支援の方法について述べる. まず, 4.1 節ではサーベイ論文の作成支援の意義について述べる. 4.2 節では, サーベイ論文を作成する上でのポイントを説明する. 4.3 節では, サーベイ論文作成に関する関連研究をいくつか紹介する. また, 4.4 節では, サーベイ論文の作成支援における参照情報の利用方法ついて説明し, 4.5 節で, サーベイ論文作成支援システムを示す.

4.1 サーベイ論文作成支援

近年, 研究者数の増加, 学問分野の専門分化と共に学術情報量が爆発的に増加している. また, 研究者が入手できる文献の量も増える一方であり, 人間の処理能力の限界から, 入手した文献全てに目を通し利用することが益々困難になってきている.

このような状況で必要とされるのは, 特定の研究分野に関連した情報が整理, 統合された文書, すなわちサーベイ論文 (レビュー) や専門図書である. サーベイ論文や専門図書を利用することで, 特定分野の研究動向を短時間で把握することが可能になる. しかし, 論文全体に対するサーベイ論文の占める割合が極端に少ないという指摘がある [10]. その理由の一つとして, サーベイ論文を作成するという作業がサーベイ論文の作者にとって, 時間的にも労力的にも非常にコストを要することが挙げられる. しかし, 今後の学術情報量の増加を考えれば, このようなサーベイ論文の需要は益々高まっていくものと思われる.

本研究ではサーベイ論文を複数論文の要約と捉えている. 本来サーベイ論文とは, 多くの論文に提示されている事実や発見を総合化, また問題点を明らかにし, 今後更に研究を要する部分を提示したものであると考えられる [10]. しかし現在の自動要約の技術から考えると, このようなサーベイ論文の自動作成は, 非常に困難であると思われる. そこで関連する複数の論文中から各論文の重要個所, 論文間の相違点が明示されている個所を抽出し, それらを部分的に言い替えて読みやすく直した後, 並べた文書をサーベイ論文と考え, そのような文書の自動作成を試みる.

本章では, その第 1 歩として, サーベイ論文作成を支援するシステムを示す. 本研究では, サーベイ論文作成支援の際, 論文間の**参照情報**に着目する. 一般に, ある論文は他の複数の論文と参照関係にあり, また論文中に被参照論文の重要個所や, 被参照論文との関係を記述した個所 (**参照個所**) がある. この参照個所を読むことで, 著者がどのような目的で論文を参照したのか (**参照タイプ**) や参照・被参照論文間の相違点が理解できる. 参照情報は特定分野の論文の自動収集や論文間の関係の分析に利用できると考えられる.

4.2 サーベイ論文作成のポイント

これまで, 単一論文の要約に関して, 論文中の重要個所を抽出する数多くの手法が提案されてきた (例えば [9, 49, 24, 61, 30]). しかし, 要約対象が複数論文の場合, 単一論文の要約とは別に考慮すべき点が出てくる. まず, 要約対象となる複数の論文をどのように収集するのか. また, 収集してきたテキスト間で内容が重複する場合, 従来の単一論文要約の手法を個々の論文に適用し並べただけでは, 個々の要約の記述が重複する可能性があり, 冗長で要約として適切ではない. そのため, 冗長な個所 (論文間の共通個所) をどのように検出し削除するかが問題となる. 一方, 冗長な個所を削除しても複数論文の要約文書としてはまだ十分であるとは言えない. 複数論文を要約するとは, それらの論文を比較し要点をまとめることであり, そのためには論文間の共通点だけでなく相違点も明らかにすることが必要であると考えられる. さらに, 共通点や相違点の情報に基づいて, 論文を分類・整理する必要がある. また, 要約文書を作成するためには, 検出された論文間の共通点や相違点を並べ, 使用する単語の統一, 接続詞の付与, “we”, “they”, “in this paper” といった照応詞の著者名への置換等, テキストを読みやすくするための処理 [42] が必要となる. 従って, 複数論文要約のポイントは図 4.1 のようにまとめることができる.

(a) 関連論文の自動収集

(b) 論文間の関係の解析

(b)-1 関連する複数論文からの情報の抽出

(b)-1-1 重要個所の抽出

(b)-1-2 論文間の共通点の検出

(b)-1-3 論文間の相違点の検出

(b)-2 収集した論文の分類・整理

(c) 論文の著者毎の文体の違い等を考慮した要約文書の生成

図 4.1: 複数論文要約のポイント

4.3 関連研究

4.3.1 サーベイ論文の自動作成に関する研究

神門は, 手がかり語を用いて論文中の各文に構成要素カテゴリの自動付与を行い, そのカテゴリを論文検索に応用している [19](図 4.1(a)). このようにして収集された特定分野の論文集合の「既存の研究」や「既存の研究の不完全さ」カテゴリの文を抽出し, それらを並べて表示することで, その分野の基本的な動向を把握するのに有用であると述べている (図 4.1(b)). 神門は, このようなカテゴリの文が「当該論文の著者の判断を通してみた, その課題に関する現状や背景を示している」と考えている. 本研究でもこのような著者の主観的な判断をサーベイ論文作成の際に利用している.

対象テキストが学術論文とは異なるが, 複数の新聞記事を対象に複数記事要約を行う試みがいくつかある [45, 46]. 要約対象が新聞記事の場合, 次のような特徴がある.

- 新聞記事は, 記事中の事実文が重要であると考えられることが多い. 従って, 客観的な正解データが作成しやすいと思われる.
- 図 4.1(c) に関して, 新聞記事では文体がある程度統一されているため, 記事間の文体

の違いをあまり意識する必要がない.

一般に, 論文には著者毎の文体の違いが存在し, しかも新聞記事を要約対象とした場合と比べてその違いが大きいため, 論文間の共通点の検出には新聞記事の場合のような各文中の個々の形態素の比較といった手法が適用しにくい. また, 論文は著者毎に異なる観点で書かれているため, 複数論文をまとめるにはどのような観点でまとめるのかが重要なポイントとなる. 本研究では, このような著者毎の観点の違いに着目している.

4.3.2 サーベイ論文の分析に関する研究

これまでに, 人間の書いたサーベイ論文を分析した研究がいくつかある [37, 64, 65, 26, 36].

一般にサーベイ論文は「ある主題の範囲の情報, 知識を統合するために, その領域の原著論文群を一定の形式の中に凝縮表現したもの」と言える. これらの情報や知識がサーベイ論文中でどのように凝縮されているかは, 同じ分野の複数のサーベイ論文を比較することで, ある程度明らかにできると考えられる. このような考え方に基づき, 村主らは, 「臨床医の情報ニーズ・情報探索行動」という分野において 7 人の研究者が 1982 年–1993 年の間に記した 7 本のサーベイ論文を対象に調査している [37]. 調査の際, 多くのサーベイ論文から参照されているその分野の代表的な論文 (以後, スター論文 [65]) に着目している.

その結果, 同じスター論文でも年と共に研究が細分化していく中で参照のされ方が変わってくるといった興味深い現象も確認されたが[1], 概して, サーベイ論文で参照する論文集合や個々の論文の参照の仕方は著者の主観に依存する部分が多く, 「サーベイ論文中の組織的な知識の蓄積・凝縮」は, 少なくともこの調査結果からは確認されていない.

このようなサーベイ論文の著者毎にばらつきが生じるのは, サーベイ論文を書くための客観的な手順が確立していないためであると考えられる. その重要性は Light ら [26], 津田ら [64], Mulrow ら [36] により指摘されており, またサーベイ論文を書くためのいくつかの手順 [26, 12] や指針 [12], 評価方法 [36, 47] などが提案されている.

サーベイ論文を書くためには, 個々の論文を読み, 論文中の情報を統合する必要があるが, このような作業を客観的に行うための方法の一例として Goldscmidt のもの [12] を図 4.2に挙げる. また, 評価方法の例として Mulrow らのものを図 4.3に挙げる.

[1]過去のサーベイ論文中では 1 度しか参照されなかったスター論文が, より新しいサーベイ論文中では複数回参照される場合などがある. これは, その分野における問題意識の変化に伴い, 同じ論文でも様々な角度から検討されることがあるからである.

(Goldscmidt, 1986) [12]((津田, 1994)[64] より抜粋)

(1) 統合すべき情報が対象としている問題と, その問題に適合する情報を定義

(2) 統合する情報の収集

(3) それらの情報の正確さの評価

(4) 目的とする情報が, 標的としている利用者達に役立つように, 正確である事を確認した上で提供

図 4.2: 情報の統合を客観的に行う方法

(Mulrow, 1987) [36]((津田, 1994)[64] より抜粋)

(1) はっきりした目的が述べられているか.

(2) 収録文献を見つけた方法や, その情報源が明らかにされているか.

(3) そのレビューに採択したり, しなかったりした事を決定した時の明確なラインが示されているか.

(4) 収録文献の情報の正確さを組織的な方法で評価しているか.

(5) 情報が組織的に統合されているか. そのときデータの限界や不一致の点が詳しく述べられているか.

(6) 情報は統合され, 重み付けがなされているのか. また計量的に分割されているのか.

(7) 関連する知見の要約がなされているか.

(8) 結果から導き出された新しい研究の糸口が示されているか.

図 4.3: サーベイ論文の評価基準

図 4.2において, (2) は図 4.1で示した複数論文要約のポイントにおける「(a) 関連論文の自動収集」に対応する. また, (3) は同じく図 4.1の「(b) 論文間の関係の解析 」と関連する. 本研究ではこれらの処理を行うために論文間の参照情報に着目しているが, 4.4 節では, 図 4.2の (2) と (3) を, 参照情報を用いてどのように実現するかについて説明する.

図 4.3に関して, サーベイ論文の評価は本研究において非常に重要な問題の 1 つである. これらの評価基準は客観的であると思われるが, 最終的な評価は人間が主観的に行わざるを得ないと考えられる. ただ, すでに 3 章でも述べたとおり, 本研究で用いる論文データベースは小規模で十分な数の論文が得られないため, サーベイ論文作成支援システムをユーザに使ってもらい, 実際にサーベイ論文を作成してもらい評価するのは, 現時点では困難であると考えられる. 将来的には大規模な論文データベースを用い, 図 4.3の基準に基づいてユーザベースの評価を行う必要がある.

4.4　サーベイ論文作成における参照情報の利用

図 4.1に複数論文要約のポイントを示したが, 本節ではその中でも特に (a) 関連論文の自動収集と (b)-1-2,3 論文間の共通点と相違点の検出における, 参照情報の利用について説明する.

