論文データベースからの研究動向情報の抽出

近藤友樹[1]　難波英嗣[1]　奥村学[2]　新森昭宏[3]　谷川英和[4]　鈴木泰山[5]
1 広島市立大学　2 東京工業大学
3 インテック・ウェブ・アンド・ゲノム・インフォマティクス　4 IRD 国際特許事務所　5 ピコラボ

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．しかし，このような動向調査には多くの時間と労力を要する．そこで，本研究では，学術論文データベースから動向情報を自動的に抽出し，可視化するシステムの構築を行う．

学術論文から動向情報の抽出を行うこれまでの試みとして，難波ら[難波 2006]の研究がある．難波らは，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 つのステップで動向情報の抽出を行っている．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．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，要素技術用語を抽出している．この結果を用い，ある分野でどのような要素技術がいつ頃から使われ始めたのか，また，ある要素技術がどのような分野で使われているのかを図として提示するシステムを構築している．

しかし，ステップ 2 において難波らが人手で作成したルールでは，十分な精度で表題解析を行うことができない，という問題がある．そこで本研究では，まず，表題解析に機械学習を取り入れ，より高い解析精度を目指す．次に，解析結果の提示方法の改良も行う．難波らのシステムには，上でも述べたとおり，ある要素技術を使う研究分野を一覧表示する機能があるが，この表示方法では，その要素技術を使う分野が数多く存在する場合，それらをすべて列挙するとユーザにとって非常に分かりにくい，という問題がある．そこで，本研究では，類似した分野をまとめて表示することで，分かりやすい提示を目指す．

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，難波らの研究の概要，問題点およびその改善方法を提案する．4 節では，改善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し，6 節で本稿をまとめる．

2. 関連研究

今井[今井 1999]は表題解析をし，その構造に基づいて論文の分類を行う手法を提案している．今井の手法は，「標準化」と「コード割当」の 2 つの処理から構成される．標準化では，動詞や機能語を手がかりに論文表題をいくつかの部分要素に分割する．コード割当では，それぞれの部分要素中の専門用語を抽出し，その用語を岩波情報科学辞典のコードと対応付けることで，論文の分類を実現している．論文表題を構造解析し主題を抽出するという点では，今井の研究と共通するが，本研究では表題解析に機械学習を取り入れている点と，主題だけでなく要素技術にも着目して処理を行う点が異なる．

研究動向の調査に関して，村田らは言語処理学会年次大会および論文誌の論文表題から名詞を抽出し，様々な側面から自然言語処理分野の研究動向の分析を行っている[村田 2005]．村田らの研究では，論文表題中の名詞はすべて等価にあつかっているが，本研究では，論文の表題を解析することで，主題を示す用語と要素技術を示す用語を区別して扱う点が異なる．

3. 研究動向情報の抽出

本節では，技術動向情報の抽出に関する先行研究[難波 2006]を 3.1 節で説明し，3.2 節でその問題点を指摘する．これらの改善方法について，3.3 節と 3.4 節で，それぞれ述べる．

3.1 論文表題解析を用いた動向情報の抽出

難波ら[難波 2006]は，手がかり語と人手で作成したルールを用いて論文表題を解析している．以下，その解析手順を説明する．まず，あらかじめ手がかり語と構造タグの対応リストを用意しておく．以下に，構造タグと手がかり語の一例を示す．

- ○ HEAD→論文の主題を示す．
- ○ METHOD→論文中で用いる要素技術を示す．(「を用いた」，「に基づく」)
- ○ GOAL→論文の目的・目標を示す．(「のための」)

次に，論文表題と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，「HMM を用いた形態素解析」という表題の場合，「を用いた」が METHOD タグの手がかり語であるため，「HMM<METHOD>形態素解析」が得られる．最後に，付与された各タグの直前の文字列を，そのタグの要素として抽出する．例えば，上の例の場合，「HMM」が METHOD タグの要素となる．ここで，この例ではこの時点で「形態素解析」という文字列には何もタグが付与されていないが，このような個所(多くの場合，論文表題中の一番末尾の名詞句)に HEAD タグを付与する．以上の処理から，最終的に HEAD が付与された個所（「形態素解析」）の要素技術として METHOD が付与された個所（「HMM」）を抽出する．

