学術論文データベースからの研究動向情報の抽出と可視化

難波英嗣[1]，谷口裕子[2]
1．広島市立大学　情報科学部
2．広島市立大学　情報科学研究科

1. はじめに

ある研究分野において，「どのような要素技術がいつ頃から使われているのか」という情報を網羅的に収集し整理することは，その分野の研究動向を概観するのに必要不可欠であるが，その作業には多くの時間と労力を要する．そこで，このような情報を学術論文データベースから自動的に抽出し，可視化するシステムの構築を目指す．

技術動向情報を抽出し，可視化するには，まず特定の分野の論文を収集し，次にそこから可視化に必要な情報を抽出する，という 2 つのステップが必要となる．本研究では，ステップ 1 は論文間の引用情報を，ステップ 2 は論文表題の解析技術を，それぞれ用いる．論文間の引用情報とは，論文間の引用・被引用関係だけでなく，ある論文が引用論文をどのような理由で引用しているのか(引用タイプ)も含めた情報のことである[難波 1999]．ステップ 1 では，キーワード検索により特定分野の論文を収集するが，この時，同時に引用情報も考慮することで，キーワードを含んでいない当該分野の論文も収集する．ステップ 2 では，収集された論文の表題から要素技術に関する情報を抽出する．多くの論文表題には「A に基づいた」や「B を用いた」などの表現が含まれる．この A や B には，ある技術を実現するための要素技術を示す用語が一般的に含まれている．そこで，論文表題を解析し，専門用語抽出器を用いて A や B から用語を抽出する．用語を抽出した論文の著作年を X 軸に，抽出された用語を Y 軸にとることで，ある分野の動向を示すグラフを作成することができる．

本論文の構成は以下のとおりである．次節では，関連研究について述べる．3 節では，技術動向情報の抽出手法を提案する．4 節では，提案手法の有効性を調べるために行った実験について述べる．5 節では，システムの動作例を示す．

2. 関連研究

近年，ある用語に関連する用語をテキスト集合から自動的に収集する研究が活発に行われている[難波 2005,小原 2004,佐藤 2003,白井 2004]．しかし，これらの研究では収集された用語が与えられた用語とどのように関係にあるのか(例えば，上位下位関係)については扱っていない．本研究では，ある用語が要素技術用語であるかどうかを自動判定する点が，これまでの研究と異なる．

技術動向情報の抽出に関して，特許を対象にした数多くのシステムがこれまでに開発されている．例えば，富士通が開発している特許業務支援システム ATMS は，その一例である．このシステムでは，関連特許をクラスタリングして提示したり，特許間の類似性と引用関係に基づいて特許フロー(流れ図)を提示したりすることが可能である．しかし，学術論文を対象にした研究やシステムの開発は，著者が知る限りこれまでに行われていない．そこで，本研究では，学術論文を対象に，技術動向の抽出と可視化を行う．

3. 技術動向情報の抽出

本節では，技術動向情報の抽出手法について述べる．抽出手法は「特定分野の論文の収集」と「収集した論文からの要素技術用語の抽出」という 2 つのステップから構成される．ステップ 1 について 3.1 節で，ステップ 2 について 3.2 節で，それぞれ述べる．

3.1 特定分野の論文の収集

本節では，まず，論文収集に用いる引用情報について説明し，次に，引用情報を用いた論文の収集方法について述べる．

学術論文中には，当該論文と被引用論文との関係について記述されている個所(引用個所)がある．引用個所から得られる情報を，本研究では引用情報と呼んでいる．引用個所からは，被引用論文の重要点や当該論文と被引用論文との相違点を明示する有用な情報が得られる．また，引用個所を読めば引用の理由が分かる．

本研究では，引用の理由を引用タイプとして以下の 3 種類に分類し，また，引用タイプの決定を自動的に行っている[難波 1999]．

- type C (問題点指摘型)
  他の論文の理論や手法等の問題点を指摘するための引用．
- type B (論説根拠型)
  既存の研究成果を用いて，新しい理論を提案したり，システムを構築したりする場合の引用．
- type O (その他型)
  type B にも type C にも当てはまらない引用．

これらの中で，type C で引用・被引用関係にある 2 論文のトピックはほぼ同一であることが分かっている[難波 1999]．そこで，次に述べる手順で特定分野の論文を収集する．まず，キーワード検索で論文データベースから特定トピックの論文を

収集し，次に，それらと type C の引用・被引用関係にある論文も収集する．こうして収集された論文集合の表題から，次節で述べる手法で要素技術に関する用語を抽出する．

3.2. 論文表題からの要素技術用語の抽出

これまでに，論文表題の分析[千田 2005]や解析[長尾 1982, 佐藤 1999]に関する研究がいくつか行われてきたが，本研究では先行研究を参考に，手がかり語を用いて論文表題を解析する．

