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

難波　英嗣[†]　奥村　学[†]

本稿では，データベースから関連する論文を自動的に収集し，人間が特定分野のサーベイ論文作成する作業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，論文の参照情報に着目する．論文の参照情報とは，論文中でその論文と参照先論文との関係について記述されている箇所 (参照箇所) から得られる情報のことで，参照先論文の重要点や，参照元と参照先論文間の相違点を明示する有用な情報が得られる．サーベイ論文作成には 2 つの処理 (1) 特定分野の論文の収集 (2) 論文間の相違点の検出が必要であると考えられる．本研究では参照情報を利用することでこれらの処理が部分的に実現可能であることを示す．具体的には，ある論文が他の論文を参照する時の参照の目的を，cue word を用いて解析し，論文の参照・被参照関係にリンク属性 (参照タイプ) を付与する．結果として，参照箇所抽出では Recall 79.6%，Precision 76.3%の精度が得られた．また，参照タイプ決定では 83%の精度が得られた．これらの参照タイプを利用し，ある特定分野の論文を自動的に収集するのに近い処理が可能になった．また，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アブストラクト，論文間の相違点の記述された参照箇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の支援に有用であることが示された．

キーワード：　複数論文の要約，参照関係，手がかり語

Towards Multi-paper Summarization Using Reference Information

Hidetsugu Nanba[†] and Manabu Okumura[†]

In this paper, we present a system to support writing a survey of the specific domain. In this system, we use reference information. Reference information includes the reference relationships between papers and the information which can be derived from the description around the citation, and be useful for understanding the difference between the referring and referred papers. To write a survey, at least two processes are necessary. One is to collect papers of some domain. Another is to make clear the differences between papers. We think the reference information is useful for these two processes. Firstly, we try to extract a fragment of texts where the author describes about a referring paper. We call the fragment “Reference Area”. Secondly, we attempt to analyze the purpose of reference. We divide that into three categories(we call these categories “Reference Types”), and develop the method to determine the type by using cue words. As a result, we got the recall of 79.6% and the precision of 76.3% in reference area extraction, and the accuracy of 83% in reference type decision. Making use of these reference types, we can collect a set of papers in the same domain. Finally, we build up a system to display the reference

graph of the papers. With the system, abstracts and reference areas of papers can be seen. Users of this system can easily collect papers of some specific domain, and also can understand the differences between the related papers.

KeyWords: *multi-paper summarization, reference relation, cue word*

1 はじめに

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 (Garvey 1979)．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．

我々はサーベイ論文を複数論文の要約と捉えており，サーベイ論文の自動作成の研究を行っている．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 (Garvey 1979)．しかし現在の自動要約の技術[1] 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箇所，論文間の相違点が明示されている箇所を抽出し，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．

本稿では，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，また論文中に参照先論文の重要箇所や，参照先論文との関係を記述した箇所 (以後，**参照箇所**) がある．この参照箇所を読むことで，著者がどのような目的で論文を参照したのか (以後, **参照タイプ**) や参照/被参照論文間の相違点が理解できる．論文の参照情報とは，このように論文間の参照・被参照関係だけでなく，参照箇所や参照タイプといった情報まで含めた物を指す．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．

本稿の構成は以下の通りである．2 章では，複数テキスト要約におけるポイントとサーベイ

† 北陸先端科学技術大学院大学 情報科学研究科, School of Information Science, Japan Advanced Institute of Science and Technology

1 近年の自動要約技術の動向に関しては (奥村，難波 1999) を参照されたい．

論文作成におけるポイントについて述べ，また関連研究を紹介する．3章では参照箇所と参照タイプについて説明する．また，参照箇所，参照タイプがサーベイ論文作成においてどのように利用できるかについて述べる． 4章では，3章で述べた考え方を基にしたサーベイ論文作成支援システムの実現方法について説明する． また，参照箇所の抽出手法，参照タイプの決定手法について述べる．5章ではそれらの手法を用いた実験結果を示す．6章では，作成したサーベイ論文作成支援システムの動作例を示す．