4.4.1　関連論文の自動収集

本研究では関連論文の自動収集に, 論文間の参照関係を利用する. 論文間の参照関係を単純に辿ることで, ある程度自動的に関連論文を収集することが可能であると考えられる. しかし, そのようにして得られた論文集合は複数分野の論文が混在してしまう可能性があり, サーベイ論文作成上望ましくない. そこで, 必要な参照関係のみを辿って論文を収集する手法が必要とされる. そのために, 参照タイプを考慮した論文収集の手法が考えられる. 著者は, type C の参照関係が論文収集に有効であると考えている. それは, 「type C の参照個所中の “既存研究の紹介” の記述が参照・被参照論文共通の問題点である」という仮定に基づいている. この仮定の妥当性を示すため, E-Print archive の論文を用いて調べた. その結果, 論文データベース中で type C の参照関係で結ばれる参照・被参照論文 31 組のうち, 29 組は参照・被参照論文共に同じ分野の論文であることが確認された[2]. 図 4.4は, 被要約対象

[2]ここで述べる分野とは, 付録に示す 58 カテゴリを指す

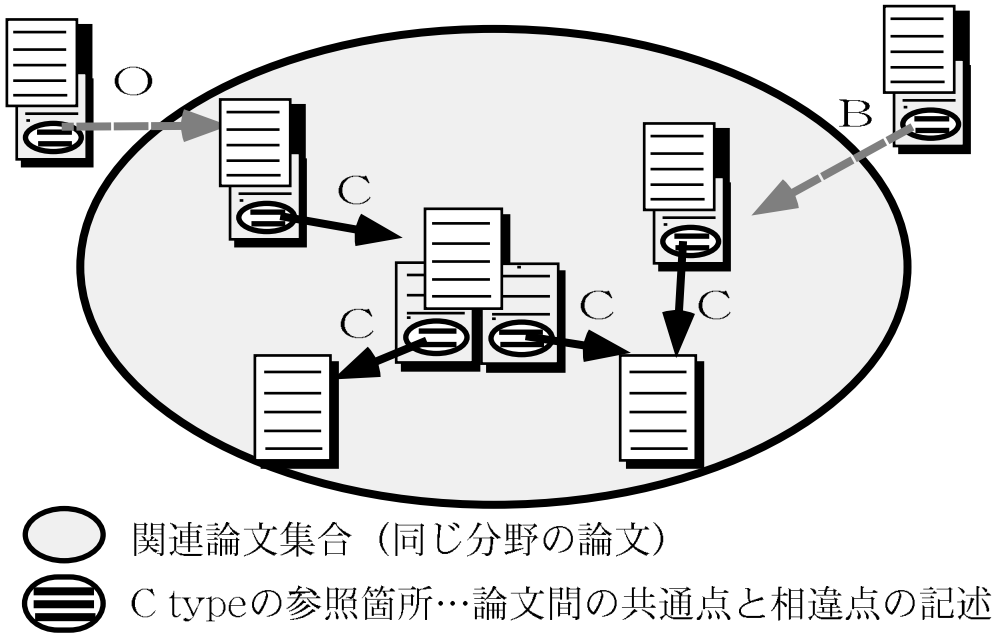


図 4.4: 論文間の共通点と相違点

論文の集合を示している (図中の楕円内の論文集合を以後, 参照グラフと呼ぶ).

4.4.2 論文間の共通点, 相違点の検出

type C の参照個所から得られる情報 (図 2.7) と, 複数論文要約のポイント (図 4.1) との関係について, 「(α) 既存研究の紹介」は「(b)-1-1 (被参照論文の) 重要個所の抽出」と「(b)-1-2 (参照・被参照) 論文間の共通点」に, 「(β) 既存研究の問題点」と「(γ) 研究の目的」は「(b)-1-3 (参照・被参照) 論文間の相違点」にそれぞれ対応している. 従って, 参照個所を抽出し提示することで, サーベイ論文作成支援が可能になると考えられる.

さて, ひとつの論文を他の複数の論文が参照する場合, 著者の観点毎に参照の仕方も異なる可能性がある. 図 2.6には, (Bond, 1996)[74] の (Murata, 1993)[78] に関する参照個所を示したが, 図 4.5に, (Murata, 1993) に関する (Bond, 1994)[75] と (Takeda, 1994)[80] の参照個所を示す.

(Bond, 1994) 中の文 (1) は図 4.1の「(α) 既存研究の紹介」「(β) 既存研究の問題点」に, 文 (2) は「(γ) 研究の目的」にそれぞれ対応する. また (Takeda, 1994) 中の文 (1)(2) が (α) に, 文 (3) が (β)(γ) に対応する. 2 つの論文の (α)(β)(γ) 同士を比較すれば, 同じ (Murata, 1993) に関しても著者毎に参照の仕方が様々であることがわかる. このように, ひとつの論

(Bond, 1994)[75] より抜粋

(1)Recently, [78] have proposed a method of determining the referentiality property and number of nouns in Japanese sentences for machine translation into English, but the research has not yet been extended to include the actual English generation.
(2)This paper describes a method that extracts information relevant to countability and number from the Japanese text and combines it with knowledge about countability and number in English.

(Takeda, 1994)[80] より抜粋

(1)Another example is the problem of identifying *number* and *determiner* in Japanese-to-English translation.
(2)This type of information is rarely available from a syntactic representation of a Japanese noun phrase, and a set of heuristic rules[78] is the only known basis for making a reasonable guess.
(3)Even if such contextual processing could be integrated into a logical inference system, the obtained information should be defeasible, and hence should be represented by green nodes and arcs in the TDAGs.

図 4.5: [Murata 93] に関する type C の参照個所

文を参照する複数の論文中の参照個所 (著者の観点の違い) を比較することはサーベイ論文作成の上で有用であると考えられる. また, 図 4.2(3) において, 収集された情報の正確さの評価は, 複数の参照論文の著者の評価を読み比べることで可能になると考えられる.

4.5 参照情報を利用したサーベイ論文作成支援システム

3 章で抽出された参照情報を用いて, サーベイ論文作成支援システムを作成した. サーベイ論文作成支援の流れを図 4.6に示す. サーベイ論文作成を支援する過程は大きく 2 つに分けられる. ひとつは論文検索過程である. 本研究では論文検索システム PRESRI(**P**aper **RE**trieval **S**ystem using **R**eference **I**nformation) を開発した. この検索システムには 2 種類の検索機能がある. ひとつはキーワード検索機能で, 論文のタイトル中の語や著者名をキーワードとして論文を検索できる. 検索結果はリスト表示される. このリスト中の個々の論文について, E-Print archive のデータベース中に参照・被参照関係の論文がある場合, 論文間の参照・被参照関係のグラフを表示することができる. このグラフを辿ることで, 論文間の参照・被参照関係を用いた検索が可能になる.

次にサーベイ論文作成支援過程について説明する. この過程では, 関連論文の収集, 関連論文の参照個所や概要の表示を行うことでサーベイ作成の支援を行う. このような機能を提供するために, 3 章で述べた参照個所の抽出や参照タイプの決定といった処理が必要とされる. 3 章で論文間の参照・被参照関係で type C のものだけを辿ることで関連論文の自動収集に近いことができることを示した. これは, 論文検索過程で示された論文間の参照・被参照関係を示したグラフを利用し, グラフ中で type C の参照・被参照関係だけ

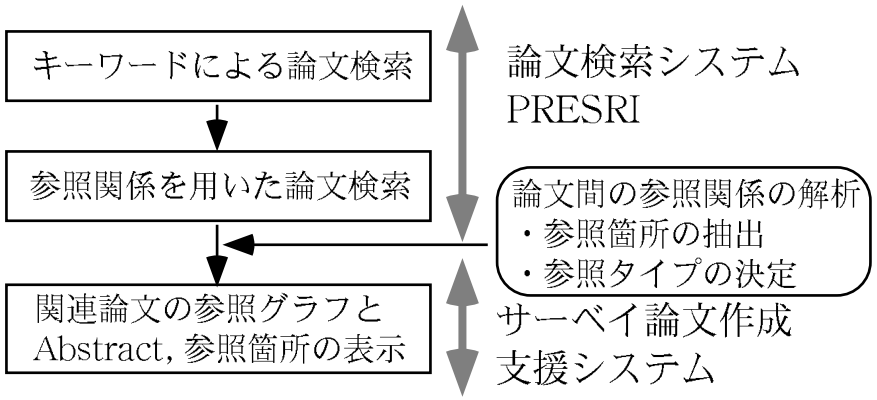


図 4.6: サーベイ論文作成支援の流れ

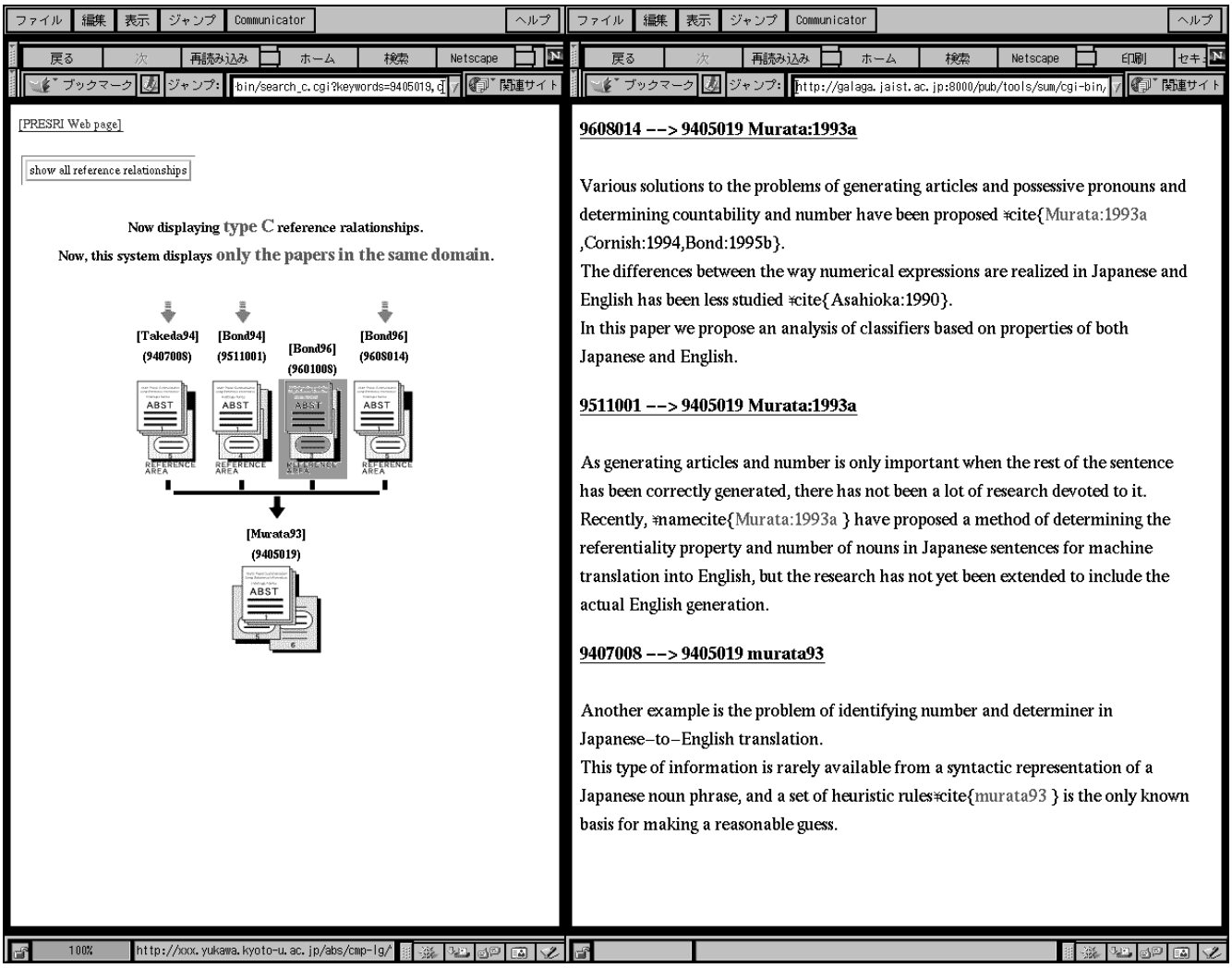


図 4.7: サーベイ論文作成支援システム

を表示することで, 実現可能であると考えられる. 図 4.7は, サーベイ論文作成支援システムの実行画面で, 左側のウィンドウは [Murata93](9405019) という論文に関する論文間の参照・被参照関係を示したグラフである. このグラフから 4 本の論文が [Murata93] を参照していることがわかる. この 4 本の論文のうち [Bond96](9601008) が黒く表示されている. これは, [Bond96](9601008) が Murata93(9405019) を type C 以外のタイプで参照しているためである. 他の 3 つの論文に関しては [Murata93](9405019) を type C で参照している. type C の参照・被参照関係の論文は関連分野の論文であると考えられ, グラフ中の "ABSTRACT" や "REFERENCE AREA"(参照個所) の個所をクリックすることで, 個々の論文の概要や参照個所を閲覧することが可能になる. 図 4.7の右側のウィンドウは 3 本の論文 [Takeda94](9407008), [Bond94](9511001), [Bond96](9608014) の [Murata93](9405019) に関する参照個所を示しており, 左側ウィンドウのグラフ中の "REFERENCE AREA" の個所をクリックした結果である[3]. このように, ひとつの論文を参照している複数の論文の参照個所を並べて表示することで, ひとつの論文に関する複数の著者の観点を直接比較することが可能となり, サーベイ論文作成において有用であると考えられる.

