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

難波 英嗣, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科
e-mail:{nanba,oku}@jaist.ac.jp

1 はじめに

学術論文調査において, 学術論文データベースを利用するという方法がある. 例えば科学技術振興事業団(JST)[1]や学術情報センター(NACSIS)[2]等から, 様々なデータベースが利用可能である. また, 日外アソシエーツの雑誌記事索引ファイルで, 幅広い分野の国内の学術論文を調査することができる. 論文検索において, よく使用される属性は, 著者名, 標題, 情報内容を表現する語句・記号などであり, 近年では, 参照文献も重要な索引語として使用されるようになってきている.

ある分野の研究の動向調査をする場合, 調査の手順としてまず, 対象となる分野の研究論文を収集し, 次にその分野の論文の中で互いに類似するものをグループ化することが必要となる. 類似する文献を分類する方法は, 付与された索引語の共出現, あるいは標題や抄録中に出現する語によるクラスタ化等, 様々な方法がある. 他に共引用を利用するという手法がある. 共引用とは, 2つの文献が同一の文献に引用されている状態を示す. 一般に単純な参照関係が文献間の類似性を表現するとは言い切れないが, 類似性を示す尺度として妥当性が高いものと考えられる. 神門らは, 共引用分析を用いて情報検索の分野の論文について調査を行っている. 神門らは, 以下に示す式(1)を用いて論文間の類似度を算出し, これを元に共引用マップを作成・分析することで情報検索分野の研究調査をしている[1]. 本研究では, この論文間の参照関係というものに着目する.

文献AとBの類似度=

$$\frac{A\text{とB}\text{が共引用された回数}}{\sqrt{A\text{の被引用回数}\times B\text{の被引用回数}}} \quad (1)$$

本研究では, ある特定の研究分野に関する複数の学術論文の差異に注目し, 論文間の参照情報を考慮して複数の関連する論文との違いを明確にする要約を自動的に作成することを試みる. 論文間の関係を解析する際, 論文の参照情報に着目する. ある論文が他の論文を参照する場合, 参照論文について記述してある箇所(参照箇所)が存在する. その箇所を読むことで, 著者がどのような目的で参照しているのか明らかになる. このようにして参照箇所から得られる情報を参照情報と呼ぶ. 参照情報を収集し整理することで, ある分野の複数の論文間の関係が明らかになり, またそれらの参照情報が要約生成に利用できると考えられる.

[1]http://www.jst.go.jp/
[2]http://www.nacsis.ac.jp/nacsis.index.html

2 複数テキスト要約における参照情報の利用

図1は論文間の参照関係を示したモデルである. 図は要約対象の論文(**target papers**)3本と, それらを参照している2本の論文から構成されている. target papersに関する記述が図の上の2本の論文中, 参照箇所に記述されている. この参照箇所を解析することで, 2本の論文がそれぞれどのような目的でtarget papersを参照しているのかがわかる. いいかえれば, 参照の目的を明らかにすれば論文間の関係が明確になると言える.

参照目的を把握するためには, その前処理として論文中から参照箇所を抽出するという作業が必要となる. 参照箇所とはどういったものであるか, またどのように参照目的を分類すれば良いかを順次述べていく.

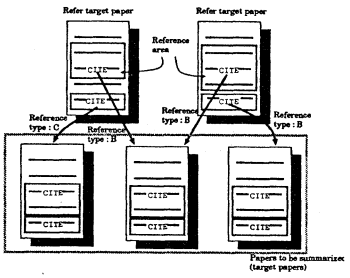

図 1: 論文間の参照関係

2.1 論文間の参照・被参照関係の解析

研究対象として e-Print archive[3] という論文データベース上の TEX ソース約 450 本を用いる. 論文間の参照情報を利用して要約を生成するには, まず要約対象となる論文ベースの参照・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.

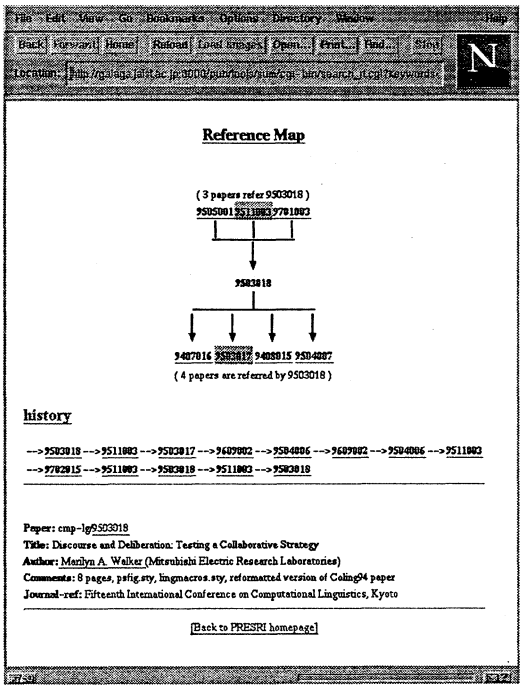

図 2: 参照関係による論文検索システム PRESRI

図 2は PRESRI(**P**aper **RE**trieval **S**ystem **U**sing **R**eference **I**nformation) というシステム[4]で, 論文間の参照関係のデータを用いて, 参照関係による論文検索が可能である.

