情 報 学 基 礎 51－11
自 然 言 語 処 理 127－11
（1998. 9. 17）

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
Email:{nanba,oku}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

[概要]
本研究では, データベース中から関連する論文を自動的に収集し, 人間がサーベイを作成する作業を支援するシステムを提案する. 本研究では, サーベイ作成支援の際, 論文の参照情報に着目する. 論文の参照情報とは, 論文が参照・被参照関係にあるというだけでなく, どのような目的で参照しているのか, という情報まで含めた物を指す. このような情報が, 特定分野の論文の自動収集や論文間の関係を分析するのに有効であると考えている.

[キーワード] 複数テキストの要約, 参照関係, cue word

Multi-paper Summarization Using Reference Information

NANBA Hidetsugu, OKUMURA Manabu

School of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-1292 Japan)

Abstract

In this paper, we show a supporting system of making a survey. When we start a research in a domain, it is easy to grasp the outline of the domain if there exists a survey. However that kind of survey doesn't always exist. Therefore, we present a method to support human to make a summary for multiple papers. To support, we use the reference relationships between papers. Then, we built up a system using this reference information.

Key Words multi-paper summariztion , reference relation, cue word

1 はじめに

学術情報の情報量の増加は, 近代科学が興った17世紀以後, 一貫してみられる傾向であるが, 「情報の洪水」や「情報の爆発」という表現によって, 第二次大戦後, 特に注目されるようになった. その要因として, 研究者数の増加, 学問分野の専門分化が挙げられる. しかし, 専門分野がより狭く小さな主題に細分化されてきているにもかかわらず, 各専門分野の研究者が入手することの出来る文献の量は増加している [5]. 一方で, 情報量が増大しても, 個々の研究者が情報入手に費す時間は一定であるため, 研究に関連する全ての情報を入手し, 利用することが困難となる. また, 仮に全ての情報が入手できても, 処理能力に限界がある.

こういった状況で必要とされるものは, 研究に関連した情報の集成の評価, 統合, 簡略化等の付加的情報処理である. この情報濃縮過程が首尾一貫した総合的な文書の形をとったものが, サーベイ論文や専門図書である. 一方, 科学論文全体に対してサーベイ論文の占める割合は極端に少ないという指摘がある [1]. それは, サーベイ論文を作成するという作業が研究者にとって, 時間的にも労力的にも非常にコストを要するからである. しかし, 今後のさらなる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るものと思われる.

本研究では, 関連する論文をデータベース中から自動的に収集し, 人間がサーベイを作成する作業を支援するシステムを提案する. 本研究では, サーベイ作成支援の際, 論文の参照情報に着目する. 論文の参照情報とは, 論文が参照・被参照関係にあるというだけでなく, どのような目的で参照しているのか, という情報まで含めた物を指す. このような情報が, 特定分野の論文の自動収集や論文間の関係を分析するのに有効であると考えている.

本稿の構成は, 2章で, "参照箇所" と" 参照タイプ" について簡単な定義を行う. またここで, 本研究における論文間の共通点と相違点の考え方についても述べる. 3章では" 参照箇所" と" 参照タイプ" を, どのような形でサーベイ作成支援に利用するかについて説明する. また, 参照箇所の抽出方法, 参照タイプの決定の手法について述べる. 4章でそれらの手法の結果, 有効性を示す. 5章では, 今回作成したサーベイ作成支援システムを示す.

2 複数論文要約における参照情報の利用

2.1 参照箇所と参照タイプ

一般に論文は複数の論文を参照する. 図1は論文間の参照関係を示したモデルである. 参照元の論文中には, 参照先の論文や参照元と参照先の論文の関係について記述された箇所が存在する. こういった箇所を読むことで, 参照の目的がわかる.