4.6 まとめ

本章では, 3 章で説明した手法により抽出された参照情報を用いて構築したサーベイ論文作成支援システムについて述べた. サーベイ論文を作成する過程は (1) 関連論文の収集, (2) 論文間の関係の解析, (3) 収集した論文の分類・整理の 3 つのステップに分けることができる. (1) について, 一般に論文間の参照・被参照関係をたどることで, ある程度特定分野の関連論文を収集できるが, 単純な参照・被参照関係だけでは, 分野外の論文も収集される可能性がある. そこで, 本研究では参照タイプに着目し, 特定の参照タイプ (type C) が付与された参照・被参照関係だけをたどって, 自動的に特定分野の論文だけを集める手法を提案した. 調査の結果, 提案手法では 94% の精度で特定分野の論文を集められることがわかった. (2) について, 論文中の参照個所には, 参照・被参照関係にある 2 論文間の関係に関する記述がある. したがって, type C の参照・被参照関係をたどって関連論文を集め, それらの参照個所を提示することで, サーベイ論文の作成支援が実現される. また, (3) については次章で述べる.

[3]図中に "REFERENCE AREA" が 3 個所あるが, いずれの個所をクリックしても右側ウィンドウの表示になる

4.7　今後の課題

提案システムの有効性を調べるために, ユーザベースの評価が不可欠であろうと考えられる. すなわち, 実際にサーベイ論文を作成する過程で有用であるか, ユーザの主観的な判断に基づいて評価してもらう必要がある. これには例えば図 4.3に示したサーベイ論文の評価基準の「(2) 収録文献を見つけた方法や, その情報源が明らかにされているか」「(7) 関連する知見の要約がなされているか」といった項目や, Oxman らのサーベイ論文の評価指標 [47] などが有用であると考えられる.

サーベイ論文の作成過程は, 4.3.2 節で紹介したサーベイ論文の分析研究からも分かるように, まだ明らかにされていない部分が多い. サーベイ論文自動作成の実現に向けて, 今後, このような分析をさらに発展させ, 人間のサーベイ論文作成の過程を明らかにし, サーベイ論文作成システムのモデル化を行う必要がある.

第 5 章

参照情報を考慮した関連論文の分類

本章では, 2 章で定義した参照情報の応用例として関連論文の分類を取り上げ, 実験によりその有効性を確認する. 5.1 節では, まず関連論文の分類の必要性について述べる. 5.2 節では, 関連研究を紹介し, その問題点について述べる. 5.3 節では, 本研究で提案する関連論文の分類手法について述べ, 5.4 節で, 実験により提案手法を評価し考察する. また, 提案手法により 4 章で説明したサーベイ論文作成支援システム PRESRI を拡張する. 5.5 節ではシステムの動作例を示す.

5.1 関連論文の分類の必要性

近年, 学術情報の爆発的な増加と共に, 数多くの電子化された論文がオンラインから入手できるようになった [25, 32]. これらの論文の書誌情報を抽出・蓄積すれば, 論文データベースとして論文検索が可能になるが, さらに, 論文の内容に基づいてあらかじめトピック毎に分類・整理しておけば, ユーザは必要な論文を効率的に入手できる. 本研究ではトピック毎の関連論文の分類を目指す.

これまで, クラスタリングやカテゴリゼーションの研究分野で, 文書を分類する様々な手法が提案されてきた [39]. その中心的な手法は, 文書中の語, 文書に付与されたディスクリプタ (キーワード) 等を用いて, 個々の文書をベクトル空間型モデル, 確率モデル等で内部表現に変換し, この内部表現により文書間の類似度を測る.

一方, 学術論文には論文間に参照・被参照関係があり, 論文の分類にはこのような参照構造が利用できる. これまで引用分析研究の分野において, 論文間の参照構造を利用し, 2 論

文間の類似度を測るいくつかの手法が提案されてきた [21, 54]. しかし, これらの手法はすべての参照・被参照関係を等価に扱っているが, 実際には Weinstock[71] や Moravcsik[35] が述べているように参照には様々な理由が存在するため, 必ずしも論文間の類似度を適切に評価できない.

そこで, 本研究では被参照論文の参照の理由 (参照タイプ) を考慮し, 参照構造を用いて論文間の類似度を測る手法を提案する. 本研究では, 2 論文間で同一論文を共に参照しており, かつそれらの参照タイプが一致している結合のみを数えるという方法で, 2 論文間の類似度を測る. この手法により, ノイズとなる結合を削減でき, また, 従来の引用分析手法と比べ, 精度の向上が期待できる.

次節では, トピック毎に関連論文を分類する手法について述べる.

5.2 関連研究

これまで, 学術論文のトピックの類似度に基づいて論文を分類するいくつかの手法が提案されてきたが, それらは大きく以下の 2 つに分けることができる.

- アプローチ 1:(語の共出現に基づく分類手法)

  トピックの似た 2 つの論文間では, 多くの語が論文間で共出現する傾向にある. このような共出現する語の数を数えることで論文間の類似度を測る [53].

- アプローチ 2:(引用分析に基づく分類手法)

  引用分析とは, 論文間の参照・被参照関係を用いて, 論文間の関係を分析する方法である. 書誌結合 (bibliographic coupling)[21] と共引用分析 (co-citation analysis)[54] は, 引用分析の代表的な手法であり, トピックの似た論文を集められることが知られている [28, 44, 72]. 書誌結合は, 論文間の関連度を測る時に, 2 論文間でどれだけ同じ論文を引用しているか, という基準に基づいている. 一方, 共引用分析は, 2 論文がどれだけ他の論文で共に引用されているか, という基準に基づいた手法である.

  従って, 発表されてから十分に時間がたっている古い論文を対象にする場合は共引用分析が適していると言える. これとは逆に, 他の多くの論文から引用されていないような新しい論文を組織化するには, 書誌結合の方が適している.

ここで, これらの 2 つのアプローチの問題点を以下に示す.

- アプローチ 1 の問題点: (語の共出現に基づく分類手法)

  2 論文全体にわたって共出現する語を調べるのは, 非常に時間がかかる. 従って, 対象とするテキストの長さを減らす必要がある.

- アプローチ 2 の問題点: (論文間の参照関係に基づく分類手法)

  引用分析に関する大半の研究は, すべての引用を等価に扱っている. しかし, 実際は Weinstock[71] が示すような様々な参照の理由が存在する. 従って, 関連論文をより正確に組織化するためには単純な参照・被参照関係だけでなく, より豊富な参照情報を考慮することが不可欠であると考えられる.

本研究では, 上記の問題を考慮した関連論文を分類するいくつかの手法を提案する.

5.3 関連論文の分類手法

提案手法 1: (語の共出現に基づく分類手法)

あらかじめ, 論文の特徴的な内容を表すパッセージを抽出し, このパッセージを用いて語の共出現を数えれば, アプローチ 1 による計算時間が短縮されると考えられる.

Kando は, 手がかり語に基づくいくつかのルールにより, 学術論文の機能構造の解析を行っている. 構造化された論文中で特定の意味役割の文 (Method and Validity) のみを用いて論文検索を行い, 論文全文を用いた検索と比べ, 精度が向上し再現率の低下も最小限にとどめることを示している [19].

三池らは, 日本語技術論文を対象に, 文書構造解析を行う BREVIDOC というシステムを開発している [34]. このシステムでは接続詞や文末表現等に着目し, 150 個の規則を用いて, 文の階層的な 2 分木を生成する. さらに, 例えば論文の背景について述べている文では「近年, –ている」といった表現が出現すると考えられるが, このような手がかり表現を用いて, 構造解析された文書から特定の役割を示す文を自動的に抽出している. 三池らは「背景」, 「話題」, 「従来の問題」, 「目的」, 「特徴」, 「結果」, 「結論」, 「課題」を示す文のみを用いて検索することで, 論文全文を検索に用いる場合よりも検索精度が向上することを, 実験により示している.

Kando の研究では, 研究の手法について記述された文が検索に有効であるという結果が得られているが, 三池らの研究では研究の背景や目的に関する記述が有効であると述べて

いる.

そこで, 本研究では, 以下の 2 種類の意味役割の文を抽出し, これらを論文の分類に利用する.

- **PURPOSE:**

  研究の目的が書かれてある個所は, 研究の分野と密接な関連があると考えられる.

- **METHODS:**

  研究の背景で用いている理論や手法も, トピック毎の分類を行う際に有用な指針になりうると思われる.

また, Kando, 三池らは共に大規模なルールを用いて文書構造解析を行っている. このような構造解析を行うことは, 論文検索ばかりでなく, テキスト自動要約等, 様々な目的に利用できるため意義深い. しかし, 論文検索の精度向上という目的に限れば, もう少し単純な手法で特定の意味役割の文を抽出しても論文構造解析に近い効果が得られると考えられる.

本研究では文書構造解析を行わず, 手がかり語だけを用いて, “PURPOSE” や “METHOD” の文を抽出する. “PURPOSE” の抽出には 5 個の手がかり語 “our work”, “Our work”, “this paper”, “This paper”, “purpose” を用いる. これらの語が出現する文は研究の目的について書かれていると考えられる. 同様に, “METHOD” の抽出には 84 個の手がかり語を用いる. この手がかり語は参照タイプの決定に用いたものであり (表 3.3参照), これらを含む文を抽出する.

次に, 論文間で抽出された文集合に共出現する語を数えることで, 論文間のトピックの類似度を計算した.

提案手法 2: (論文間の参照関係に基づく分類手法)

2 論文間で同一論文を共に参照しており, かつそれらの参照タイプが一致している場合のみ数えるという方法 (以後, BCCT: Bibliographic Coupling using Citation Types) で, 2 論文間の類似度を測る.

本研究では引用分析に共引用分析ではなく書誌結合を用いている. なぜならば, 本研究で用いる論文集合 (E-Print archive) は比較的新しい論文を多く含んでおり, 従って他の論文からあまり参照されていないためである.

また, 実験では “BCCT-C” と “BCCT-BCO” という 2 種類の手法を用いている.

“BCCT-BCO” はすべての参照タイプ B, C, O を考慮した書誌結合である. “BCCT-BCO” は, 2 論文が共通に参照する論文が存在しても, 異なる参照タイプで参照していれば, 類似度に反映しない点で従来の書誌結合と異なる.

また, “BCCT-C” は type C のみを考慮した書誌結合である. type C に着目した理由は, type C とは既存の研究の問題点を指摘する参照であり, 2 論文間で多くの type C の参照が一致すれば, これらの論文の著者は共通の問題意識を持っていると考えられるからである.

5.4 関連論文の分類手法の評価

5.4.1 評価方法

文書集合

前節で述べた提案手法の有効性を調べるために, いくつかの実験を行った. 近年, NTCIR, Cranfield, Medlars, CACM 等, 大規模な情報検索テスト・コレクションが作成されている. しかし, これらは論文抄録を検索対象にしており, 学術論文全文を検索対象にしたテスト・コレクションは作られていない. そこで本研究では, 既存のテスト・コレクションに比べると小規模ではあるが, E-Print archive の “The Computation and Language” に関する TEX 形式の論文データ 395 本を用いる.

正解セットと検索クエリ

関連論文の分類システムを評価するために, 395 論文を用いて正解データセットを作成した. これらの 395 本の論文を人手で 58 のカテゴリに分類した (付録).