以下に示す 2 つの例を用いて解析手順を説明する．まず，あらかじめ表 1 に示す手がかり語と構造タグの対応リストを用意しておく．次に，論文表題と表 1 の手がかり語を比較し，表題中で一致する文字列を構造タグに置き換える．例えば，以下の解析例 1 において，論文表題(1)には下線で示す 2 種類の手がかり語が含まれているため，これらを構造タグに置き換えると(2)が生成される．(2)において，“GOAL”タグの直前の文字列(字幕生成)を“GOAL”タグで，“RESTRICT”タグの直前の文字列(ニュース番組)を“RESTRICT”タグで挟むことで(3)を得る．これを解析結果として出力する．

[解析例 1]

(1)「ニュース番組における字幕生成のための文短縮」

↓

(2)ニュース番組<RESTRICT cue=“における”>字幕生成<GOAL cue=“ための”>文短縮

↓

(3)<RESTRICT cue=“における”>
ニュース番組</RESTRICT>
<GOAL cue=“ための”>字幕生成</GOAL>
文短縮

[解析例 2]

(1) “Sentence extraction using support vector machine”

↓

(2) Sentence extraction <METHOD cue=“using”> support vector machine

↓

(3) Sentence extraction
<METHOD cue=“using”>support vector machine</METHOD>

英文表題の場合も同様であるが，(2)から(3)を生成する時に，タグの直前ではなく，直後の文字列を挟む点だけが異なる．

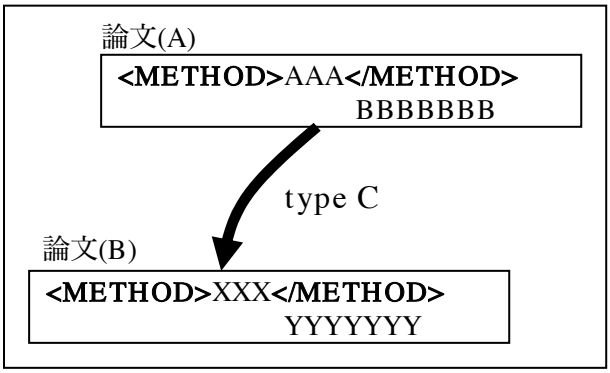
なお，論文表題の構造解析を行うため，英文表題用に 31 個，日本語用に 165 個の手がかり語を使用した．また名詞句間の関係を表すタグは英語用に 11 種類，日本語用に 10 種類設定した．

表 1 表題解析用手がかり語と構造タグの例

構造タグ	手がかり語(英，日)	
METHOD	by based on using	による に基づく を用いた
RESTRICT	in of on	における の に関する
GOAL	for towards	のための に向けて
CONJ	and or	と，や 及び

これらの解析結果から，解析例 2 の場合“support vector machine”が“sentence extraction”の要素技術であることがわかる．また，解析例 1 の場合「文短縮」の目的が「字幕生成」であることがわかるが，この関係は，別の観点から見ると「文短縮」が「字幕生成」の要素技術であると捉えることもできる．そこで，表 1 に示すタグのうち，“METHOD”と“GOAL”に着目し，要素技術用語を抽出する．

次に，表題解析と引用情報を組み合わせた抽出方法について，以下の例を用いて説明する．今，用語「YYYYYYY」の要素技術を収集する場合を考える．この場合，論文(B)からは「XXX」が抽出される．ここで，もし，論文(B)と type C で引用・被引用関係にある論文(A)が存在すれば，論文(A)中で“METHOD”タグが付与されている用語「AAA」も「YYYYYYY」の要素技術用語として抽出する．その理由は，3.1 節でも述べたとおり，type C で引用・被引用関係にある論文のトピックはほぼ同一であると考えられるからである．



4. 実験

3 節で述べた手法の有効性を調べるために実験を行った．

4.1. 実験に用いるデータ

引用論文データベース

実験には，PostscriptおよびPDF形式のフルテキスト論文約 12,000 件を用いる．これらのうち，

約 8,000 件はACLが提供するACL Anthology[1]に含まれるもの，残りの 4,000 件は，国内外の自然言語処理研究者や自然言語処理系研究室のWebページから収集したものや，自然言語処理関連の国際会議の予稿集(CD-ROM)から抽出した論文データから構成されている．

これらのデータから論文データベースを構築する．まず各論文データから，その論文の書誌情報(タイトル，著者名，所属，キーワード，アブストラクト)と，参考文献が抽出される[阿辺川 2003]．次に引用タイプの自動判定が行われる[難波 1999]．最後に抽出された書誌情報間で同定処理を行い，すべての書誌情報および引用情報をデータベースに格納する．結果として，約 58,000 件の書誌情報を含む引用論文データベースが構築された．