なお，難波らは、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

を使用している．また，構造タグは英語用に 11 種類，日本語用に 10 種類設定している．

3.2 難波らの研究の問題点

以下に，難波らの手法の問題点を指摘する．

[問題点 1]

難波らの表題解析手法では，例えば，以下の表題は正しく解析できない．

> 中国語形態素解析に対する SVM とコスト最小法の比較実験

この表題からは，「比較実験」が論文の主題(HEAD)として抽出される．しかし，この場合，「比較実験」よりも「SVM」や「コスト最小法」の方が論文の主題として適切であると考えられる．3.3 節では，この改善方法について述べる．以後，この例における「比較実験」を「形式的な主題」，「SVM」や「コスト最小法」を「真の主題」と呼ぶことにする．

[問題点 2]

難波らのシステムでは，要素技術を中心とした分野を抽出し，表示することができるが，その要素技術を使う分野が数多く存在する場合，それらをすべて列挙するとユーザにとって非常に分かりにくい，という問題がある．このような場合，類似した分野をひとつのグループにまとめて表示することで，わかりやすい表示が可能になると考えられる．表示方法の一例を図 1 に示す．図 1 は，要素技術「Hidden Markov Model(HMM)」が使われる分野の一覧を示した例である．図において，HMM を用いている 6 分野が，音声を対象にした分野なのか，テキストを対象にした分野なのかによって分類されている．この具体的な改善方法については 3.4 節で説明する．

HMM

[音声]

- continuous speech recognition (1988)
- speech recognition (1988)
- 音声合成 (2002)

[テキスト]

- 日本語形態素解析 (1995)
- 形態素解析 (1995)
- Summarization (2001)

図 1 要素技術(HMM)を中心とした分野の表示例

3.3 論文表題からの要素技術の抽出

本研究では，YamCha[1]と CRF++[2]を使用して論文表題解析を試みる．METHOD，GOAL，OTHER タグについては機械学習を用いて論文表題のタグ付けを行うが，HEAD タグは，機械学習＋人手ルールに基づく主題検出という 2 段階手法でタグ付与を行う場合（手法 1)と機械学習だけでタグ付与を行う場合（手法 2)の 2 通りの方法について比較する．HEAD タグが付与される個所は，論文表題の主題となる部分であるが，難波らの手法[難波 2006]のように，論文表題の末尾の名詞句に HEAD タグを付与すると，前節で述べたような形式的な主題に HEAD タグが付与される場合もある．もし，形式的な主題の検出ができれば，HEAD タグ付与の精度が向上すると考えられる．この 2 種類の手法の詳細については後述する．

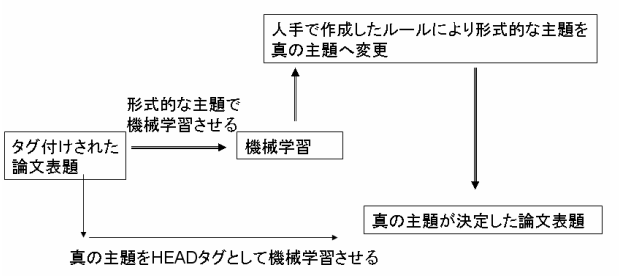


図 2 表題解析の 2 種類の解析の流れ

図 2 に，表題解析の流れを示す．まず，論文表題に形式的な主題が付与されるよう機械学習させ，HEAD が付与された文字列を収集し，リストを作る．これを不要語候補リストと呼ぶ．この中から出現回数順の上位 1000 件について技術的な用語を人手で省いた不要語リストを作る．さらに不要語リスト中の用語を部分文字列に含む名詞が不要語候補リストに存在するか再度確認し，もし存在すればリストに加える[3]．以下は不要語リストの例である．現在，不要語リストとして 6482 語が登録されている．

