様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 6 月 18 日現在

研究種目：若手研究(B)
研究期間：2007 ～ 2009
課題番号：19700152
研究課題名（和文）連接語に着目した専門用語の体系化および技術動向分析への応用
研究課題名（英文）Organization of technical terms by adjacent words and its application to technical trend analysis

研究代表者
難波　英嗣　（NANBA HIDETSUGU）
広島市立大学・情報科学研究科・講師
研究者番号：50345378

研究成果の概要（和文）：
本研究では、専門用語を、その用語が持つ性質や種類によって体系化する手法を提案し、実験によりその有効性を確認した。また、1993 年～2002 年の公開特許公報から抽出した約 2000 万文から、専門用語シソーラスを構築した。さらに、この手法を用い、ある分野の技術文書集合から技術動向の直感的な理解を可能にする技術動向マップを自動的に生成するシステムの改良を行った。

研究成果の概要（英文）：
We proposed a method that organizes technical terms by their properties, and confirmed its effectiveness by experiments. Using this method, we constructed a thesaurus from 20,000,000 sentences in unexamined Japanese patent applications published during 1993-2002. We also improved a system that creates technical trend map from a set of technical documents using this method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,100,000	0	1,100,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,200,000	630,000	3,830,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：技術動向支援、専門用語

１．研究開始当初の背景

近年、学術情報量が爆発的に増加し、専門家は自分の専門分野の最新動向を把握するために、絶えず膨大な量の文献を読まなければならない状況に直面している。また、研究分野の専門分化に伴って、ある分野の知識を得るために、さらに複数の別の分野についても知らなければならないということも、もはや一般的になりつつある。バイオテクノロジー、半導体、情報科学のように研究・開発・

製品化のサイクルが非常に短い分野では、論文だけでなく、特許等、他のジャンルの文献にも注意を払う必要がある。しかし、入手した文献全てに目を通し利用することは益々困難になっている。

そこで、研究代表者は、ある分野の技術動向を自動抽出する研究を行っている。一般に、多くの論文表題には「Aに基づいたB」や「Cを用いたD」などの表現が含まれる。ここでAやCは、ある研究課題BやDを実現するための要素技術を示す用語であると考えられるので、論文表題を解析することにより、要素技術を示す用語を抽出することは可能である。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軸にとれば、ある分野の動向を示すグラフを作成することもできる。

研究代表者はこのような手法に基づいて、ある分野の技術動向分析ができるシステムのプロトタイプシステムをすでに作成しているが、現時点では、要素技術として抽出される用語に、様々な性質や種類のものが混在しているという問題点がある。以下にその一例を示す。

- アルゴリズムやモデルを示す用語(例，隠れマルコフモデル)
- データやコーパスを示す用語(例，World Wide Web，タグ付きコーパス)
- 装置を示す用語(例，電子顕微鏡)

分野によっては，100 以上もの要素技術用語が提示される場合もあり、用語を性質毎に分類できる機能が必要となっている。

２．研究の目的

本研究では、以下の２点に取り組む。

(i) 専門用語の体系化

専門用語を，その用語が持つ性質や種類によって分類する手法を提案し、実験によりその有効性を確認する。

(ii) 技術動向分析システムの改良

1 の結果を利用し、上で述べた技術動向分析システムの問題点を改善する。具体的には、図 3(4 節)に示すような技術動向マップ提示システムを構築する。

３．研究の方法

(i)専門用語の体系化

例えば、「形態素解析」「機械翻訳」「データベース」「形態素」という 4 つの専門用語のうち「形態素解析する」や「機械翻訳する」は存在するが、「データベースする」や「形態素する」という表現は存在しない。隣接用語「する」をとりうる専門用語の多くは、何らかの入力があり、それを処理して出力するものであると考えられる。このような専門用語を対象に、格フレーム辞書の構築を行う。これまでに、「見る」や「食べる」などの一般的な動詞を対象に、大規模な文書集合から格フレーム辞書を構築するという研究は行われているが、本研究では、それをサ変動詞「する」に隣接可能な専門用語に適用する点が従来研究と異なる。

一般に、専門用語+「する。」という表現を含む文は「A を B に C する。」という構造になっている場合が多く、このような文のヲ格を抽出すればそれがある専門用語の入力に、またニ格を抽出すればそれが出力になっていると考えられる。例えば、「機械翻訳」の場合、A として「英語」や「英語文書」などが、A として「日本語」や「日本語文書」等が収集できると思われる。同様に形態素解析の場合、B として「文書」「テキスト」「文」等が収集できる。

(ii)技術動向分析システムの改良

研究代表者は、これまでに「サポートベクトルマシンを用いたテキスト自動要約」といった論文表題から、「を用いた」のような手掛かり句に着目することで、この論文の主題として「テキスト自動要約」、要素技術として「サポートベクトルマシン」を抽出する手

法を提案している。この手法を英語論文にも拡張し、日英論文を対象にした言語横断技術動向分析システムを構築する。英語論文表題の構造は、日本語論文の表題と比べ多様であり、日本語論文の表題解析手法と同様の方法では十分な解析精度が得られない、という問題があった。そこで、英語論文表題の構造を解析する際、機械翻訳技術と日本語論文の表題構造解析技術も併せて用いることにより、精度の向上を試みる。