2　サーベイ論文作成

2.1　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，論文中の重要箇所を抽出する数多くの手法が提案されてきた (例えば (Edmundson 1969; Kupiec, Pedersen, Chen 1995; Teufel, Moens 1997; Mani, Bloedorn 1998))．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，冗長な箇所 (論文間の共通箇所) をどのように検出し削除するかが問題となる．一方，冗長な箇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，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，readability を上げるための処理が必要となる．従って，複数論文要約のポイントは図 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) 関連する複数論文からの情報の抽出
  - $(b)-1$ 重要箇所の抽出
  - $(b)-2$ 論文間の共通点の検出
  - $(b)-3$ 論文間の相違点の検出
- (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 1　複数論文要約のポイント

2.2　関連研究

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 (Kando 1997)(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 (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，Yamamotoら (Yamamoto, Masuyama, Naito 1995)，船坂ら (船坂, 山本, 増山 1996)，稲垣ら (稲垣，早川，田中 1998)，柴田ら (柴田, 上田, 池田 1997)，McKeownら (McKeown, Radev 1995)，Maniら (Mani, Bloedorn 1997) はいずれも，複数の新聞記事を対象に複数記事要約を試みている[2]．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 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．

3　サーベイ論文作成における参照情報の利用

3.1　参照箇所と参照タイプ

図 2の 5 文は (Bond, Ogura, Ikehara 1996) 中で (Murata, Nagao 1993) を参照している文の前後数文を抜粋したものである (Bond, Ogura, Ikehara 1996)，(Murata, Nagao 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) では，参照先論文 (Murata, Nagao 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られている．文 (3) では，参照先論文の問題点の指摘がなされている．そして文 (4) では，参照元論文 (Bond, Ogura, Ikehara 1996) がその問題点を考慮した論文であると述べている．

2 これらの論文のサーベイについては，(奥村，難波 1999) の 5 章を参照されたい

in (Bond, Ogura, Ikehara 1996)

(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, Nagao 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.

参照箇所 文 (2)〜(4)

図 2　type C の参照箇所

ここで，参照元論文 (Bond, Ogura, Ikehara 1996) と参照先論文 (Murata, Nagao 1993) の関係は文 (2)〜(4) を読めばわかる．このように参照元/参照先論文の関係が明示されている箇所を**参照箇所**と呼ぶ．参照箇所を読むことで参照元論文の参照の理由 (**参照タイプ**) や，参照元/参照先論文の関係が容易に理解できる．我々は，Weinstock の 15 種類の参照の理由 (Weinstock 1971) の分類を基に，参照タイプを次に示す 3 種類に分類する．

- **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 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 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)** とする．

我々は，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．type C の参照箇所からは，図 3のような情報が得られる．図 2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) は参照元論文の著者の観点から見た参照先論文の一種の要約であると考えられ，同時に参照元/参照先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箇所であると捉えることもできる．文 (2) では，参照元/参照先論文の両方が，

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

図 3　type C の参照箇所中の記述

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．このように，type C の参照箇所には論文間の共通点や相違点に関する事項が書かれているため，サーベイ論文作成に有用であると思われる．

3.2　サーベイ論文作成における参照情報の利用

3.2.1　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．我々は，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 の参照箇所中の “既存研究の紹介” の記述が参照元/参照先論文共通の問題点である」という仮定に基づいている．この仮定がどの程度正しいか，次章で説明する論文データベースを用いて調べた．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元/参照先論文 31 組のうち，29 組は参照元/参照先共に同じ分野の論文であることが確認された．図 4は，被要約対象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)

3.2.2　論文間の共通点，相違点の検出

type C の参照箇所から得られる情報 (図 3) と，複数論文要約のポイント (図 1) との関係について，(α) は (b)-1,2 に，(β) と (γ) は (b)-3 にそれぞれ対応している．従って，参照箇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には，(Bond, Ogura, Ikehara 1996) の (Murata, Nagao 1993) に関する参照箇所を示したが，図 5 に，(Murata, Nagao 1993) に関する (Bond, Ogura, Ikehara 1994) と (Takeda 1994) の参照箇所を示す．(Bond, Ogura, Ikehara 1996) 中の文 (1) は図 1の $(\alpha)(\beta)$ に，