実験方法は, まず, 395 本の論文から任意に 1 論文を選択する. これを検索クエリと見なし, 論文集合から検索クエリと同一カテゴリの論文を集めることを試みる. 検索システムは入力クエリに関する論文を集め, クエリに対して適合度の高い順に検索結果として論文の一覧を返す. このような手順を 395 回繰り返し, 395 本の論文それぞれについて関連論文を集める. これらと人手による分類を比較し, 検索システムの性能を評価する.

検索エンジン

ベクトル空間型モデルを用いて, 検索エンジンを作成した. 提案システムは, Brill の品詞タギングツール [4] を用い, パッセージから名詞のみを抽出しインデックスを作成する. 次にコサイン距離で論文間の類似度を計算する.

分類手法

実験は 8 種類の手法を用いて行った. 語の共出現を用いる手法は, 5.3 節で説明した “METHOD”, “PURPOSE” という 2 つの提案手法の他に, 論文表題 (“TITLE”) や概要 (“ABST”) を加えた. これらは論文の著者により作成された, 論文の特徴を表すパッセージと考えることができる. また, 各パッセージがどの程度, 原論文の内容を反映しているのかを調べるため, 論文全文 (“FULL”) を用いた分類も行う.

- **“FULL”, “TITLE”, “ABST”:**
  論文全文, タイトル語と概要中の語を用いた語の共出現.
- **“METHOD”, “PURPOSE”(提案手法):**
  手がかり語により抽出された文中に含まれる語を用いた語の共出現.
- **“NBC”:**
  書誌結合.
- **“BCCT-C”, “BCCT-BCO”(提案手法):**
  参照タイプが type C の時のみ結合を数える書誌結合 (BCCT-C) とすべての参照タイプを考慮した書誌結合 (BCCT-BCO).

5.4.2 評価

以下の評価尺度を用いて 8 種類の組織化手法の有効性を調べた.

- 上位 n 文書の精度
- フォールアウト
- 計算コスト

再現率 - 精度は, 情報検索の分野では最も一般的に用いられている評価尺度である. この尺度は検索エンジンの有効性の全体的なバランスを見る上では良い指針となるが, 本研究では評価には再現率は用いていない. 何故ならば, 実験に用いるテスト・コレクションではクエリによっては再現率が計算出来ないものがあるからである. 付録にも示したとおり, 58 カテゴリの中には 1 カテゴリに 1 論文しか含まないものがある. このような論文が検索クエリになった場合, 同一カテゴリの他の論文は存在しないため, 再現率の分母は 0 となり計算できない. 従って, 本研究では再現率の代わりにフォールアウトと精度で 8 手法の比較を行う.

精度とフォールアウトは次の式で与えられる.

$$\text{精度 } (Precision) = \frac{\text{検索システムにより集められた論文の中で正解の論文数}}{\text{検索システムにより集められた論文総数}}$$

$$\text{フォールアウト } (Fallout) = \frac{\text{検索システムにより検索された論文の中で不正解の論文数}}{\text{クエリと異なるカテゴリの総論文数}}$$

フォールアウトは検索エンジンのエラーを測る尺度であり, フォールアウト値が小さいほど良いシステムであると言える. 精度とフォールアウトを算出する際, TREC で使われている trec_eval というツール [62] を利用した. 通常は, trec_eval に正解文書セットと検索システムの出力結果を与えると, 再現率が (0 %, 10 %, 20 %, ..., 100 %) の 11 点における精度が計算される. ここで, 正解文書セットの代わりに正解と不正解を反転させた文書セットを trec_eval に与えると, フォールアウトが (0 %, 10 %, 20 %, ..., 100 %) の 11 点における 1-精度が計算される.

上位 n 文書の精度による評価の結果を図 5.1に, フォールアウトによる評価の結果を図 5.2に, 計算コストによる評価の結果を表 5.1に, それぞれ示す.

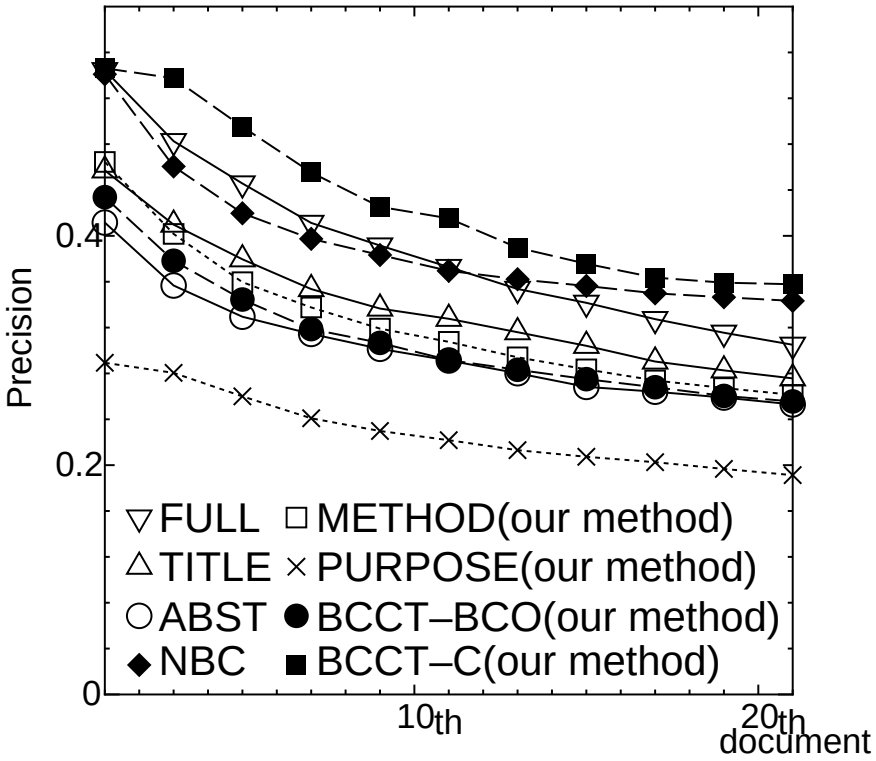


図 5.1: 上位 n 論文の精度の比較

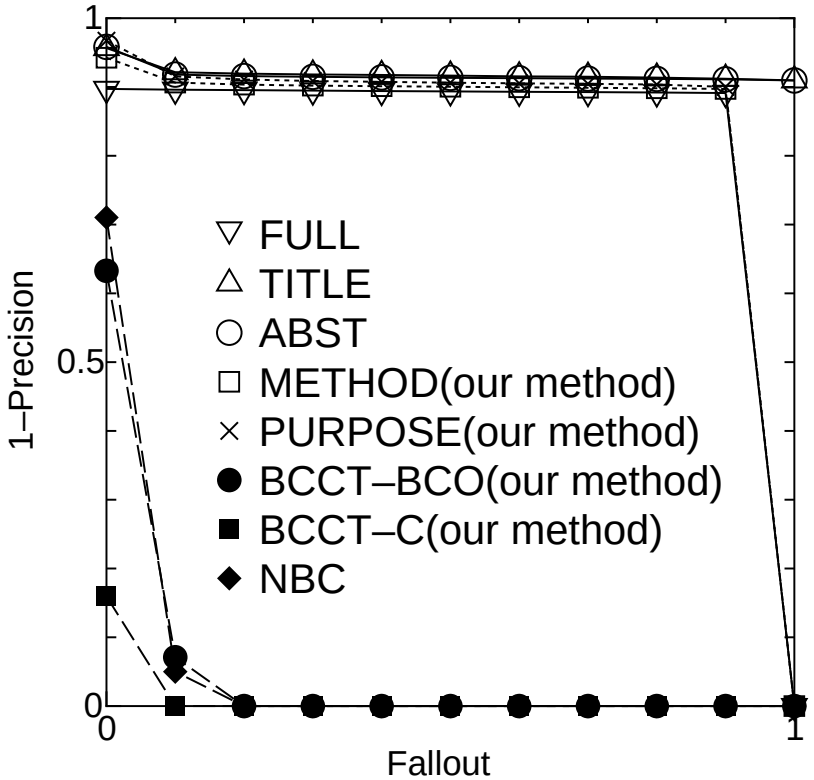


図 5.2: フォールアウトと精度による分類手法の比較

表 5.1: 計算コストによる比較 (クエリあたり)

手法	計算時間 (秒)
FULL	232
TITLE	0.25
ABST	1.2
METHOD (提案手法)	8.1
PURPOSE (提案手法)	0.77
BCCT-BCO (提案手法)	14
BCCT-C (提案手法)	1.3
NBC	14

5.4.3 考察

上位 n 論文の精度による評価

“FULL” と “NBC” が論文全体の情報を用いた分類手法, その他が論文の部分情報を用いた分類手法であると考えれば, 後者の手法で前者の精度を上回っているのは “BCCT-C” だけである. 特に “BCCT-C” は上位 1 位から 5 位の間において他の手法との精度の差が顕著に表れている. これは, “BCCT-C” が論文の中でも特にトピックに関連する情報を高い精度で抽出できていると考えることができる[1].

一方, 2 論文間で type C の (書誌) 結合が存在しなければ, 当然 “BCCT-C” で関連論文を収集することはできない. すなわち, “BCCT-C” は高い精度が得られる反面, 高い再現率が得られないため, 精度を犠牲にしてもより多くの関連論文を収集するには “TITLE” や “NBC” の方が “BCCT-C” よりも適していると言える.

また, “BCCT-C” による分類の失敗の原因の多くは, 参照タイプの判定の誤りに関するものであったが, その他に, 2 論文が同一論文を type C で参照していても, 2 論文が被参照論文について述べているポイントがずれている場合, 異なる分野の論文を集める場合があった. 図 5.3に “BCCT-C” による分類の失敗例を示す.

[1]今回実験に用いたテストコレクションでは, 1 論文あたり平均約 18.1 本の論文を参照している. また, type C での参照は 1 論文あたり約 2.9 本となっている.

(Scheler, 1996)[79] の (Church, 1988)[76] に関する type C の参照個所
「冠詞を含む名詞句の意味素性の解析, 生成, 文法チェック」

> The different logical forms of the sentences can be represented by a set of sentential operators, which are defined in first-order logic. These sentential operators can be used as atomic semantic features, which are consequently sufficient in representing the logical meaning of a sentence with respect to the chosen semantic dimensions. This approach is significantly different from POS or sense-tagging systems such as (Yarowsky92) (Schmid94) (Jelinek85) **(Church, 1988)** (Brill93).

(Heeman, 1997)[77] の (Church, 1988)[76] に関する type C の参照個所
「品詞タギングと言語モデルの結合」

> The final probability distributions are similar to those used for POS tagging of written text (DeRose88:cl) **(Church, 1988)**. However, these approaches simplify the probability distributions as is done by previous attempts to use POS tags in speech recognition language models.

図 5.3: “BCCT-C” の失敗例

図 5.3は, Scheler[79] と Heeman[77] の Church[76] に関する type C の参照個所を示したものである. (Scheler, 1996) と (Heeman, 1997) はそれぞれ, 「冠詞を含む名詞句の意味素性の解析, 生成, 文法チェック」と「品詞タギングと言語モデルの結合」に関する論文である.

Scheler の研究では, 名詞句を分類する 5 つの観点 (dimension)(“Generalized quantification”, “Anaphoric relation”, “Reference to discourse objects”, “Boundedness”, “Active involvement”) を設定し, 名詞句を自動的に分類する手法を提案している. (Scheler, 1996) の参照個所中の記述 (図 5.3上) によれば, これらの観点は一階述語論理の形式で定義されているが, 同時に文オペレータという異なる形式でも表されており, この文オペレータを意味素性の代わりとして用いる点が (Church, 1988) をはじめとする品詞あるいは意味タグを付与するシステムと異なる, と述べている.

この記述は, 参照論文 (Scheler, 1996) と被参照論文 (Church, 1988) との思想的な違いを

述べているので, type C の参照であると考えられるが, (Church, 1998) の問題点を明示的に述べているわけではない.