専門用語リスト

論文表題からの専門用語の抽出は，表題構造解析を行った後，専門用語リストとの照合により行う．この処理は，3.2 節の解析例 1 や 2 を見る限りでは，表題構造解析により分割された文字列を専門用語と見なせば不要のように思われるかもしれない．しかしこの方法では，例えば「統計的手法に基づく機械翻訳方式」は「<METHOD cue="に基づく">統計的手法</METHOD>機械翻訳方式」と解析されるため，「機械翻訳方式」の要素技術が「統計的手法」となる．その場合，「機械翻訳」の要素技術としてこの表題からは「統計的手法」が抽出されないことになる．そこで，専門用語リストとの照合というステップを入れる．

専門用語リストの作成を行うため，中川らの開発したTermExtract[2]というツールを利用した[中川 2001]．中川らの抽出手法は，「多くの異なり語と連接する名詞から構成される複合語は重要語である」という考え方に基づいている．このツールをテキスト集合(ある専門分野のコーパス)に適用すると，重要度と共に専門用語リストが出力される．本研究では，重要度が低いものをリストから除去し，さらに残ったものの中から専門用語として適切であると考えられるものを人手で選定した．その結果，日本語の用語 512 語，英語の用語 1210 語を得た．これらに，言語情報処理ポータルで公開されている用語集[3]，言語処理学会Webページで公開されている日英対訳用語集[4]，NTTコミュニケーション科学基礎研究所の用語集[5]を加え，日本語 2906 語，英語 2839 語を収録した用語リストを得た．

4.2 実験方法

今回の実験では，自然言語処理分野の用語でも，特に「自動要約」や「機械翻訳」などの応用分野のものを対象とする．このような用語の集合(以後，入力用語リスト)を，以下に述べる手順で自動的に収集する．4.1 節で述べた専門用語リスト中の任意の用語を取り出し，その用語の後ろに「システム」をつけた用語が論文データベース中に存在すれば，入力用語リストに入れる．例えば，「形態素解析」という用語の場合，この用語の後ろに「システム」をつけた「形態素解析システム」という用語がデータベース中に存在するため，入力用語リストに入れる．しかし，「形態素」という用語に「システム」をつけた「形態素システム」という用語はデータベース中に存在しないため，入力用語リストの中には入れない．英語の用語についても同様に，ある用語の後ろに“system”という用語をつけた用語がデータベース中に存在すれば入力用語リストに入れる．以上述べた手順で得られた入力リストを用いて動向情報を抽出したところ，結果が得られたものが，日本語用で 38 語，英語用で 53 語あった．

評価

以下の 4 つの手法で収集できた用語数を調べる．

- 論文表題から METHOD タグを用いて抽出
- 論文表題から GOAL タグを用いて抽出
- type C で引用関係にある論文の表題から METHOD タグを用いて抽出
- type C で引用関係にある論文の表題から GOAL タグを用いて抽出

4.3 結果

結果を表 2 に示す．表 2 からわかるとおり，(入力) 1 用語あたり平均 6.86 語抽出できており，そのうち，引用を利用して抽出できた用語数は全体の約 1 割を占めている．

表 2 抽出された要素技術用語数(平均値)

手法		抽出用語数
表題から抽出	METHOD	5.93
	GOAL	0.32
引用情報利用	METHOD	0.17
	GOAL	0.43
計		6.86

4.4 考察

実際の論文結果を見たところ，論文表題からの専門用語の抽出ミスが多いことがわかった．例えば，「最大エントロピーモデルに基づく形態素解析」という表題からは，「形態素解析」の要素技術として，「最大エントロピーモデル」が抽出されるべきである．しかし，この用語が 4.1 節で述べた専門用語リストに含まれておらず，システムは代わりに部分文字列である「エントロピー」を抽出してしまった．この問題を解決する方法について，今後検討する．

[1] http://acl.ldc.upenn.edu/
[2] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
[3] http://www.kc.t.u-tokyo.ac.jp/NLP_Portal/glossary/index.html
[4] http://www.pluto.ai.kyutech.ac.jp/NLP/term.csv
[5] http://www.kecl.ntt.co.jp/mtg/resources/lingdic.txt