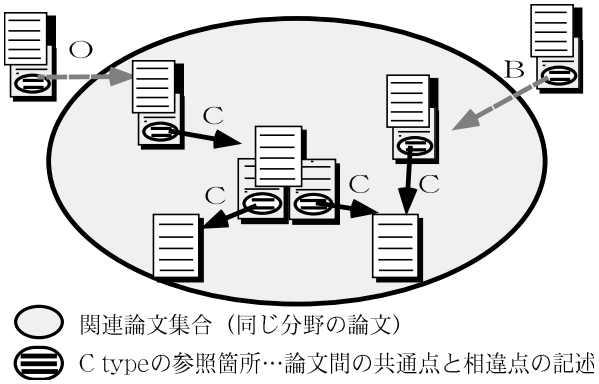

図 4　論文間の共通点と相違点

文 (2) は (γ) にそれぞれ対応する．また (Takeda 1994) 中の文 (1)(2) が (α) に，文 (3) が $(\beta)(\gamma)$ に対応する．2 つの論文の $(\alpha)(\beta)(\gamma)$ 同士を比較すれば，同じ (Murata, Nagao 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，ひとつの論文を参照する複数の論文中の参照箇所 (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．

4　サーベイ論文作成支援システム

前章で，ひとつの論文を他の複数の論文が参照する場合は，著者の観点毎にまとめる必要があることについて述べた．しかし同じ事項を述べる場合でも，論文の著者毎に用いる単語や文体等が異なるため，形態素同士の比較といった単純な手法では著者毎の観点の同一性は明らかにできない．また，著者の使用する単語や文体等の違いは，著者の観点の分類だけでなく，サーベイ論文生成時にも問題がある．論文間の共通点や相違点を検出して並べただけでは readability に欠けるため，ひとつのサーベイ論文として非常に読みづらくなると思われる．readability を向上するためには使用する単語や文体の統一といった処理を必要とするが，それには高度な言い替えの処理技術が必要になると考えられる．

そこで，本稿ではサーベイ論文の自動作成ではなく，サーベイ論文作成システム実現の第 1 歩として，関連論文を自動収集し，関連論文間の相違点や各論文の ABSTRACT が表示できるサーベイ論文作成支援システム作成を試みた．

4.1　論文間の参照・被参照関係の解析

サーベイ論文作成の対象として e-Print archive [3]という論文データベースの “The Computation and Language” の分野の論文の TEX ソース約 450 本を用いる．論文間の参照情報を利

3 http://xxx.lanl.gov/cmp-lg/

in (Bond, Ogura, Ikehara 1994)

(1)Recently, (Murata, Nagao 1993) 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.

in (Takeda 1994)

(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(Murata, Nagao 1993) 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 5　(Murata, Nagao 1993) に関する type C の参照箇所

用して要約を作成するには，まず要約対象となる論文データベース中の論文間の参照・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 450 本の TEX ソース間の参照関係が明らかにできる．