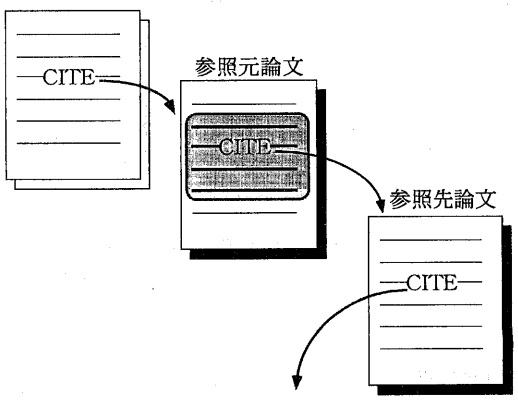

図 1: 論文間の参照関係

論文中で他の参照文献について記述された箇所の例を図2に示す. 図2の囲みの中の5文は [7] の論文中の [6] に関する記述である. この例の場合, 文 (1) に [6] のようなルールベースの tagging の研究がされている, といった記述がある. 文 (3),(4) では [7] 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, [7] は [6] 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. さらに文 (5) で, 文 (3)(4) で指摘した問題点についてどのような手法が必要とされるのか, ということが記述してある. このように参照元の論文 [7] と参照先の論文 [6] の関係について記述された箇所を本研究では参照箇所と呼ぶ.

(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 statistical approaches, we need a method that can combine both statistical and rule-based approaches [Tapanainen94].

図 2: C type の参照箇所

我々は参照箇所を解析することで, 参照の目的 (以下, 参照タイプと呼ぶ) が明らかにできると考えている. この参照の目的を本研究では**参照タイプ**と呼ぶ. 参照タイプを以下の 3 種類に分類した. 各タイプの詳細な説明については [3] を参照されたい.

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

先ほどの図 2の例の場合, C type の参照になる.

2.2 論文間の共通点と相違点

複数の論文をまとめてひとつの要約を作成する際, 論文間の相違点と共通点を明らかにすることが重要である. 本研究では 3 種類の参照タイプ B, C, O の中でも, 特に C type の参照に着目している. 論文間の差異に関する情報は C type の参照箇所から得られると考えている. 先ほどの [7] の例では, 文 (3),(4) で [6] の研究の問題点を指摘し, 文 (5) でその問題に対し必要とされる手法について述べている. これは, [7] と [6] の相違点であると言える.

また, 論文間の共通点についても, この C type の参照が利用できると考えている. 論文の参照には様々な目的があり, 単純に参照・被参照の関係だけで論文を収集すると論文集合中に複数のジャンルのものが含まれる可能性がある.

一方, 既存の研究の問題点を指摘するような C type の参照であれば, 参照元の論文も参照先も同じ分野である可能性が極めて高いのではないか, と考えている. 実際に C type の参照について 32 の参照箇所参照元の論文と参照先の分野を調べたところ, 94%(29 参照箇所) のものが同じ文やであることがわかった. 従って C type の参照関係をたどれば自動的に同一分野の論文が収集できると考えられる.

3 サーベイ作成支援システム

3.1 論文間の参照・被参照関係の解析

対象テキストとして E-Print archive [1] という論文データベースの"The Computation and Language" 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.

この参照・被参照関係を利用した検索システム PRESRI(**P**aper **RE**trieval **S**ystem using **R**eference **I**nformation)[2]を構築した. PRESRI では, 2 種類の検索機能を提供している. 1 つはキーワード検索機能で,

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

著者名, タイトル語から検索する. その結果が一覧表示される. 表示された論文リストの中で, 参照・被参照関係にある論文が E-Print archive 中に存在するものについては, 続けて参照関係を用いた検索が可能である.

3.2 参照関係の解析

サーベイ作成には, 大まかに (i) 特定分野の文献の収集, (ii) 収集された文献の分析, (iii) 分析結果に基づいた文書の生成, という 3 ステップが必要になる.

(i) 特定分野の文献の収集について. 収集の基本的な考え方はすでに 2.2 節で述べた. 参照タイプを考慮して論文収集を行うためには, E-Print archive 中の論文のすべての参照タイプを決定しなければならない. その前処理として個々の論文中から参照箇所を抽出する作業が必要になる.