2.2 参照箇所

参照箇所について [4] の論文を例に挙げて説明する。囲みの中の 5 文は [4] の中で [3] について記述された箇所である. 例えば, この例の場合, 文 1 に [3] のようなルールベースの tagging の研究がされている, といったことが書かれてある. 文 3,4 では [4] がルールベースの tagging の問題点を指摘している. この箇所から, [4] は [3] を既存の研究の問題点を指摘するために参照していることが分かる.

> **1. Recently, rule-based approaches are re-studied to cope with the limitations of statistical approaches by learning the tagging rules automatically from the corpus [Brill94].**
> **2. Some systems even perform the POS tagging as part of syntactic analysis process [Voutilainen95].**
> **3. However, the rule-based approaches alone are in general not robust to handle the unknown words , and is not flexible to adjust to the new tag-sets and languages.**
> **4. Also the performance is usually no better than the statistical counterparts [Brill94].**
> **5. To gain flexibility and robustness and also to overcome the limited window range of statistica l approaches, we need a method that can combine both statistical and rule-based approaches [Tapanainen94].**

このような箇所を読むことで, 著者がどういった目的でその論文を参照したのかがわかる. 本研究では, これを参照箇所 (reference area) と呼ぶ. 尚, 参照箇所抽出の際, 研究対象として TEX ファイルを使用するため, 論文を参照する際に用いる TEX の cite コマンドが利用できる.

参照箇所からどのような目的で他の論文を参照しているのか明らかにすることは複数論文の要約を生成する上で重要なタスクである. そこで, 参照の目的をいくつかに分類する. またこれらを参照タイプ (**reference type**) と呼ぶ.

2.3 参照タイプ

論文の参照タイプを以下の 3 つに分類した.

論説根拠型 (*Btype*)	ある理論を提案する場合や仮定をする場合, その根拠となる論文
問題点指摘型 (*Ctype*)	他の論文の理論や手法等の問題点を指摘する
その他型 (*Otype*)	*Btype* にも *Ctype* にも分類が難しい論文

参照箇所からこれらのタイプを決定することにより, ある分野の論文間の関係を明らかにする. 先の参照箇所の例の場合, 参照タイプは B type となる.

[3] http://xxx.lanl.gov/cmp-lg
[4] http://galaga.jaist.ac.jp:8000/pub/tools/sum/

3 論文間の参照関係解析の手法

3.1 参照箇所抽出

参照箇所抽出の際, 文間の結束性に着目した. それらの結束性は大まかに, (1) 照応詞 (2) 接続詞 (3)1 人称代名詞 (4)3 人称代名詞 (5) その他結束性のある語, の 5 つに分類される.

これらの結束性を考慮して, 7 クラス, 86 個の cue word list を作成した. cue word list は, 参照箇所コーパス 200 箇所から抽出した uni-,bi-,tri-gram を, 人手で分類整理して作成した. 7 クラスの cue word list を表 1に示す.

表 1: 7 クラスの cue word list

list name	a part of cue word
we.cue(9)	we,We,our,Our,us,I,my,My,me
this.cue(10)	In this, On this, In these, On these
base.cue(16)	base, basis, adopt, apply
and.cue(8)	and, furthermore, additionally,
but.cue(15)	however, but, In spite of
they.cue(18)	they, their, them, he, his, him
other_C(10)	difference between, different

cue word を用いて参照箇所抽出を試みた. TEX の cite コマンドの前後数文から, 参照している論文に関する記述箇所を抽出するというタスクで, 11 種類のルールを作成した. その一部を以下に示す. 基本的には, cite の前後の文に cue word が出現すれば, その文も参照箇所として抽出する. なお, ルールについて抽出過程のある段階で抽出の候補として選択されている文の中で一番最初にあるものを FIRST SENTENCE, 一番最後にあるものを LAST SENTENCE とする.

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

1-6 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.

1-7 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.

3.2 参照タイプ決定

抽出された参照箇所から, cue word を利用して表層的な解析を行い, 参照タイプの決定を試みる. タイプの決定には cite と cue word の並びを考慮して 12 種類のルールを作成した. ルールの一部を表 2に示す.

表 2: 参照タイプ決定のルール

No.	適用ルールの内容
B-2	cite の文に base.cue がある場合 B
B-3	cite の文に we.cue がある場合 B
C-1	cite の文に however.cue がある場合 C
C-3	cite の文以降 however.cue がある場合 C
B-7	cite の文より前に we.cue がある場合 B
B-8	cite の文より前に this.cue がある場合 B

4 実験

4.1 参照箇所抽出実験

評価を以下に示す Recall と Precision で行う.