図 6は，TEX ファイル (Bond, Ogura, Ikehara 1996)(cmp-lg/9608014) の参考文献の記述の一部を抜粋したものである．(Bond, Ogura, Ikehara 1996) は論文中で (Murata, Nagao 1993) 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 7はそのリストの一部を抜粋したものである．(Bond, Ogura, Ikehara 1996)(cmp-lg/9608014) が (Murata, Nagao 1993)(cmp-lg/9405019) を参照しているという情報を得るには，図 6と図 7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箇所から単語 (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

```
\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 6　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 7　e-Print archive 論文リスト中の書誌情報の一例

問題となるが，参考文献の記述形式に着目する．齊藤ら (齊藤 1993) 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 6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 6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%の精度が得られた．

4.2　参照箇所の抽出

参照箇所の抽出とは，citation の出現する段落において，citation のある文と文間のつながりが強いと考えられる文を，citation 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) 接続詞 (3)1 人称代名詞 (4)3 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我々は考え，これらの 6 つの分類を考慮し，cue word を用いて参照箇所の抽出を試みた．cue word としてどのような語を用いるかは，人間が主観的に決める方法もあるが，本研究では，参照箇所コーパスの n−word gram 統計をとり半自動的に cue word を得ることを試みた．n−word gram 統計の結果を分類し

表 1　参照箇所抽出用 cue word の例

(1) 照応詞	In this, On this, Such
(2) 接続詞	But, However, Although
(3) 1人称	We, we, Our, our, us, I
(4) 3人称	They, they, Their, their, them
(5) 副詞	Furthermore, Additionally, Still
(6) その他	In particular, follow, For example

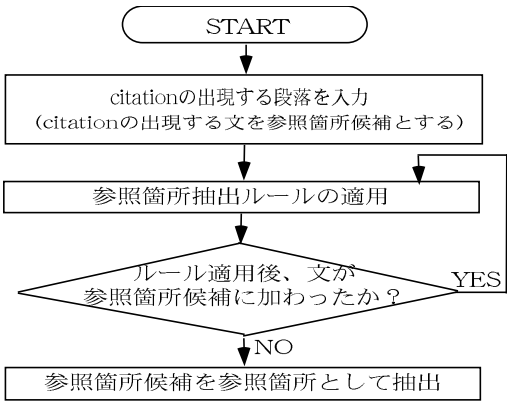

図 8　参照箇所抽出の手順

整理することで，最終的に 86 個の cue word を得た．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 1に cue word の例を示す．

次に，cue word を用いた参照箇所抽出の手順を図 8に示す．入力は，予め citation 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箇所抽出を行う．参照箇所抽出ルールとは，「参照箇所候補となる文の前後の文に cue word が出現すれば，その文も参照箇所候補に含める」といったものである．参照箇所抽出ルールの例を図 9 に示す．

図 9において，変数 $first_sentence とは参照箇所候補の最初の一文の文番号，$last_sentence は最後の一文の文番号を意味する．図 9に示すようなルールを 11 種類作成し，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箇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．調査の結果，11 種類のうちで 10 種類のルールを組み合わせた場合，参照箇所抽出精度が最も高くなり，この組み合わせで参照箇所抽出を行うことにした．

4.3　参照タイプの決定

参照箇所中で，例えば citation の後の文が “However” で始まるような場合，参照元論文の著者は参照先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，citation の

```
# cue word の設定
@this_cue=('For this','For these','On this','On these','In this',
           'In these','This','These');
@however_cue=('However','however','But','In spite of','in spite of'…);

# 照応詞に関する参照箇所抽出ルール
foreach $cue (@this_cue){  # ルール 1
   if($paragraph[$first_sentence]=~/$cue/){$first_sentence--}
}

# 接続詞に関する参照箇所抽出ルール
foreach $cue (@however_cue){ # ルール 2
   if($paragraph[$first_sentence]=~/$cue/){$first_sentence--}
}
foreach $cue (@however_cue){ # ルール 3
   if($paragraph[$last_sentence]=~/$cue/){$last_sentence++}
}
…
```

図 9　参照箇所抽出ルールの例

前に “We use” や “We adopt” といった語が出現する場合，参照元論文は参照先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 cue word list を作成し，次に cue word と citation 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，cue word 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 (Biber, Finegan 1994)．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，B type については，“Introduction”，“Experiment” の section に注目した．e-Print archive の論文約 450 本から section 毎に n−word gram をとり，次に cost criteria (Kita, Kato, Omoto, Yano 1994) を利用することで cue word 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，cue word として選びだした．cue word の一部を表 2，表 3に示す．

次に参照タイプ決定ルールについて説明する．参照タイプ決定は，表 2，表 3に示す cue word を用いてルールを作成した．参照タイプ決定には，本節の始めでも述べたような citation と cue word 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

表 2　type C 決定用 cue word の例

Although,	Though,	,although
However ,	however, their	however, the
but the	but it	But they
In spite of	Instead of	But instead
does not	did not	was not
should not	has not	were not
not require	not in effect	not provide
difficult to	more difficult	a difficult

表 3　type B 決定用 cue word の例

based mainly on	basis	is based on
the basic	used in	uses? of
used by	to use a	can use
that can	We can	We use
which can be	follow	useful for
available in	available for	applied to
the application of	application to	We adopted
extend the	extended to	For this

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 いずれのタイプも割り振られなかった参照箇所を type O とする．ルールは各 cue word 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 10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箇所を配列として，また配列中の citation の位置を入力値として受け取り，参照タイプ B，C，O を値として返す．

5　実験

5.1　参照箇所の抽出

前章で述べた手法の有効性を評価するため，参照箇所抽出実験を行った．評価は式 (1)(b=1) に示す F-measure(van Rijsbergen 1979) を用いて行う．