[不要語リスト(一部)]
研究　検討　開発　影響　解析　一考察　評価
考察　実験的研究　一検討　応用　効果　提案

このようにして作成された不要語リストを用い，表題解析を行う．本研究では，2 種類の方法で表題解析を試みる．手法 1 は，上で述べた形式的な主題にも HEAD タグを付与する解析器の出力結果を修正して正しい解析結果を得る方法である．この方法では，HEAD タグが付与された文字列が不要語リストに含まれていれば，HEAD タグを OTHER タグに書き換え，その個所よりも前に存在する名詞句を HEAD として抽出する．もし新たに HEAD として抽出された個所が不要語リストに含まれていれば，さらに前の名詞句を HEAD として抽出する．

手法 2 は，真の主題に人手で HEAD タグを付与したデータを用意し，機械学習を行う方法である．この時，論文表題中の語句が不要語リストに含まれているかどうかも素性のひとつとして用いることで，形式的な主題に誤って HEAD タグが付与されることがなくなると考えられる．

3.4 要素技術を中心とした分野の分類

ある要素技術を用いた複数の分野を分類するひとつ

[1] http://chasen.org/~taku/software/yamcha/
[2] http://www.chasen.org/~taku/software/CRF++/

[3] 例えば不要語リスト中の用語「特性」と「実験」から作られる「特性実験」や，「研究 2」などの数値が連結した語などを不要語候補リストから探す．

の方法として，その分野では「何を処理するのか」,「システムの入出力は何か」という観点で分野(用語)を分類する．以下にその方法を説明する．分野を示す用語の中には，用語の直後に「する.」を加えることで動詞になるものがある．例えば,「形態素解析」や「機械翻訳」といった用語に「する.」を加えると「形態素解析する.」や「機械翻訳する.」という表現が得られる．このような表現は，実際に使われることがある．こうした用語の多くは，何らかの入力があり，それを処理して新たなものを出力する用語であると考えられる[4].

ここで，このような文は「A を B に C する．」という文構造になっている場合が少なくない．この時，ヲ格(A)とニ格(B)を抽出すれば，それが C の入力と出力になっていると考えられる．例えば，C が「機械翻訳」の場合， ヲ格から「日本語文」や「文書」や「文字列」などが，ニ格から「英語」などが抽出できる．同様に，「形態素解析する.」の場合，ヲ格から「日本語文書」などが抽出できる．そこで，用語ごとに入出力情報を抽出し，それらを比較することで，入出力が似た用語同士をグルーピングすることが可能になる．

なお，ヲ格以外にもカラ格からも入力情報が，ニ格以外にもマデ格やヘ格から出力情報が得られる．さらに，専門用語に「する.」を付け加える以外にも「をする.」を付け加えた場合(例えば,「機械翻訳をする」)にも，同様に格情報から入出力情報が得られる．

この他，用語の分類方法として，上位，下位関係にも着目する．例えば,「音声認識」という用語の前に「大語彙連続」が付け加わった「大語彙連続音声認識」という用語は「音声認識」の下位語であるが，入出力に関する性質は,「音声認識」と基本的にほぼ同じであると思われる．そこで，分類対象となる用語の中に，上位，下位関係のある用語が含まれる場合には，これらを同一グループにまとめる．

4. 実験

3 節で述べた提案手法の有効性を調べるため，実験を行った．

4.1 実験に用いるデータ

表題解析には，NTCIR ワークショップ 1, 2 言語横断検索タスクのデータを用いる．このデータは，1988～1997 年の抄録データベースであり，国内 65 学会の発表論文約 45 万件を含んでいる, このうち半数以上は日英対訳になっている．これらのデータから無作為に抽出した日本語論文表題 1000 件に METHOD, GOAL, HEAD，OTHER タグを人手で付与したデータを機械学習に用いる．なお，HEAD タグは，形式的な主題の場合には type= “no”という属性が付与されている．この時，真の主題を示す個所には，その個所に付与されているタグに type= “head”という属性が付与されている．