3.2.1 参照箇所の抽出

参照箇所抽出の際, 文間の結束性に注目した. 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.

参照箇所抽出のための cue word list を作成した. cue word list は, 人手で作成した参照箇所コーパスから n-word gram をとり, それらを分類整理した. 表 1 に参照箇所抽出用 cue word list の一部を示す.

表 1: 参照箇所抽出用 cue word の一部

(1) 照応詞	In this	On this	Such
(2) 接続詞	But	However	Although
(3)1 人称	We	Our	us
(4)3 人称	They	Their	them
(5) その他	Furthermore	Additionally	Still

また cue word を用いた参照箇所抽出のルールを 11 種類作成した. 参照箇所抽出ルールは論文中の citation の前後の文に参照箇所抽出用 cue word が出現すれば, その文までを参照箇所候補とし, 参照箇所候補が変化しなくなるまで再帰的にルールを適用する.

3.2.2 参照タイプの決定

3 種類の参照タイプのうち, 特に C type の参照に注目した.

C type の参照は, 論文中の"Introduction", "Related Work", "Discussion" に多く出現すると考えており, E-Print archive の論文約 450 本において, これらの 3 種類の section の n-word gram をとり, Cost Criteria [2] を用いて cue word の候補のリストを作成した. その中から C type 決定用の cue word を 70 個選択した. B type については, "Introduction", "Experiment" の section から同様の手法で cue word 候補のリストを作成し, B type 決定用の cue word を 52 個選択した. 今回用いた cue word の一部を表 2 , 表 3に示す.

表 2: C type 決定用 cue word の一部

Although,	Though,	,although
However ,	however, their	however, the
but the	but it	[Bb]ut they
In spite of	Instead of	[Bb]ut 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: B type 決定用 cue word の一部

based mainly on	basis	is based on
the basic	used in	uses? of
used by	to use a	can use
that can	[Ww]e can	[Ww]e use
which can be	[Ff]ollow	useful for
available in	available for	applied to
the application of	application to	[Ww]e adopted
extend the	extended to	[Ff]or this

これらの cue word を用いた参照タイプ決定ルールを作成した. ルールは大まかに以下の 2 種類に分けることができる.

- 参照箇所中でcitationの出現する文以降にC type用のcue wordが出現すればC type
- 参照箇所中でcitationの出現する文以前にB type用のcue wordが出現すればB type

参照タイプ決定ルールの一部を図3に示す.

C type 決定ルール

- citationの出現した文の次の文に"However, "が現れたらC type
- citationの出現した文の2文後に"However, "が現れたらC type
- citationの出現した文の3文後に"However, "が現れたらC type
- …

B type 決定ルール

- citationの出現した文の2文前に"based mainly on, "が現れたらB type
- citationの出現した文の1文前に"based mainly on, "が現れたらB type
- citationの出現した文に"based mainly on, "が現れたらB type …

図 3: 参照タイプ決定ルール

参照タイプを決定するためのルールが, 全部で約470ステップある. システムに参照箇所を入力してルールの条件文とマッチした時点で参照タイプ決定処理を修了する. いずれのルールにもマッチしなかったものは参照タイプO typeを返す.

4 実験

4.1 参照箇所の抽出

評価を以下に示すRecallとPrecisionで行う.

$$Recall = \frac{\text{抽出された文のうち正解のものの数}}{\text{参照箇所コーパスの抽出すべき文の総数}} \quad (1)$$

$$Precision = \frac{\text{抽出された文のうち正解のものの数}}{\left(\begin{array}{l}\text{参照箇所抽出ルールにより}\\ \text{抽出された文の総数}\end{array}\right)} \quad (2)$$

実験用データとして100個の参照箇所, 評価用として50個を用意した. まず, 実験用データを用いて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4に示す.