一方, (Heeman, 1997) では (図 5.3下), (Church, 1988) について, 「音声認識における品詞タグの取り扱いの時と同様, (Church らの研究では) 確率分布を単純化しすぎている」と (Church, 1988) の問題点を明示的に述べている.

このように, (Scheler, 1997) と (Heeman, 1997) では, 同じ type C でも (Church, 1988) の言及の仕方が全く異なっており, このような場合, "BCCT-C" では失敗している.

図 5.1において, "BCCT-BCO" は "NBC" よりも精度が低かった. 失敗の原因を調査したところ, type B の (書誌) 結合が論文をトピック毎に分類する際あまり有効ではなく, また場合によっては分類を阻害する方向に作用することが判明した. "The Computation and Language" の分野において, 形態素解析器や構文解析器などのツールは多くの研究で汎用的に使われる. 従って, 2 つの論文がこのようなツールについて書かれた論文を共に type B で参照していても, トピック毎の論文の分類には有用ではない.

参照タイプを考慮した書誌結合の手法として, 前節で説明した "BCCT-C" や "BCCT-BCO" の他にも "BCCT-B" という手法も事前に考えられた. しかし, 予備調査の段階で, 先に述べた理由により type B の (書誌) 結合がトピック毎の分類に向かないことが判明したため, 比較手法に "BCCT-B" を入れなかった. しかし, 実際には "BCCT-B" ばかりでなく, "BCCT-BCO" においても type B の結合が, その精度を下げる要因になった.

図 5.1において, 語の共出現による 3 手法, "ABST", "METHOD(提案手法)", "TITLE" について詳細に調べた. 3 手法についてそれぞれの平均精度の上位 5 件まで, 10 件まで, 15 件まで, 20 件までの手法毎の値を算出し, 表 5.2にまとめた. 表 5.2において, 全般的に "METHOD" が "ABST" を上回り, 特に上位 5 件においてその差が顕著に表れている (9.4 %). すなわち, 計算機で論文を分類する上では, 人間が作成した概要よりも METHOD の方が有用であることを示している.

一方, "METHOD" よりも "TITLE" の分類精度が若干上回っている. 一般に, 論文表題には論文の内容を表す代表語が多く含まれていると言われており, それは今回の実験結果にも表れていると考えられる. しかし, 論文表題が非常に短い場合は, "TITLE" よりも "METHOD" の方が有効であると考えられる.

あらゆる手法の中で "PURPOSE" の精度が一番低かった. 論文から抽出される文数が少なかった (PURPOSE:平均 4.9 文, METHOD:平均 31.7 文) というのが, その理由の 1 つと

表 5.2: 上位 n 論文の精度の比較

ranking	ABST	METHOD (提案手法) (METHOD/ABST)	TITLE (TITLE/METHOD)
5th	0.3606	0.3944	0.4144
		+9.4 % ($\frac{0.3944}{0.3606}-1$)	+5.1 % ($\frac{0.4144}{0.3944}-1$)
10th	0.3258	0.3413	0.3625
		+4.8 % ($\frac{0.3413}{0.3258}-1$)	+6.2 % ($\frac{0.3625}{0.3413}-1$)
15th	0.2937	0.3112	0.3269
		+6.0 % ($\frac{0.3112}{0.2937}-1$)	+5.0 % ($\frac{0.3269}{0.3112}-1$)
20th	0.2797	0.2888	0.3018
		+3.3 % ($\frac{0.2888}{0.2797}-1$)	+4.5 % ($\frac{0.3018}{0.2888}-1$)

して挙げられる. また, “PURPOSE” に含まれる語は, 論文の内容を良く表わしている場合もある. しかし, 多くの場合抽象的すぎるか, 論文の非常に具体的な記述で, 代表語として適切でないものが多く含まれていた.

同様の結果が Kando により報告されている [19]. Kando は, “Method and Validity” と “Evidences” という意味役割が振られた文 (本論文の “METHOD” に相当する) と “Research Topic” の意味役割の文 (本論文の “PURPOSE” に相当する) を用いて検索を行った結果, どの文書も “Research Topic” の意味役割がふられた文が 1 文以上存在していたにもかかわらず, “Method and Validity” と “Evidences” を用いた解析精度が “Research Topic” の解析精度を上回ると報告している.

フォールアウトによる評価

図 5.2において, 書誌結合に基づく 3 つの手法 (“BCCT-BCO”, “BCCT-C”, “NBC”) でいずれも良い結果が得られている. 3 つの中でも特に “BCCT-C” が一番優れている. これは, あらゆる参照の理由の中で, type C が関連論文を集める上では重要な参照の理由であることを示している. また, 図 5.1では, “NBC” と “TITLE” はほぼ同程度の精度が得られていたが, システム全体で比較した場合, “NBC” の方が “TITLE” よりもトピックの異なる論

文を収集しない, という面で優れていると言える.

語の共出現に基づく手法のフォールアウト値が高い理由は, 書誌結合に基づく手法に比べ, 検索システムが収集する論文数が多いからである. すなわち, 関連論文を漏れなく集めるには語の共出現に基づく手法が適しているが, なるべく高い精度で関連論文を集める場合には書誌結合に基づく手法, 特に “BCCT-C” が適していると言える.

計算コストによる比較

最後に, 8 種類の分類手法の計算コストを比較した (表 5.1). 計算時間は, クエリ毎にトピックの類似度を計算するのに要した時間である. これには品詞タギングや “METHOD”, “PURPOSE” の文抽出に要した時間は考慮していない.

8 手法の中で上位 n 論文の精度では “FULL” と “NBC” は比較的良い精度が得られていたが, 計算コストの面では “FULL” と “NBC” が最も遅かった.

以上をまとめると, より高い精度でかつ妥当な計算コストで関連論文を集めるためには提案手法である “BCCT-C” が最も適していると言える.

5.5 “BCCT-C” の応用 - サーベイ論文作成支援システムの拡張 -

4 章で説明したサーベイ論文作成支援システム PRESRI を提案手法の 1 つである “BCCT-C” を用いて拡張した.

図 5.4の左側のウィンドウは, “CID:8001780” (Pereira, 1992)[81] に関する論文間の参照・被参照関係を示したグラフである. データベース中ではこの論文を 8 本の論文が参照している. このうち 3 本は type C で (“DID:9512002”, “DID:9504034”, “DID:9606014”), 3 本は type B で (“DID:9611002”, “DID:9604008”, “DID:9605036”), 2 本は type O で (“DID:9605012”, “DID:9606027”), それぞれ参照している.

さらに, 図 5.4の左側のウィンドウ中央において, type B で (Pereira, 1992) を参照する 3 本の論文のうち, “DID:9604008”[82] と “DID:9605036”[83] はひとつのグループにまとめられている. これは, この 2 つの論文には type C で共通に参照する論文があり (BCCT-C), PRESRI がこの 2 つの論文は似たようなトピックの論文であると判定したためである. 実際この 2 論文は同一著者により書かれた構文解析に関するものであり, “DID:9611002”[84]

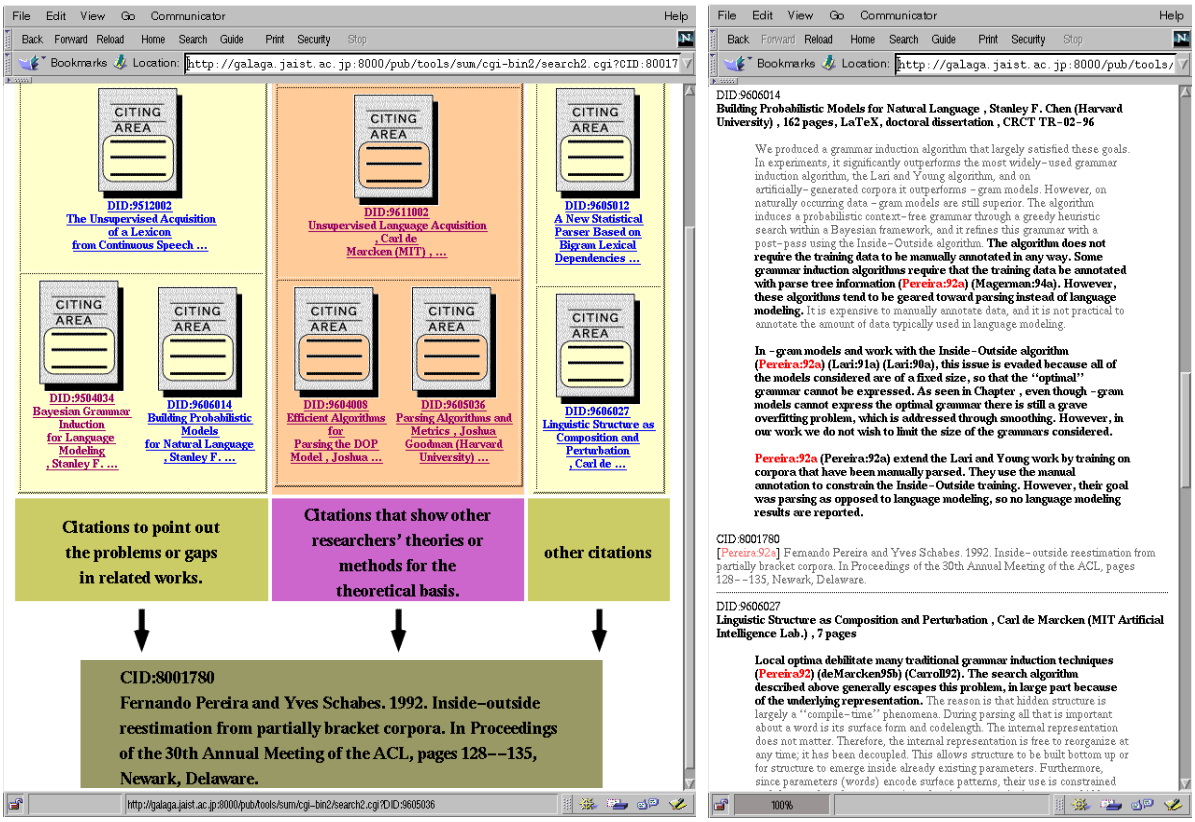


図 5.4: サーベイ論文作成支援システム

のトピック (語彙の自動学習) とは異なる. すなわち, (Pereira, 1992) の手法が構文解析や語彙の自動学習といった異なる分野で用いられていることが, この図より明らかになる.

また, 参照論文中で被参照論文の提案手法がどのように用いられているのかは, 参照論文中の対応する参照個所を読めば分かる. グラフ中の "CITING AREA" の個所をクリックすると, 論文中の参照個所が表示される. 図 5.4の右ウィンドウでは, 論文中で "CID:8001780" の参照を含んでいる段落を並べて表示している. 表示された段落の中で太字になっている個所が参照個所である. ここでは "DID:9606014" と "DID:9606027" の論文の著者が "CID:8001780" について述べている個所が示されている.

PRESRI を実装するにあたり, NEC Research Institute の Steve Lawrence 博士より提供していただいた論文データベース ResearchIndex[25] のソースの一部を参考にさせていただいた. ResearchIndex は World Wide Web 上に存在する Postscript 形式および PDF 形式の論文データを集めて作成されたシステムである. このシステムでも PRESRI と同様, キーワード検索と論文間の参照関係を用いた検索機能を提供している. また, PRESRI のように複数の参照論文を分類することは行っていないが, 書誌結合を用いて関連論文をユーザに提示する機能を提供している.

5.6 まとめ

本章では, 2 章で定義した参照情報の応用例として, 関連論文の分類を取り上げた. 関連論文を分類するためには, 2 つの論文間のトピックの類似性を測るための尺度が必要になる. これまで学術論文を自動的に分類するための尺度がいくつか提案されてきた. それらは語の共出現に基づく手法と論文間の参照・被参照関係に基づく手法の 2 種類に分けられる. 前者の手法では, 論文全文にわたって共出現する語を調べると, 計算コストが非常にかかるという問題がある. また, 後者の手法に関するこれまでの研究は, 論文間の参照・被参照関係は参照の理由を考慮しておらず, 従って関連のない論文も集めてしまうという問題がある.