```
sub reference_type_decision($@){ # 参照タイプ決定ルーチン
  ($citeline,@ra)=@_; # $citeline : citationの位置
                      # @ra       : 参照箇所，1行1文のリスト
# type C 決定用ルール
  for($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 10　参照タイプ決定ルーチンの一部

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

ここで，P，Rは以下により計算される．

$$R(Recall) = \frac{\text{抽出された文のうち正解のものの数}}{\text{参照箇所コーパスの抽出すべき文の総数}} \tag{2}$$

$$P(Precision) = \frac{\text{抽出された文のうち正解のものの数}}{\left(\begin{array}{l}\text{参照箇所抽出ルールにより}\\ \text{抽出された文の総数}\end{array}\right)} \tag{3}$$

実験用データとして，citation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箇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，参照箇所は最大でもcitationの含まれる段落全体までとした．そのうちルール作成用を100個，評価用を50個とした．ルールについては4.2節で述べた通りである．また，ルール作成用データを用いて，参照箇所抽出用の11種類のルールの最適な組み合わせを得た．この組み合わせで評価用データに対して実験を行った．結果を表4に示す．

本手法の有効性を示すために，2つのベースラインを考慮した．citationの含まれる文のみを参照箇所として抽出した場合，その文は必ず参照箇所である．この時F-measureは0.575(Recall/Precision: 40.4/100.0 %)であった．一方，citationのある段落全体を参照箇所として抽出した場合，参照箇所として抽出されうる文はすべて含まれてしまう．この時F-measureは0.534(Recall/Precision: 100.0/36.4 %)であった．

表 4　参照箇所抽出精度

	Recall(%)	Precision(%)	F-measure
本手法 (ルール作成用)	90.9	76.9	0.833
本手法 (評価用)	79.6	76.3	0.779
ベースライン 1	40.4	100.0	0.575
ベースライン 2	100.0	36.4	0.534

表 5　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
	B	1	**105**	13	88.2
	O	3	8	**103**	90.3

参照タイプ決定精度 90.1(%)

表 4に示すように，本手法の F-measure 値は 2 つのベースラインの値を上回っており，従って参照箇所抽出手法の有用性が示されたと言える．

5.2　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箇所抽出と同様，Recall，Precision を用いた．式 (4)(5) は type C のタイプ決定精度の評価方法である．

$$Recall = \frac{\left(\begin{array}{l}\text{ルールを用いて } typeC\text{に決定された}\\ \text{参照箇所のうち正解の数}\end{array}\right)}{\text{参照箇所コーパス中の } typeC\text{参照の数}} \tag{4}$$

$$Precision = \frac{\left(\begin{array}{l}\text{ルールを用いて } typeC\text{に決定された}\\ \text{参照箇所のうち正解の数}\end{array}\right)}{\text{ルールを用いて } typeC\text{に決定された参照箇所の数}} \tag{5}$$

実験用データとして，参照箇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．ルール作成用データにおけるタイプ決定精度を表 5に，評価用データにおける精度を表 6に示す．