表 4: 評価用コーパスの参照箇所抽出精度

	Recall	Precision
訓練用	0.909	0.769
評価用	0.796	0.763

4.2 参照タイプの決定

参照タイプ決定実験の評価方法としてRecall, Precisionを用いた. 以下はC typeのタイプ分類精度の評価方法である.

$$Recall = \frac{\left(\begin{array}{l}\text{ルールを用いて } Ctype \text{ に分類された}\\ \text{参照箇所のうち正解の数}\end{array}\right)}{\text{参照箇所コーパス中の } Ctype \text{ 参照の数}} \quad (3)$$

$$Precision = \frac{\left(\begin{array}{l}\text{ルールを用いて } Ctype \text{ に分類された}\\ \text{参照箇所のうち正解の数}\end{array}\right)}{\text{ルールを用いて分類された参照箇所の数}} \quad (4)$$

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.

表 5: C type の分類精度

	Recall	Precision
訓練用	0.813	0.867
評価用	0.818	0.750

表 6: B type の分類精度

	Recall	Precision
訓練用	0.686	0.772
評価用	0.825	0.647

現在, B type 分類用の cue word の選定及びルール作成の途中であるため, B type の分類精度は十分なものが得られていない (表 6) が, C type 分類に関しては (表 5) 十分な精度が得られていると言える.

5 サーベイ作成支援システムの構築

論文検索システム PRESRI を改良し, サーベイ作成支援システム (図 4) を作成した. システムは Perl で実装し, また, データは LaTeX の論文ファイルを jlatex2html で html 化して利用した.

システムは 2 つのウィンドウから構成される. ひとつは論文間の参照関係のグラフを示したウィンドウで, もうひとつは論文のアブストラクト, 参照箇所を表示するウィンドウである. 参照グラフウィンドウ中の"ABSTRACT" や"REFERENCE AREA"(参照箇所) をクリックすることで, もう 1 枚のウィンドウに論文のアブストラクトや参照箇所が表示される.

このシステムでは, C type の参照関係をたどることで, 関連する文献のアブストラクトや参照箇所を見ることができる. また, 切替えスイッチにより, C type の参照だけでなくすべての参照関係を見ることも可能である.

6 考察

参照タイプの決定について, これまで cue word として uni-gram をいくつか用いていたが [3], 今回はほとんど用いていない. それは uni-gram が参照タイプの決定の際, ノイズの要因となる可能性が高いためである. 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についても C type に判定されていた. そこで今回は bi-gram 以上 10-gram までのものの n-word gram をとり, Cost Criteria で cue word の候補を獲得した. cue word の候補から実際に用いる cue word を決定するのは人手で決めざるを得ないが, 例えば, "In spite of" の場合, その sub-sequence となる" In spite", "spite of", "In", "spite", "of" といった文字列の頻度は" In spite of" の頻度よりも大きいため, 単純な n-gram 統計では sub-sequence が" In spite of" よりも上位にランクされてしまう. 一方 Cost Criteria では"In spite" や" spite of" といった sub sequence の頻度は" In spite of" の頻度から引かれるため, コスト計算を行った結果も"In spite of" が他の sub-sequence に比べ, かなり上位にランクされており, cue word を選択する上でも効率的であった. また, 先の例で挙げた "not" について, 今回は cue word として" not"1 語でなく, 助動詞+"not"(e.g. do not, did not , can not, could not etc.) や be 動詞+"not" を cue word として選択することにより, "not only" のような場合の参照タイプ決定ミスがほとんどなくなった.

7 結論

本研究では, 関連する文献集合からサーベイを作成するための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, 相違点を明らかにするために, 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) が明らかにできることを示した.

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 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用ルールと参照タイプ決定ルールを作成した. 実験の結果, cue word を利用することで参照タイプを決定するのに十分な精度が得られることがわかった.

また, 論文データベース中から特定分野の論文を収集する際, 論文の参照タイプを考慮し, 参照箇所, 及び論文のアブストラクトをユーザに提示するという形でサーベイ作成支援をするシステムを作成した.