[4] 例えば「形態素」という専門用語に「する．」を付け加えた「形態素する.」という表現は存在しない．「形態素」は何か入出力のある用語ではない．

4.2 実験方法

機械学習の入力データを表 1 に示す．表の 1 列目は形態素を示す．ただし，形態素の品詞が名詞か未知語で連続する場合にはひとつにまとめる．表の 2 列目は形態素の品詞を示す．なお，形態素解析器には MeCab を用いる[5]．3 列目は，難波ら[難波 2006]が表題解析に用いた手がかり語のうち，METHOD タグの付与に関するものと一致すれば 1，なければ 0 を素性として用いる．同様に，4 列目は，難波らが GOAL タグの付与に用いた手がかり語の有無を素性として用いる．5 列目は人手で付与した正解タグデータである．なお，表 1 に示す例は，不要語リストを作成する時に，形式的な主題に HEAD タグを付与する場合の素性であり, 3.3 節で述べた手法 2 を実施する場合には，さらに，各形態素が不要語リストに含まれているかどうかも素性として用いる．

表 1 機械学習への入力データ例

電気回路演習用CAI	未知語	0	0	B-OTHER
と	助詞	0	0	I-OTHER
その	フィラー	0	0	I-OTHER
改良	名詞	0	0	B-HEAD

4.3 結果

形式的な主題を HEAD として機械学習させた結果を表 2 に示す．なお，この結果は，真の主題を HEAD とした時の精度を示している．表からわかるとおり，形式的な主題を抽出すると真の主題の抽出精度は 44% 程度と，非常に低い値になっている．次に，3.3 節で述べた手法 1，手法 2 それぞれで解析した結果を表 3，表 4 にそれぞれ示す．

表 2　形式的な主題を HEAD とした時の解析結果(%)

構造タグ	Precision	Recall	F 値
GOAL	64.3	94.7	76.6
HEAD	44.6	40.5	42.4
METHOD	89.5	91.4	90.5
OTHER	82.0	83.9	82.9
全体	74.8	74.8	74.8

表 3　手法 1 の解析結果(%)

構造タグ	Precision	Recall	F 値
GOAL	72.0	94.7	81.8
HEAD	77.4	72.8	75.0
METHOD	90.9	88.4	89.6
OTHER	91.4	93.1	92.3
全体	88.4	88.4	88.4

表 4　手法 2 の解析結果(%)

構造タグ	Precision	Recall	F 値
GOAL	79.0	79.0	79.0
HEAD	77.5	75.3	76.4
METHOD	87.3	83.8	85.5
OTHER	92.2	93.3	92.7
全体	88.9	88.9	88.9

[5] http://mecab.sourceforge.net/

4.4 考察

表 2 の結果と手法 1，2 による結果(表 3 と 4)を比較すると，HEAD の精度と再現率が 30%以上向上していることがわかる．このことから，今回作成した不要語リストの有効性が確認できる．

手法 1 と手法 2 を比較すると，HEAD の解析精度は手法 2 の方が Precision 値は 3%程度高いものの，METHOD に関しては手法 1 の方が Precision 値で 3%，Recall 値で 5%程度良い，という結果になった．動向情報の抽出には HEAD と METHOD の情報が共に必要であるため，手法 1 と 2 のどちらが良いか，この結果からは簡単に結論付けることはできない．ただ，手法 1 に関して，真の主題を検出するルールは不要語リストと照合するという非常に単純なものであるが，例えば抽出している HEAD が，どの程度 HEAD らしいか(例えば，他の論文表題から同一の文字列が HEAD として抽出されている件数)を考慮することで，主題の検出精度を向上できるのではないかと思われる．

5. システム動作例

図 3 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 3 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 3 の「接続コスト最小法」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 1991 年に 1 件，1993 年に 1 件発表されており，これらは図 3 中で「●」として表示され，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 3 では，「確率モデル」(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「確率モデルによる自由発話の形態素解析, 1994, 言語・音声理解と対話処理研究会(略称 SIG-SLUD), (人工知能学会)」が例示されている．