表 6　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
	B	2	**25**	5	78.1
	O	1	5	**46**	88.5

参照タイプ決定精度 83.0(%)

タイプ決定精度について考察する．過去の研究 (難波 1998) では cue word として uni-gram を多く用いていたが，今回は cue word 選定の際，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．例えばこれまでは “not” や “but” といった語を cue word として用いていたが，“ not only 〜 but also” のように “not” や “but” が明らかに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 cue word 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．一方で，cue word のような表層的な情報のみを用いたタイプ決定方法は精度的にほぼ限界に達していると思われ，これ以上の精度向上には意味処理等を行う必要があると考えられる．

6　サーベイ論文作成支援システムの構築

4 章に基づき，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 11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．以前の研究 (難波 1998) で作成した論文検索システム PRESRI(**P**aper **RE**trieval **S**ystem using **R**eference **I**nformation)[4]を利用して論文検索を行う．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，関連論文の参照箇所や ABSTRACT 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，前章の参照箇所抽出や参照タイプ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

4 http://galaga.jaist.ac.jp:8000/pub/tools/sum

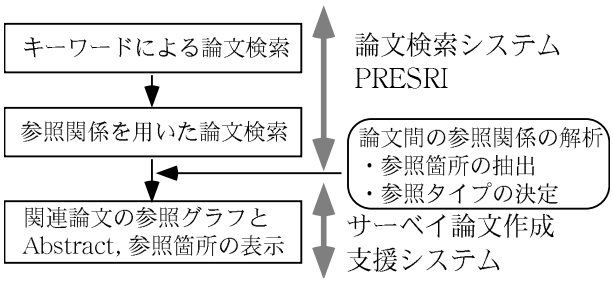

図 11　サーベイ論文作成支援の流れ

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・被参照関係だけを表示することで，実現可能であると考えられる．図 12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”(参照箇所) の箇所をクリックすることで，個々の論文の ABSTRACT や参照箇所を閲覧することが可能になる．図 12の右側のウィンドウは 3 本の論文 [Takeda94](9407008)，[Bond94](9511001)，[Bond96](9608014) の [Murata93](9405019) に関する参照箇所を示しており，左側ウィンドウのグラフ中の “REFERENCE AREA” の箇所をクリックした結果である[5]．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箇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．尚，このシステムは Perl で実装し，また CGI を用いることで World Wide Web 上からの利用が可能となっている．

7　おわりに

本研究では，関連する論文集合からのサーベイ論文自動作成を目指し，その第 1 歩としてサーベイ論文作成支援システムを構築した． 本研究では，複数の論文間の共通点，相違点を検出するために，論文間の参照情報に着目した．ある論文中の他の論文について記述してある箇所 (参照箇所) を論文中から自動的に抽出し，その箇所を解析することで，論文の参照の目的 (参照タイプ) が明らかにされる．