本章では, 前者については, 手がかり語を用いて論文の内容を表す文を抽出し, 抽出された文のみを用いて語の共出現を調べるという方法で, 計算コストを抑える手法を提案した. また, 後者に関しては, 3 章で述べた手法で自動的に決定された参照タイプを考慮し書誌結合という手法を拡張した "BCCT" という手法を提案した.

実験では, 提案手法といくつかの既存の手法を比較した. 比較の際には, 「上位 n 件の精度」, 「フォールアウト」, 「計算コスト」という 3 種類の尺度を用いた. その結果, 提案

手法の 1つである “BCCT-C” が「上位 n 件の精度」と「フォールアウト」において最も良い分類精度が得られ, また「計算コスト」の面でも妥当な計算速度が得られた.

また, “BCCT-C” を用いて 4 章で説明したサーベイ論文作成支援システム PRESRI を拡張した. 拡張された PRESRI の動作例においても, その有効性が確認された.

5.7 今後の課題

本章では, 関連論文をトピック毎に分類する手法を 4 種類提案し, その中で参照タイプを用いた書誌結合が分類精度においても計算コストの面からも有用であることを示した. 一方, サーベイ論文の作成においては, 収集した同一トピックの論文をさらにサブトピックに分類する必要がある. このような場合において, 本章の提案手法が既存の手法と比較してどの程度有効であるのか, 調査する必要があると考えられる.

逆に, より粗い分類を行う上で, 提案手法がどの程度有効であるかについても, 調査する必要がある. 今回実験で用いた文書集合では, 考察でも述べたように形態素解析器や構文解析器に関する type B の参照は分類精度を下げる要因になった. しかし, 論文中で形態素解析器や構文解析器に関する論文を参照していることは, その論文が “The Computation and Language” の分野の論文であるかどうかを判断する上で重要な情報になりうる. おそらく, type B の参照は, Document Frequency (論文集合からどれだけ参照されているか) と深い関係があり, “BCCT” で結合数を数える際, (type B の参照数 / DF) のように補正することで, 分類精度の改善が可能になると推測される.

本研究では, 参照個所を分類する際, 5.2 節で述べた以下のような仮定を前提にしている.

> 1 つの論文中の個々の参照個所はその論文の構成要素であり, その論文全体の目的や方法と密接に関連している. 従って参照個所の内容は, その部分だけの局所的な内容だけでなく, その参照個所を含む論文全体の目的や方法との関連の中で捕らえる必要がある. そこで, 本研究では 1 論文中の全参照個所は同じトピックであると仮定した.

一方で, 「参照個所は参照論文の著者の観点から見た被参照論文の要約であり, 従って 1 つの論文中の個々の参照個所は, その個所だけに着目すればむしろ被参照論文のトピックと密接な関係にある」という考え方もできる. 5.1 節で示したように, ある論文を参照する複数の参照論文の参照個所を分類する, という特殊な目的においては, この考え方を当ては

めるとすべての参照個所が同一トピックとなってしまい, 分類の意味をなさなくなる. しかし, 参照個所を独立したパッセージと捉え, 一般的にパッセージ分類を行う場合においては, 参照個所は参照論文と同一トピックであると考えるか, 被参照論文と同一トピックであると考えるかの議論は, パッセージの定義に関する基本的な事項であり, 今後の課題の 1 つとして検討する余地がある.

第 6 章

結論

本論文では, 当該論文と被参照論文との関係を明示する個所 (参照個所) を論文中から抽出し, さらに参照個所から参照の理由 (参照タイプ) を自動的に判別する手法を提案した. また, 参照個所や参照タイプをサーベイ論文の作成支援や関連論文の分類に利用し, その有効性を示した.

本研究では, まず参照情報を言語的な情報を用いて自動的に抽出する手法を提案した. 参照情報の抽出は, 論文中の参照個所の抽出と, 参照タイプの決定という 2 つの処理がある. 本研究では参照個所の抽出を, 参照のある文と文間のつながりが強いと考えられる文を, 参照の前後の文から抽出する処理であると考え, 手がかり語を用いて参照個所の自動抽出を行った. その結果, 再現率 80 %, 精度 76 %の抽出精度が得られた. また, 抽出された参照個所を手がかり語を用いて解析し, 参照タイプを明らかにした. 実験の結果, 参照タイプ決定では 83 %のタイプ決定精度が得られた.

抽出された参照情報の応用例として, サーベイ論文作成支援を取り上げ, サーベイ論文作成支援システムを作成した. このシステムでは論文データベース中から特定分野の論文を自動収集し，関連論文間の相違点や個々の論文の概要が閲覧可能である. ひとつの論文を参照する複数の論文の参照個所を並べて表示することで，著者間の参照が直接比較できるため，サーベイ論文作成の際に有用である.

また, 既存の引用分析手法の改良例として, 関連論文の分類を取り上げた. 書誌結合は引用分析の代表的な手法であるが, 書誌結合ではすべての参照を等価に扱っているため, 十分な分類精度が得られていない. そこで, 本研究では参照情報を利用した関連論文の組織化手法を提案し, 実験により, 提案手法と書誌結合を含む既存の手法を比較した. その結果, 分

類精度において提案手法が最も優れており, また計算コストの面でも, 他の手法と比較して提案手法が十分高速であることが分かった.

今後の課題

本論文で提案した, 参照情報抽出の手法は手がかり語に基づいており, 他分野の論文への適用は可能であると思われるが, 今後は, 実際にいくつかの分野の論文を用いて実験を行い, その適用性を確認する必要があると考えられる.

本研究で取り扱った論文はすべて英語で記述されている. 一方, 4 章で述べたサーベイ論文の作成支援には, サーベイ論文の網羅性を考えると, 英語以外の言語で書かれた論文にも対応する必要があると考えられる. 本論文で提案する参照情報の抽出手法は, 手がかり語に基づいている. 手がかり語の抽出は, 統計的な手法に基づき半自動的に行っているため, 対象言語が異なっても参照情報抽出ルールの作成自体はそれほど困難を伴わないと考えられる.

また, 各言語毎に参照タイプ決定ルールを作成すれば, 複数の言語で書かれた論文データベースにおいて, 5 章で提案した関連論文組織化の手法 "BCCT-C" がそのまま適用できる. 一般に複数言語で書かれた論文集合を組織化するには, 機械翻訳の技術が必要不可欠である. また, 機械翻訳の精度が, 関連論文組織化の精度に直接影響してくる. しかし, 学術論文が対象の場合, すべての研究分野において専門用語の対訳辞書が存在するとは限らず, また, 仮に存在してもメンテナンスに非常にコストがかかる. これに対し, 本論文の提案手法では, 最初にある言語を対象にした参照情報抽出ルールを作成してしまえば, その後はメンテナンスの必要がない. このような理由からも, 提案手法の他言語の論文への拡張は非常に有用であると考えられる.

本論文では, 研究対象として学術論文を取り扱っているが, 提案手法の学術論文以外のテキストへの拡張も考えられる. 例えば, 特許は, 本論文の提案手法を最も適用しやすい文書の一つであると考えられる. その理由は, 特許文書の構造がある程度学術論文と似ていること, 他の特許との関係や違いを明確にする必要性から, 特許文書中の参照の前後は参照情報抽出の手がかりとなる表現が現れる可能性が高いことなどが挙げられる.

また, 特許の申請の際には予め他の関連特許を網羅的に調べる必要があるが, 本論文の 4 章で提案したサーベイ論文作成支援システムは, このような目的にも有用であると考えられる. また, 関連特許の調査の際, 複数の言語で記述された特許を調べる必要があるが,

前節でも述べたとおり本論文の 5 章で提案する関連論文組織化の手法は，複数言語の適用にも適していると考えられるため，このような調査にも有益であると推測される．

また，提案手法のウェブ文書への適用も考えられる．ウェブ文書は，文書間にハイパーリンクが張られており，他の複数の文書と参照・被参照関係にある．2 章でも紹介したように，引用分析の技術はすでにいくつかウェブ文書の検索や分類に応用されている．ウェブ文書は，学術論文や特許などと比べ，記述形式や文書の長さがまちまちである．また，リンク先の文書に関する記述も多くのリンク集に見られるようにほとんど記述が存在しないものから，ウェブ文書全体がリンク先の文書について記述されているものまで様々である．

一方，ウェブ文書は全体で 10 億ページ以上はあると言われており，こうした膨大な数の文書から必要な情報を効率的に見つけ出すための手段の一つとして，本論文で提案するような参照情報の抽出技術の確立が必要であると考えられる．

本論文で取り上げた研究とともに，参照情報は様々な応用研究において有効な情報として利用できる．近年，増大を続ける電子化文書を有効利用し，より良い人間の支援を実現するために，これらの研究についても今後検討していく必要がある．

謝辞

本研究を行なうに当たり, 終始, 御指導ならびに御鞭撻を賜わりました奥村 学助教授に深甚なる感謝の意を表します.

国立情報学研究所の神門 典子助教授には, 副テーマ研究において熱心な御指導, 御助言を頂きました. また本論文における研究を含め, 幾つかの研究に対し大変有意義な議論をして頂きました. 深く感謝致します.

東条 敏教授, 石崎 雅人助教授, 京都大学大学院情報学研究科の佐藤 理史助教授, および国立情報学研究所の影浦 峡助教授には, 本研究に対する適切な御助言, 御指導を頂きました. 深く感謝致します.

また, 日頃から有益な御助言をいただき, 多面に渡って励ましていただいた島津明教授, 望月 源助手に感謝致します.

慶應義塾大学文学部の上田修一教授には, 参照情報に関する貴重な御意見を頂き, また引用文脈分析の関連研究を紹介して頂きました. 深く感謝致します.

また, 本論文をまとめるに当たって御協力いただいた島津・奥村研究室の諸兄に厚く御礼申し上げます.

論文データの提供およびサーベイ論文作成支援システム PRESRI の公開を快く承諾して下さった E-Print archive administrator の方々に感謝致します.

PRESRI の実装にあたって, NEC Research Institute の Steve Lawrence 博士, C. Lee Giles 博士, Texas 大学の Kurt Bollacker 博士から提供していただいた論文検索システム ResearchIndex のソースコードおよびインターフェースを一部参照させていただきました. 深く感謝致します.

最後に, 常に著者を励まし応援してくれた両親に感謝致します.

付録 (「5 章 関連論文の分類」の実験に用いた正解セット)

分類カテゴリは, 自然言語処理の分野のいくつかの教科書 [38, 59, 60] の構成 (章立て), 言語処理学会年次大会発表論文集 (第 3 回 – 第 6 回) のセッションの分類を参考にした. 素性構造, 単一化 (“feature structure, unification, TAG, HPSG”) は, 構文解析 (“parsing”) に含むもの [38, 59] と, 意味解析に含むもの (“semantic analysis, word sense disambiguation”)[60] の 2 通りあった. また, 言語処理学会年次大会では開催される年度によって単一化が単独のセッションになっている年と構文解析のセッションに含まれる場合があった. 本研究では, 組成構造, 単一化は構文解析や意味解析とは独立したカテゴリとして考える.