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

$$Precision = \frac{\text{抽出された文のうち正解のものの数}}{\text{参照箇所抽出ルールにより抽出された文の総数}} \quad (3)$$

実験用コーパスとして 100 個の参照箇所, 評価用 50 個を用意した. まず, 実験用コーパスを用いて 11 種類のルールの組み合わせ 2^{11}通りの中で最も Recall, Precision の値が高くなるものを選んだ. その組み合わせで評価用コーパスを用いて実験を行った. 結果を表 3 に示す.

表 3: 評価用コーパスの参照箇所抽出精度

Recall	Precision
0.796	0.763

4.2 参照タイプ決定実験

参照タイプ決定実験の評価方法として以下のものを用いた.

$$Coverage = \frac{\text{ルールを用いて分類された参照箇所の数}}{\text{全ての参照箇所の数}} \quad (4)$$

$$Accuracy = \frac{\left(\begin{array}{c}\text{ルールでタイプが分類された}\\ \text{もののうち正解の数}\end{array}\right)}{\text{ルールを用いて分類された参照箇所の数}} \quad (5)$$

12 種類のルールを用いた参照タイプ決定の精度を表 4に示す.

表 4: 参照タイプ決定精度

参照タイプ決定精度	
Coverage	Accuracy
0.562	0.784

5 考察

5.1 参照箇所抽出

照応詞の問題に関して. cite コマンドの前後の文で this.cue に含まれる語が出現した場合, その前文に先行詞があると過程してルールを作成した. しかし, Paice が指摘しているように, 先行詞は前文ばかりでなく, 後方照応や同一文内に先行詞がある場合, あるいは先行詞が複数文や前の段落全体にわたるケースもある [2]. Paice は, この問題に対して多くのルールを作成して照応解析を試みているが十分な精度が得られていない. 本研究では, 参照箇所というテキスト中で極めて限定された箇所を文抽出の対象としているため, 照応問題について Paice ほど一般的に処理する必要はないものと考えられる. したがって, 今回作成したルールのように「this.cue の先行詞は前文である」という単純な仮定でも, 比較的良い抽出精度が得られたものと考えられる.

5.2 参照タイプ決定

本節では, 参照タイプ決定ルールの coverage について述べる. 実験で用いた参照箇所コーパスのうち 58.5% の参照タイプが決定可能となった. しかし, 残りのものについては決定できず, 現段階では要約生成処理の対象から外れる. ある分野の要約を作る時に, その分野の中心的な論文が存在する. 「中心的」の定義は色々あると思われるが, 定義のひとつに「参照される回数の多いものがその分野の中心的な論文である」という考え方がある. 本研究において, 参照関係の多いその分野で中心的と考えられる論文が, 参照タイプ決定ルールを用いてタイプ決定できなかった場合その論文が生成される要約からもれてしまい, ある特定分野の要約としてはあまりふさわしくないと考えられる. 本研究での要約対象論文は, 現在は e-Print archive 上のものだけであるが, 将来的には他のデータソースからの情報収集ということも考えている. coverage の低さは, 論文ソースの拡大によってある程度まではカバーできるが, 重要な論文が要約対象に成りうるかどうかという点で現在の手法を検討していく必要があるものと考えられる. また後でも述べるが要約の評価で, 要約対象となる論文の種類 (品質) という点についても今後考えていく必要がある.

6 結論

本研究では, 参照関係にある複数の論文からひとつの要約を生成するための前処理として, 論文間の関係を明らかにする手法を提案した.

参照箇所の抽出実験において, 論文の引用箇所の前後の文間の結束性を考慮して 11 種類のルールを作成した. このルールを用いることで, 人手で作成した評価用データにおいて約 80%の精度で参照箇所を抽出できるようになった. また, cue word (一種の手がかり語) の並びを考慮して, 12 種類のルールを作成した. このルールにより評価用のすべての参照箇所データ (200 参照箇所) のうち約 6 割はタイプ決定可能になった. タイプが決定できたもののうち, 約 80%は人が割り振った正解データと一致した.

これらの結果から, 複数の論文からひとつの要約を生成する際, 参照情報が利用できることが明らかになった.

7 今後の課題

今後の課題として, 参照タイプを用いて, 複数の target papers とそれらを共引用する論文の要約生成を試みる. 複数の論文をまとめる際に, 著者名から同一研究チームの論文であるかどうか判断する等, 参照情報以外の知識の利用についても検討していく.

参考文献

[1] 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. 1991.

[2] Chris D. Paice "Constructing Literature Abstracts by Computer: Techniques And Prospe cts". Information Processing & Management. Vol.26 No.1, pp. 171–186. 1990.

参照情報の説明に用いた論文

[3] E. Brill. "Some advances in transformation-based part-of-speech tagging". In Proceedings of the AAAI'94. 1994. (http://xxx.lanl.gov/ps/cmp-lg/9406010)

[4] Geunbae Lee, Jong-Hyeok Lee, Sanghyun Shin. "TAKTAG: Two-phase learning method for hybrid ". statistical/rule-based part-of-speech disambiguation". (http://xxx.lanl.gov/ps/cmp-lg/9504023).