5 図中に “REFERENCE AREA” が 3 箇所あるが，いずれの箇所をクリックしても右側ウィンドウの表示になる

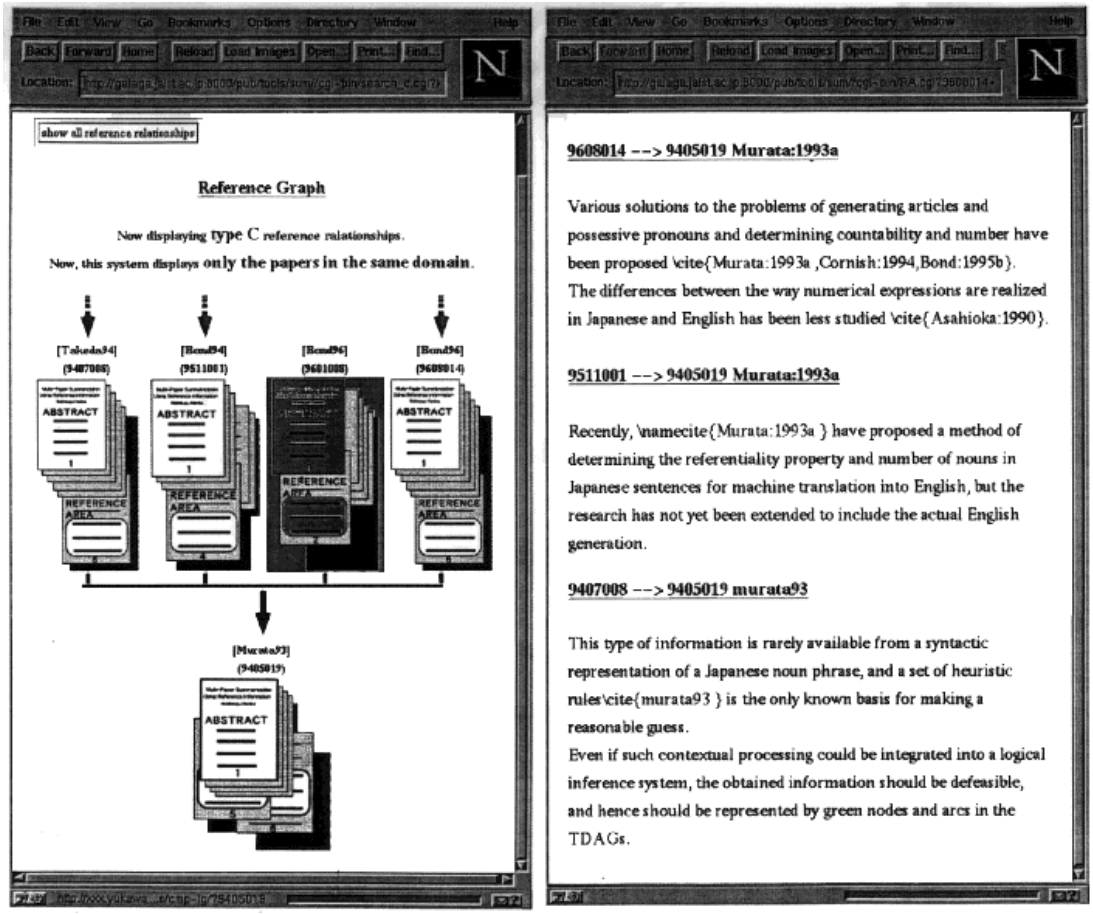

図 12　サーベイ論文作成支援システム

参照箇所の抽出と参照タイプの決定には，cue word を利用した．cue word の選定には (Kita, Kato, Omoto, Yano 1994) らの提唱する cost criteria という手法を利用し，得られた cue word を用いて参照箇所抽出ルールと参照タイプ決定ルールを作成した．その結果，参照箇所抽出精度は Recall，Precision 共に 80%弱，参照タイプ決定は 83%の精度が得られた．

また，サーベイ論文作成支援をする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 ABSTRACT 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箇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と考えられる．

謝辞

本研究にあたり，御指導を賜わりました学術情報センターの神門典子助教授に心から感謝致します．また，論文データの提供および論文検索システム 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.

参考文献

Biber,D., Finegan, E. (1994). "section13: Intra-textual variation within medical research articles." *Corpus-Based Research into Language*. Oostdijk & de Haan(eds.) Amsterdam, Rodoph. 201–221.

Edmundson, H.P. (1969). "New Methods in automatic abstracting." *Journal of ACM*, Vol.16, No.2, 264–285.

船坂貴浩, 山本和英, 増山繁 (1996). "冗長度削減による関連新聞記事の要約." 情報処理学会研究報告 自然言語処理, 97-NL-114-7, 39–46.

Garvey, W.D. /津田 良成 監訳 (1979). 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.

稲垣博人, 早川和宏, 田中一男 (1998). "類似意味内容の統合による伝達型電子化文書要約方式の提案." 情報処理学会 第 57 回 全国大会, (2), 153–154.

Kando, N. (1997). "Text-level Structure: Implications for Information Retrieval and the Potential for Genre Analysis." *British Computer Society IR SG Annual Colloquium*. (http://www.rd.nacsis.ac.jp/~kando/kando.ps).