8 今後の課題

将来的には, サーベイ作成支援という形ではなく, サーベイを生成するシステムを構築したいと考えてい

る. しかしそれには非常に数多くの問題を解決しなければならない.

- Readability の問題. 参照箇所中の we, our, us や they their them は, 著者名に置換しなければならない.
- 参照タイプ C type について, 既存研究の問題点の指摘と一口にいっても, 例えばアルゴリズムの問題, 実装方法の問題等, 様々である. そこで, 現在用いている type に新たに sub type を設定する.
- 参照元の論文の参照箇所を示すだけでなく, 参照先の論文中でそこに対応する箇所を抽出する.

参考文献

[1] William D. Garvey/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.

[2] Kenji Kita, Yasuhiko Kato, Takashi Omoto, Yoneo Yano. "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

[3] 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.

[5] 津田 良成. "図書館・情報学概論 第二版" . 勁草書房

参照情報の説明に用いた論文

[6] 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)

[7] 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).

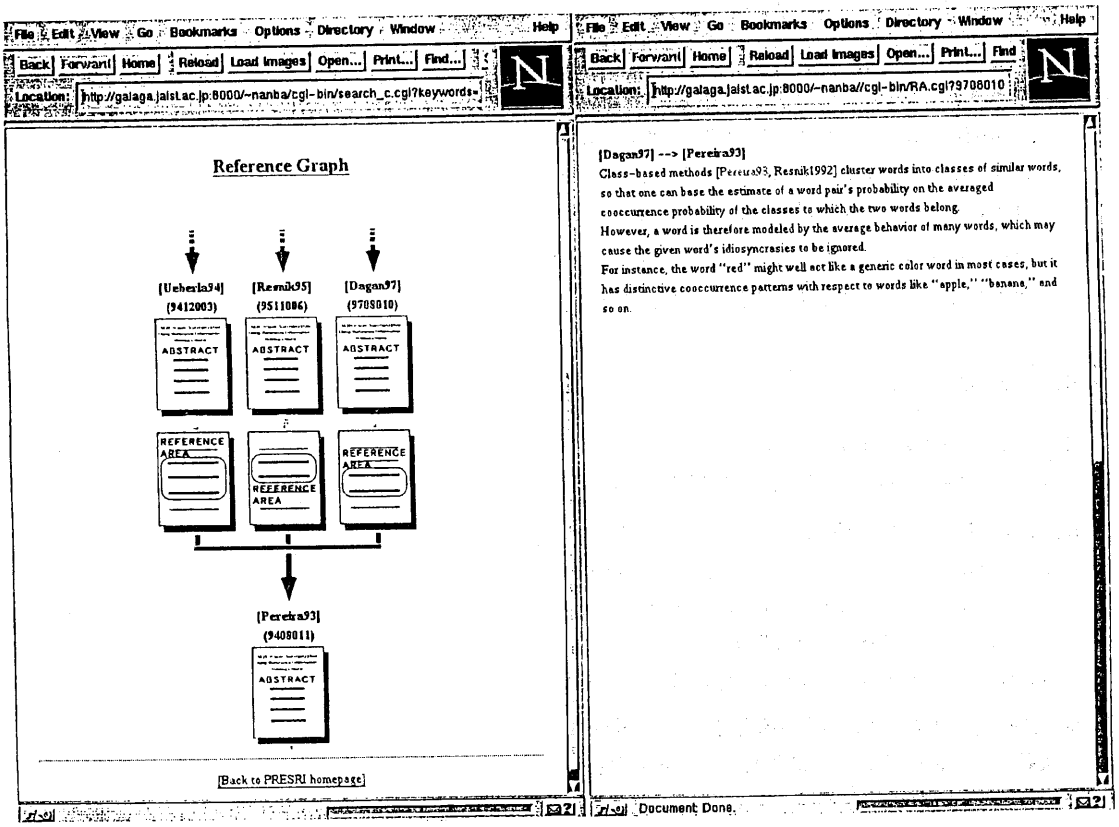

図 4: サーベイ作成支援システム