図 4 は，図 3 の「HMM」をクリックした時に表示される画面である．この図は，HMM が使われた分野一覧を示しており，類似する分野はひとつのクラスにまとめられている．例えば，クラス 1 において，「音声認識」の下位語である「孤立単語音声認識」は同じグループとしてまとめられている．また，「音声合成」と「音声認識」は，いずれも「音声信号」と「文字列」が共通の入力となっているため，ひとつのクラスにグルーピングされている．なお，入出力情報を得るために，今回はコーパスとして特許公開公報 1993～2002 年を用いた．

6. おわりに

本研究では，難波ら[難波 2006]の手法をベースに，動向情報を抽出する手法を提案し，システムを実装した．提案手法は，論文の形式的な主題を検出するため，半自動的に不要語リストを作成し，これを用いて表題解析を行う．その結果，(形式的でない)主題の検出精度が約 30%以上向上した．

また，サ変名詞型の専門用語に着目し，格解析を用いて，その用語の入出力情報を自動収集し，入出力の類似性という観点から用語を分類する手法を提案した．

図 3「形態素解析」に関する要素技術の一覧表示

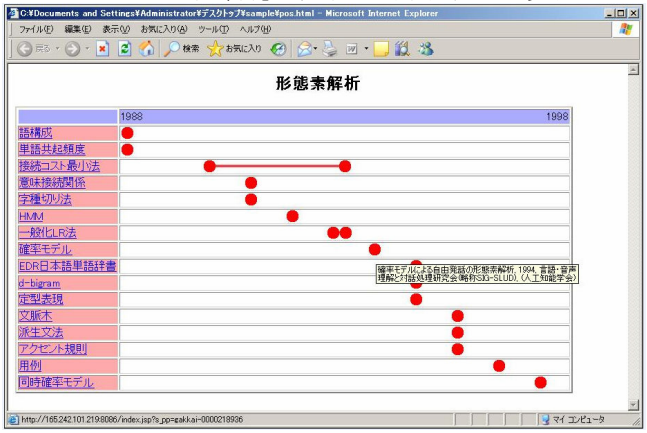
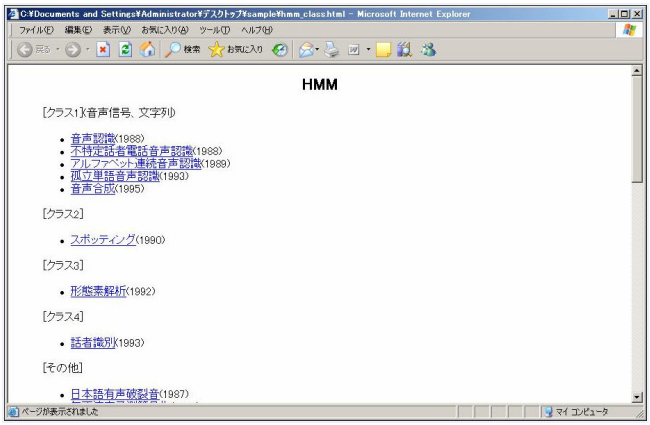


図 4 HMM が使われる分野のクラス別表示



7. 謝辞

本研究で用いた論文データおよび特許データは，国立情報学研究所の許可を得て，NTCIR テストコレクションを利用させていただいた．本研究は，NEDO 産業技術研究助成事業の支援を受けて行われた．

参考文献

[難波 2006] 難波 英嗣 谷口 裕子：「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」．言語処理学会 第 12 回年次大会 併設ワークショップ「言語処理と情報可視化の接点」，pp.35-38，2006

[村田 2005] 村田 真樹 一井 康二 馬 青 白土 保 井佐原 均：「過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査」．言語処理学会 第 11 回年次大会，2005

[今井 1999] 今井 俊：「表題解析による科学技術論文の自動分類」．北陸先端科学技術大学院大学修士論文，1999