Kita, K., Kato, Y.,Omoto, T., Yano, Y. (1994). "A Comparative Study of Automatic Extraction of Collocation from Corpora: Mutual Information vs. Cost Criteria." *Journal of Natural Language Processing*, 1(1), 21–33.

Kupiec, J., Pedersen, J., Chen, F. (1995). "A Trainable Document Summarizer." *SIGIR'95*, 68–73.

Mani, I., Bloedorn, E. (1997). "Multi-document Summarization by Graph Search and Matching." *AAAI'97*, 622–628.

Mani, I. and Bloedorn, E (1998). "Machine Learning of Generic and User-focused Summarization." *In Proc. of the 15th National Conference on Artificial Intelligence*, 821–826.

McKeown, K., Radev, D.R. (1995). "Generating Summaries of Multiple News Articles." *SIGIR'95*, 74–81.

難波英嗣 (1998). "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 (http://www.jaist.ac.jp/~nanba/study/master_thesis.ps)

奥村学, 難波英嗣 (1999). "テキスト自動要約に関する研究動向 (巻頭言に代えて)." 自然言語処理, 6(6), 1-26.

齊藤陽子 (1993). "引用文献の記述形式の実態と基準." 書誌索引展望, 17(4).

柴田昇吾, 上田隆也, 池田裕治 (1997). “複数文章の融合.”情報処理学会研究報告 自然言語処理, 97-NL-120-12, 77–82.

Teufel, S. and Moens, M. (1997). “Sentence extraction as a classification task.” Intelligent Scalable Text Summarization Proceeding of a Workshop ACL'97, 58–65.

塚田政嘉, 黒川恭一 (1998). “絞り込み用キーワードの抽出.”情報処理学会研究報告 自然言語処理, 98-NL-128-19, 135–141.

van Rijsbergen. (1979). Information Retrieval(2nd Edition). *Butterworths, London.*

Weinstock N. (1971). Citation indexes, in Kent A. (Ed.), *Encyclopedia of Library and Information Science, New York: Marcel Dekker*, Vol.5, 16–41.

Yamamoto, K., Masuyama, S,. Naito S. (1995). “An Empirical Study on Summarizing Multiple Texts of Japanese Newspaper Article.” *NLPRS'95*, 461–466.

参照情報の説明に用いた論文

Bond, F., Ogura, K., Ikehara, S. (1996). “Classifiers in Japanese-to-English Machine Translation.” *COLING'96* 125–130. (http://xxx.lanl.gov/ps/cmp-lg/9608014).

Bond, F., Ogura, K., Ikehara, S. (1994). “Countability and Number in Japanese-to-English Machine Translation.” *COLING'94*, 32–38. (http://xxx.lanl.gov/ps/cmp-lg/9511001).

Murata, M., Nagao, M. (1993). “Determination of referential property and number of nouns in Japanese sentences for machine translation into English.” *TMI-93*, (http://xxx.lanl.gov/ps/cmp-lg/9405019).

Takeda, K. (1994). “Tricolor DAGs for Machine Translation.” *In Proceedings of ACL'94* (http://xxx.lanl.gov/ps/cmp-lg/9407008).

略歴

難波 英嗣 (学生会員): 1972年生．1996年東京理科大学理工学部電気工学科卒業．1998年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．同年同大学院博士後期課程，現在に至る．自然言語処理，特にテキスト自動要約に関する研究に従事．情報処理学会，人工知能学会 各学生会員．

奥村 学 (正会員): 1962年生．1984年東京工業大学工学部情報工学科卒業．1989年同大学院博士課程修了．同年，東京工業大学工学部情報工学科助手．1992年北陸先端科学技術大学院大学助教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，語彙的知識獲得に関する研究に従事．情報処理学会，人工知能学会，AAAI，ACL，認知科学会，計量国語学会各会員．

(1998 年 11 月 5 日 受付)
(1999 年 2 月 8 日 再受付)
(1999 年 4 月 23 日 採録)