また, 今回設定したカテゴリ数 (全 58) は自然言語処理の教科書の構成と比べるとかなり多い. 「形態素解析」「構文解析」「意味解析」といった “The Computation and Language” の研究分野におけるいくつかの典型的な分野に含まれない新しい分野の論文が少なからず存在するのが理由の一つである. また, E-Print archive “The Computation and Language” の論文データベースには, 分野外と考えられる論文がいくつか含まれていたことも, カテゴリ数が増えた理由の一つである. E-Print archive は, 論文の著者が自発的にデータベースに登録する形式をとっている. 従って著者が論文を誤ったカテゴリに登録しても, その著者が気づかない限り, 論文が第三者によって削除されることはない. “The Computation and Language” は自然言語処理や計算言語学と呼ばれる研究分野の論文を含むと一般的に考えられるが, 中にはプログラミング言語やコンパイラに関連する論文も含まれていた. これらの論文を削除した上で実験を行うことも考えられたが, このような論文を検出することも重要であると考え, 分類対象から削除しなかった.

以下は, 5 章の実験に用いた論文集合と, そのカテゴリである. 個々の論文は E-Print archive における登録番号で表記している.

- **parsing (not including unification, HPSG, TAG etc.)**

  9706001 9606016 9710005 9708013 9604019 9706003 9605003 9405028 9507003 9601002 9506021 9505040 9502017 9605018 9704009 9502004 9706002 9504026 9607020 9605012 9404003 9502021 9404008 9605036 9701004 9408004 9504030 9505042 9405009 9405022 9405023 9807006 9406029 9406031 9410014 9604009 9708008 9702009 9404007 9508002 9605038 9502024 9604008 9503023 9504034 9505006 9505031 9604013 9604017 9605016 9605023 9606017 9606020 9611001 9607001 9607035 9705006 9705009 9706004 9709001 9709010 9508009 9409008 9411021 9502031 9606011 9606014 9807007

- **discourse and dialogue**

  9405002 9609006 9410005 9705002 9405010 9502023 9706011 9502014 9503008 9704013 9707009 9605007 9708005 9505043 9706012 9706020 9410006 9502018 9503018 9505039 9701003 9504007 9505032 9505001 9801002 9409012 9505038 9505025 9705003 9606010 9512003 9608007 9608008 9608009 9609005 9609007 9705004 9708001 9708003 9503017 9605001 9504006 9511003 9609002 9612002 9612003 9612004 9702007 9703004 9606031 9407010 9407009 9704004 9704005 9707014 9702015 9407011 9709006 9408015 9806019 9406006 9704008 9406004 9706019 9405013

- **semantic analysis, word sense disambiguation**

  9605009 9702008 9703003 9605013 9503025 9605029 9601004 9406026 9606003 9503024 9505011 9505034 9510007 9610001 9511007 9601007 9607031 9607032 9706013 9706028 9708010 9511006 9712007 9712008 9712006 9806014 9702010 9704007 9706008 9706010 9502009 9707016 9502028 9708011 9408011 9605014 9807004 9607028 9405001 9409004 9607017 9505019

- **feature structure, unification, TAG, HPSG**

  9709014 9504009 9502003 9411025 9512005 9605015 9605005 9505033 9609001 9406040 9502005 9404009 9504012 9506004 9507001 9504029 9502022 9505009 9408016 9503005 9502015 9404010 9505028 9709011 9505030 9805008 9405020 9503022 9708012 9404001 9706022 9707010 9806017 9606006 9411012 9610003 9707012 9503021 9404011 9603002

- **machine translation**

  9808003 9607011 9604020 9607027 9706026 9805005 9805006 9607009 9703005 9704001 9504027 9701002 9505045 9510008 9508006 9705015 9705007 9405019 9407008 9410009 9511001 9601006 9601008 9608014 9608019 9706024 9706025 9405035 9706027

- **tagging, morphological analysis**

  9410012 9707015 9503009 9606021 9704011 9406010 9604012 9506024 9604022 9504023 9604025 9606005 9407001 9503004 9505026 9505035 9705011 9705014 9705016 9706005 9707003 9710002 9807013 9502038 9507004 9607021 9504002 9504024

- **generation**

  9405004 9604024 9709005 9411031 9506022 9411032 9504013 9505008 9707001 9605002 9607015 9607014 9607026 9704012 9708002 9712001

- **speech recognition, phonology**

  9408010 9603001 9605028 9607023 9412005 9707020 9607013 9707011 9708007 9611002 9406034 9702003 9607036 9604015 9512002 9603005 9606027 9608020 9608021

- **categorization, classification, clustering**

  9707002 9602004 9503002 9709007 9706006 9709004 9710008 9606004 9609003 9705005 9412003 9703001 9606002

- **learning**

  9406003 9801003 9801004 9509001 9509002 9705012 9705010 9606030 9505012 9405018

- **tagset**

  9506005 9406023 9506006 9505010 9604005

- **knowledge base**

  9411011 9702014 9703002 9508011 9704010

- **logic**

  9504028 9405031 9404005

- **information extraction**

  9705013 9702006 9706023

- **interface**

  9503016 9611006

- **sentence boundary disambiguation**

  9411022 9704002

- **information retrieval**

  9608003 9808002

- **metaphor**

  9607034

- **text summarization**

  9411023

- **dictionary**

  9605024

- **others (1 カテゴリ中に 2 論文を含むもの)**

  9606029 9605032

  9502032 9506013

  9502039 9506025

  9801001 9611004

- **others (1 カテゴリ中に 1 論文しか含まないもの)**

  9805003 9404002 9507002 9412008 9501005 9505004 9505014 9505036 9505041 9506002 9506018 9506023 9506026 9508001 9604011 9604021 9605010 9605020 9606008 9606009 9607016 9607018 9607019 9608001 9608002 9702004 9702005 9706021 9710003 9710007 9807008 9406030

参考文献

[1] Amitai, E., "InCommonSense - Rethinking Web Search Results", *IEEE, International Conference on Multimedia and Expo (ICME-2000)*, 2000. http://www.mri.mq.edu.au/~einat/incommonsense/publications.html

[2] Biber, D. and Finegan, E., "Section 13: Intra-textual Variation within Medical Research Articles", *Corpus-Based Research into Language*, Oostdijk & de Haan(eds.), Amsterdam, Rodoph, pp.201–221, 1994.

[3] Bonzi, S., "Characteristics of a Literature as Predictors of Relatedness between Cited and Citing Works", *Journal of American Society Information Science*, Vol.33, No.4, pp.208–216, 1982.

[4] Brill, E., "Some Advances in Rule-based Part of Speech Tagging", *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, pp.722–727, 1994.

[5] Brin, S. and Page, L., "The Anatomy of a Large-scale Hypertextual Web Search Engine", *Proceedings of 7th International World Wide Web Conference*, pp.14–18, 1998.

[6] Chakrabarti, S., Dom, B.E., Gibson, D., Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A.S., "Mining the Link Structure of the World Wide Web", *IEEE Computer*, Vol.32, No.8, pp.60–67, 1999.

[7] Chubin, D.E., and Morita, S.D., "Content Analysis of References: Adjunct or Alternative to Citation Counting?", *Social Studies of Science*, Vol.5, pp.423–441, 1975.

[8] Church, K., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of the Second Conference on Applied Natural Language Processing*, pp.136–143, 1988.

[9] Edmundson, H.P., "New Methods in Automatic Abstracting", *Journal of ACM*, Vol.16, No.2, pp.264–285, 1969

[10] Garvey, W.D. / 津田 良成 監訳, "コミュニケーション -科学の本質と図書館員の役割-", 敬文堂, 1979.

[11] Garfield, E., "Citation Indexes to Science: A New Dimension in Documentation Through the Association of Ideas", *Science*, No.122, pp.108–111, 1955.

[12] Goldscmidt, P.G., "Information Synthesis.: Practical Guide", *Health Services Research*, Vol.21, No.2, pp.214-237, 1986.

[13] Gross, P.L.K., Gross, E.M., "College Libraries and Chemical Education", *Science*, No.1713, pp.385–389, 1927. "大学図書館と化学教育", 竹内比呂也訳, 情報学基本論文集 I, 勁草書房, pp.151–158, 1989.

[14] 原田昌紀, "サーチエンジンにおける検索結果のランキング", *bit*, Vol.32, No.8, pp.8–14, 共立出版, 2000.

[15] Herlach, G., "Can Retrieval of Information from Citation Indexes be Simplified?: Multiple Mention of a Reference as a Characteristic of the Link between Cited and Citing Articles", *Journal of the American Society Information Science*, Vol.29, No.6, pp.308–310, 1978.

[16] Honda, T., Mochizuki, H., HO, T.B., and Okumura, M., "Generating Decision Trees from an Unbalanced Data Set", *Proceedings of the 9th European Conference on Machine Learning*, pp.68–77, 1997.

[17] 神門典子, 野末道子, 榛田倫子, 村上匡人, 谷津真理子, 上田修一, "情報検索分野の構造：引用調査による下位領域の発展過程の分析", *Library and Information Science*, No.29, pp.39–65, 1991.

[18] 神門典子, “原著論文の機能構造の分析とその応用 - C型肝炎論文を対象とした基本動向記述文の抽出とその前提としての構成要素カテゴリ自動付与の試み -”, 図書館学会年報, Vol.40, No.2, pp.49–61, 1994.

[19] Kando, N., “Text-level Structure: Implications for Information Retrieval and the Potential for Genre Analysis”, *British Computer Society IR SG Annual Colloquium*, 1997.

[20] 神門典子, “3. 全文検索を高度化する技術：情報検索とテキスト構造”, 「全文検索 技術と応用」, 学術情報センター編, pp.33–74, 1998.

[21] Kessler, M.M., “Bibliographic Coupling between Scientific Papers”, *American Documentation*, Vol.14, No.1, pp.10–25, 1963.

[22] Kita, K., Kato, Y.,Omoto, T., and Yano, Y., “A Comparative Study of Automatic Extraction of Collocation from Corpora: Mutual Information vs. Cost Criteria”, *Journal of Natural Language Processing*, Vol.1, No.1, pp.21–33, 1994.

[23] 窪田輝蔵, “科学を計る - ガーフィールドとインパクト・ファクタ -”, インターメディカル, 1996.

[24] Kupiec, J., Pedersen, J., Chen, F., “A Trainable Document Summarizer”, *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.68–73, 1995.

[25] Lawrence, S., Giles, L., Bollacker, K., “Digital Libraries and Autonomous Citation Indexing”, *IEEE Computer*, Vol. 32, No. 6, pp.67–71, 1999.

[26] Light, R.J., Pillemer, D.B., “Summing Up: the Science of Reviewing Research”, Cambridge, Harverd University Press, 1984.

[27] Lipetz, B., “Improvement of the Selectivity of Citation Indexes to Science Literature through Inclusion of Citation Relationship Indicators”, *American Documentation*, Vol.16, No.2, pp.81–90, 1965.

[28] Liu, M., "Progress In Documentation The Complexities of Citation Practice: A Review of Citation Studies", *Journal of Documentation*, Vol.49, No.4, pp.370–409, 1993.

[29] Mani, I., Bloedorn, E., "Multi-document Summarization by Graph Search and Matching", *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI'97)*, pp.622–628, 1997.

[30] Mani, I. and Bloedorn, E., "Machine Learning of Generic and User-focused Summarization", *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98))*, pp.821–826, 1998.

[31] 真弓育子, "文学研究における引用活動: シェークスピア研究を題材とした引用カテゴリー調査", *Library and Information Science*, No.22, pp.119–128, 1984.

[32] McCallum, A., Nigam, K., Rennie, J., and Seymore, K., "A Machine Learning Approach to Building Domain-Specific Search Engines", *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, pp.662–667, 1999.

[33] 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.

[34] 三池誠司, 住田一男, "文書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, FI-34-3, pp.17–24, 1994.

[35] Moravcsik, M.J., and Murugesan, P., "Some Results on the Function and Quality of Citations", *Social Studies of Science*, Vol. 5, pp.86–92, 1975.

[36] Mulrow, C.D. et al., "The Medical Review Article: State of the Science", *Annals of Internal Medicine*, Vol.106, pp.485-488, 1987.

[37] 村主千賀, 津田良成, "レビュー論文にみる知識の蓄積と統合:「臨床医の情報ニーズ・情報探索行動」に関するレビューの調査に基づいて", *Library and Information Science*, No.35, pp.1–40, 1996.