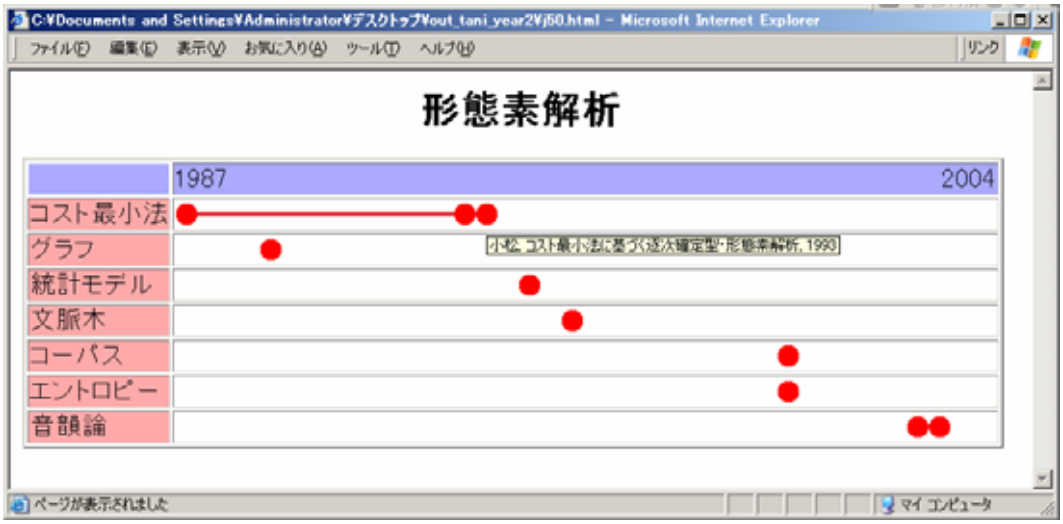


図 1 システムの動作例(「形態素解析」の要素技術)

5. システム動作例

図 1 にシステムの動作例を示す．図は，「形態素解析」という用語をシステムに入力した時の解析結果を示している．図において，左端に「形態素解析」の要素技術名が列挙してあり，その用語が論文表題中で使われた年が，各技術の右側に示してある．例えば図の「コスト最小法」の場合，この用語を論文表題に含んだ形態素解析に関する論文が 1987 年に 1 件，1993 年に 2 件発表されている．これらは図中で「●」として表示されており，その間が直線で結ばれている．ユーザが●上にカーソルを重ねると，その論文の書誌情報がポップアップ表示される．図では，「コスト最小法」 (一番右端の●)にカーソルを重ねた時のポップアップ表示として「小松，コスト最小法に基づく逐次確定型・形態素解析，1993」が例示されている．

6. おわりに

本研究では，「機械翻訳」や「自動要約」などの用語を与えると，論文データベースからその要素技術を示す用語を抽出し，それらを年代順にまとめてグラフとして出力する手法を提案した．日本語の用語 38 語，英語用語 53 語を入力として，用語毎に要素技術を示す用語を抽出したところ，1 用語あたり平均 6.86 語収集できた．

参考文献

[阿辺川 2003] 阿辺川 武, 難波 英嗣, 高村 大也, 奥村 学 (2003) “機械学習による科学技術論文からの書誌情報の自動抽出” 情報処理学会研究報告自然言語処理，NL-157，pp. 83-90.

[小原 2004] 小原 恭介，山田 剛一，絹川 博之，中川 裕志 (2004) “ウェブを利用した関連用語収集” 第 3 回情報科学技術フォーラム (FIT2004).

[佐藤 1999] 佐藤 理史 (1999) “論文表題を言い換える” 情報処理学会論文誌，Vol. 40，No. 7，pp. 2937-2945.

[佐藤 2003] 佐藤 理史，佐々木 靖弘 (2003) “ウェブを利用した関連用語の自動収集” 情報処理学会研究報告 自然言語処理，NL-153，pp. 57-64.

[白井 2004] 白井 清昭，菅井 俊介，平野 健児，星 正人 (2004) “ポータルサイト自動作成の試み” 言語処理学会第 10 回年次大会，pp. 624-627.

[千田 2005] 千田 恭子，篠原 靖志，奥村 学 (2005) “タイトルの文型が読者の関心に及ぼす影響の分析” 自然言語処理，Vol. 12，No. 2，pp. 87-107.

[長尾 1982] 長尾 真，辻井 潤一，矢田 光治, 柿元 俊博 (1982) “科学技術論文表題の英和機械翻訳システム” 情報処理学会論文誌, Vol. 23，No. 2，pp. 202-210.

[中川 2001] 中川 裕志，湯本 紘彰，森 辰則 (2001) “出現頻度と連鎖頻度に基づく専門用語抽出” 自然言語処理，Vol. 10，No. 1，pp. 27-45.

[難波 1999] 難波 英嗣，奥村 学 (1999) “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.

[難波 2005] 難波 英嗣 (2005) “論文間の引用情報を利用した関連用語の自動収集” 言語処理学会 第 11 回年次大会.