[38] 長尾真 編 "自然言語処理", 岩波講座ソフトウェア科学 15, 岩波書店, 1999.

[39] 永田昌明, 平博順, “テキスト分類 - 学習理論の「見本市」 -” 情報処理, Vol.42, No.1, pp.32–37, 2001.

[40] 中山茂, “歴史としての学問”, 中央公論社, 1974.

[41] 難波英嗣, “論文間の参照情報を考慮した学術論文要約システムの開発”, 北陸先端科学技術大学院大学 修士論文, 1998.

[42] Nanba, H. and Okumura, M., “Producing More Readable Extracts by Revising Them”, *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pp.1071–1075, 2000.

[43] Narin, F., Gabriel, P., Hofer, G.H., “Structure of the Biomedical Literature”, *Journal of the American Society for Information Science*, Vol.27, No.1, pp.25–45, 1976. “生物医学文献の構造”, 神門典子訳, 情報学基本論文集 I, 勁草書房, pp.189–227, 1989.

[44] Narin, F., Olivastro, D. and Stevens, K.A., “Bibliometrics / Theory, Practice and Problems”, *Evaluation Review*, Vol.18, No.1, pp.65–76, 1994.

[45] 奥村学, 難波英嗣, “テキスト自動要約に関する研究動向”, 自然言語処理, Vol.6, No.6, pp.1–26, 1999.

[46] 奥村学, 難波英嗣, “テキスト自動要約に関する最近の話題”, 北陸先端科学技術大学院大学 情報科学研究科 Research Report.IS-TM-2000-001, 2000.

[47] Oxman, A.D., Guyatt, G.H., “The Science of Reviewing Research”, *Annals of the New York Academy of Sciences*, Vol.703, pp.125–134, 1993.

[48] Page, L., “The PageRank Citation Ranking: Bringing Order to the Web”, *Proceedings of the ASIS Annual Meeting*, 1998.

[49] Paice, C.D., “Constructing Literature Abstracts by Computer: Techniques And Prospects”, *Information Processing & Management*, Vol.26, No.1, pp.171–186, 1990.

[50] Peritz, B.C., “A Classification of Citation Roles for the Social Sciences and Related Fields”, *Scientmetrics*, Vol.5, pp.303–312, 1983.

[51] Rull,I., "Citation Analysis of a Scientific Career: A Case Study", *Social Studies of Science*, Vol.9, pp.81–90, 1979.

[52] 齊藤陽子, "引用文献の記述形式の実態と基準", 書誌索引展望, Vol.17, No.4, 1993.

[53] Salton, G., McGill, M. J., "Introduction to Modern Information Retrieval", New York, McGraw-Hill, 1983.

[54] Small, H., "Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents", *Journal of the American Society for Information Science*, Vol.24, pp.265–269, 1973.

[55] Small, H., "Cited Documents as Concept Symbols", *Social Studies of Science*, Vol.8, pp.327–340, 1978.

[56] Spiegel-Rosing, I., "Bibliometric and Content Analysis", *Social Studies of Science*, Vol.7, pp.97–113, 1977.

[57] Swanson, R.W., "A Work Study of the Review Production", *Journal of the American Society for Information Science*, pp.70–72, 1976.

[58] Swanson, D.R., "Undiscovered Public Knowledge", *Library Quarterly*, Vol.56, pp.103–118, 1986.

[59] 田中穂積, "自然言語解析の基礎", 産業図書, 1989.

[60] 田中穂積 監修, "自然言語解析 – 基礎と応用 –", 電子情報通信学会, 1999.

[61] Teufel, S., "Argumentative Zoning: Information extraction from scientific text", *PhD thesis, University of Edinburgh*, 1999. http://www.cogsci.ed.ac.uk/ simone/t.ps

[62] trec_eval, "ftp://ftp.cs.cornell.edu/pub/smart".

[63] 津田良成 編, "図書館・情報学概論", 勁草書房, 第二版, 1990.

[64] 津田良成, 村主千賀, "レビュー論文における収録文献の選択：臨床医の情報ニーズ・情報探索行動に関する 3 つのレビュー論文の比較", *Library and Information Science*, No.32, pp.1–16, 1994.

[65] 津田良成, “「スター論文」の貢献: (2) 先ず読む論文の選択”, あいみっく, No.17, No.2, pp.32–40, 1996.

[66] 上田修一 他, “文献間の類似度を測定する尺度としての共引用の妥当性についての評価: 「情報学」関連文献を事例として”, 平成 2 年度慶応義塾大学文学部学事振興基金による研究 (共同研究) 「主題表現のマッピングの手法に関する研究」報告書, 慶応義塾大学文学部図書館・情報学科, 1991.

[67] 牛澤典子, “被引用文献の概念シンボル化 - 医学雑誌論文を事例として -”, *Library and Information Science*, No.30, pp.133–146, 1992.

[68] van Rijsbergen, “Information Retrieval (2nd Edition)”, *Butterworths*, London, 1979.

[69] Voos, H. and Daraev, K.S., “Are All Citations Equal?: or, Did We op, cit. Your Idem?”, *Journal of Academic Librarianship*, Vol.1, No.6, pp.19–21, 1976.

[70] 鷲崎誠司, 村本達也, “ハイパーリンクの構造を利用した検索結果の選択手法”, 情報処理学会研究報告 情報学基礎, 99-FI-55-10, pp.73–80, 1999.

[71] Weinstock N. , “Citation Indexes, in Kent A. (Ed.)”, *Encyclopedia of Library and Information Science*, New York, Marcel Dekker, Vol.5, pp.16–41, 1971.

[72] White, H.D. and Mccain, K.W., “Bibliometrics”, *Annual Review of Information Science and Technology (ARIST)*, Vol.24, pp.119–186, 1989.

[73] 山崎茂明, “インパクトファクターとは何か: 正しい理解と研究への生かし方”, 第 1 回北里大学図書館セミナー, 1998. http://mlib.kitasato-u.ac.jp/homepage/seminar1.html

参照情報の説明に用いた論文

[74] Bond, F., Ogura, K., and Ikehara, S. “Classifiers in Japanese-to-English Machine Translation”, *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pp.125–130, 1996. http://xxx.lanl.gov/ps/cmp-lg/9608014

[75] Bond, F., Ogura, K., Ikehara, S. 1994 . "Countability and Number in Japanese-to-English Machine Translation", *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, pp.32–38, 1994. http://xxx.lanl.gov/ps/cmp-lg/9511001

[76] Church, K., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of the Second Conference on Applied Natural Language Processing*, 1988.

[77] Heeman, P.A. and Allen, J.F. "Incorporating POS Tagging into Language Modeling", *Proceedings of Eurospeech'97*, 1997. http://xxx.lanl.gov/ps/cmp-lg/9705014

[78] Murata, M. and Nagao, M., "Determination of Referential Property and Number of Nouns in Japanese Sentences for Machine Translation into English", *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'93)*, pp.218–225, 1993. http://xxx.lanl.gov/ps/cmp-lg/9405019

[79] Scheler, G., "With Raised Eyebrows or the Eyebrows Raised ? A Neural Network Approach to Grammar Checking for Definiteness", *FKI-215-96*, 1996. http://xxx.lanl.gov/ps/cmp-lg/9606017

[80] Takeda, K., "Tricolor DAGs for Machine Translation", *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94)*, 1994. http://xxx.lanl.gov/ps/cmp-lg/9407008

PRESRI の説明に用いた論文

[81] Pereira, F. and Schabes, Y., "Inside-outside Reestimation from Partially Bracket Corpora", *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92)*, pp.128–135, 1992.

[82] Goodman, J., "Efficient Algorithms for Parsing the DOP Model", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996. http://xxx.lanl.gov/ps/cmp-lg/9604008

[83] Goodman, J., "Parsing Algorithms and Metrics", *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL'96)*, 1996. http://xxx.lanl.gov/ps/cmp-lg/9605036

[84] Carl de Marcken, "Unsupervised Language Acquisition ", *PhD thesis, MIT*, 1996. http://xxx.lanl.gov/ps/cmp-lg/9611002

[85] Carl de Marcken, "The Unsupervised Acquisition of a Lexicon from Continuous Speech", *MIT AI Memo*, No.1558/CBCL Memo, No.129, 1995. http://xxx.lanl.gov/ps/cmp-lg/9512002

[86] Stanley, F.C., "Bayesian Grammar Induction for Language Modeling", *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pp.228–235, 1995. http://xxx.lanl.gov/ps/cmp-lg/9504034

[87] Stanley, F.C., "Building Probabilistic Models for Natural Language", *Doctoral dissertation, Harvard University*, CRCT TR-02-96, 1996. http://xxx.lanl.gov/ps/cmp-lg/9606014

発表論文一覧

論文

1. 難波英嗣, 神門典子, 奥村学, “論文間の参照情報を考慮した関連論文の組織化”, 情報処理学会論文誌 (投稿中).

2. 難波英嗣, 奥村学, “論文間の参照情報を考慮したサーベイ論文作成支援システムの開発”, 自然言語処理, Vol. 6, No. 5, pp.43–62, 1999.

国際会議

1. Nanba, H., Kando, N., and Okumura, M., “Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation”, *Proceedings of the 11th ASIS & T SIG/CR Classification Research Workshop: Classification for User Support and Learning*, pp.117–134, 2000.

2. Nanba, H. and Okumura, M. “Producing More Readable Extracts by Revising Them”, *Proceedings of the 18th International Conference on Computational Linguistics*, pp.1071–1075, 2000.

3. Nanba, H. and Okumura, M. “Towards Multi-paper Summarization Using Reference Information”, *Proceedings of the 16th International Joint Conferences on Artificial Intelligence*, pp.926–931, 1999.

4. Okumura, M., Mochizuki, H., and Nanba H., “Query-biased Summarization Based on Lexical Chaining”, *Proceedings of Pacific Association for Computational Linguistics*

'99, pp.324-334, 1999.

その他の発表論文

1. 奥村学, 難波英嗣, “テキスト自動要約に関する最近の話題”, 北陸先端科学技術大学院大学 情報科学研究科 Research Report.IS-TM-2000-001, 2000.

2. 難波英嗣, 神門典子, 奥村学, “論文間の参照情報を考慮した関連論文の組織化”, 情報処理学会研究報告, 2000-NL-137, pp.94, 2000.

3. 難波英嗣, 奥村学, 神門典子, “論文間の参照情報を考慮した学術論文要約システムの開発”, *Synsophy* 第 *12* 回研究会, 1999.

4. 難波英嗣, 奥村学, “書き換えによる抄録の読みやすさの向上”, 情報処理学会研究報告, 99-NL-133, pp.53–60, 1999.

5. 奥村学, 難波英嗣, “テキスト自動要約に関する研究動向”, 自然言語処理, Vol. 6, No. 6, pp.1–26, 1999.

6. 奥村学，難波英嗣, “テキスト自動要約に関する研究動向”, 「自然言語処理と情報提示技術」講習会資料, 1999.

7. 難波英嗣, 奥村学, “論文間の参照情報を考慮した学術論文要約システムの開発”, 情報処理学会研究報告, 98-NL-127, pp.79–86, 1998.

8. 難波英嗣, 奥村学, “論文間の参照情報を考慮した学術論文要約システムの開発”, 言語処理学会第 4 回年次大会, pp.638–641, 1998.

9. 難波英嗣, 奥村学, “観点に基づいた新聞記事の重要文選択に関する心理実験と考察”, 言語処理学会第 4 回年次大会併設ワークショップ「テキスト要約の現状と課題」, pp.30–35, 1998.

10. 奥村学, 難波英嗣, “テキスト自動要約技術の現状と課題”, 北陸先端科学技術大学院大学 情報科学研究科 Research Report, IS-RR-98-0010I, 1998.