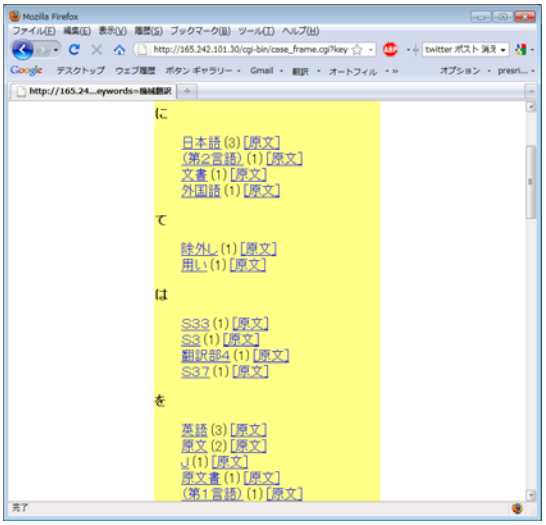
また、特許と論文を対象にした技術動向分析支援システムの構築を行う。上述のとおり、一般に、特許や論文などの表題や概要中には、「Aを用いた」や「Bに基づく」などの表現が含まれている。このAやBは、ある技術を実現するための要素技術を示す用語であることが多い。一方、表題中の末尾周辺の名詞句は、その論文のテーマ(主題)を示していることが多い。そこで、特許や論文の表題や概要を解析して要素技術と主題を抽出し、同一の主題を持つ特許と論文から抽出された要素技術を縦軸に、各文献の著作年を横軸にとって表示することにより、その主題を中心とした要素技術の変遷を知ることができる。さらに、「エラー率が低減」や「処理速度が向上」などの効果に関する表現もあわせて提示できれば、特定分野の技術動向を効率的に把握することができる。

4．研究成果

(i) 専門用語の体系化

本研究では、1993年～2002年の公開特許公報から「する。」を含む約 2000 万文を抽出・構文解析器し、専門用語の格フレーム辞書を構築した。また、Web上でこの格フレーム辞書を提示できるシステムを構築した。図1は、「機械翻訳」に関する格フレームを提示している。この図は、「機械翻訳する。」という文字列を含む文を特許から収集し、「機械翻訳する」という文節と係り受け関係にある文節をまとめたものである。各文節に併記してある数値は出現頻度を示している。このうち、ヲ格「英語」や「原文」などは機械翻訳の入力を、ニ格「日本語」や「(第２言語)」は機械翻訳の出力を、それぞれ示しており、構築された格フレームが正しく動作していることが確認できる。

図１「機械翻訳」に関する格フレーム



(ii) 技術動向分析システムの改良

英語表題の構造解析では、精度 78.0%、再現率 75.2%の解析精度が得られ、提案手法の有効性が確認された。

図２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２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２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２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 3 は、図 2 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、「専門用語の体系化」の欄で述べた手法を用いた。

図 2「形態素解析」に関する要素技術の一覧表示

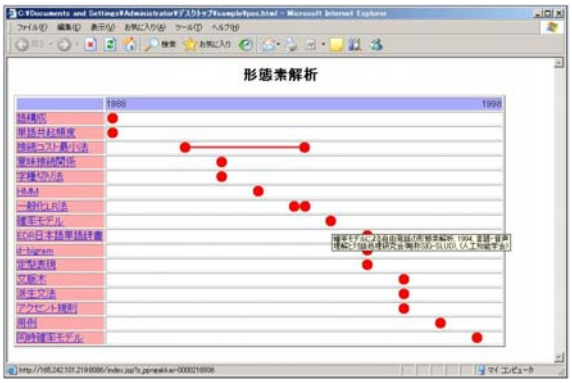
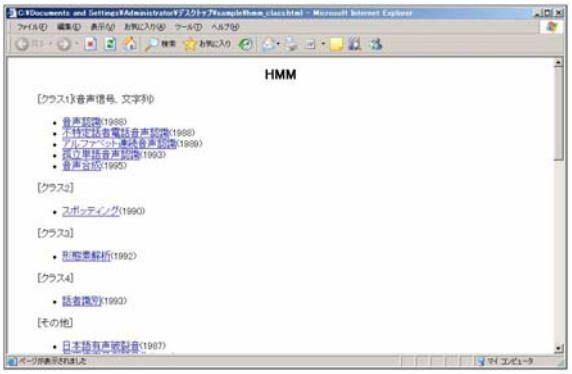


図 3 HMM が使われる分野のクラス別表示



５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）
① 難波英嗣、竹澤寿幸、2 種類の翻訳システムを用いた学術論文の特許分類体系への自動分類、情報処理学会論文誌データベース、査読有、Vol。2、No。3、2009、76-86
② 難波英嗣、奥村学、新森昭宏、谷川英和、特許と論文を対象にした技術動向分析、Japio Year Book、査読無、2007、184-191

〔学会発表〕（計 6 件）
① 近藤友樹、難波英嗣、竹澤寿幸、特許と論文からの技術動向マップの自動構築、言語処理学会第 16 回年次大会、査読無、2010、114-117
② Tomoki Kondo, Hidetsugu Nanba, and Toshiyuki Takezawa, Technical Trend Analysis by Analyzing Research Papers' Titles、Proceedings of the 4th Language & Technology Conference, 査読有, 2009, 234-238
③ Hidetsugu Nanba and Toshiyuki Takezawa, Classification of Research Papers into a Patent Classification System Using Two Translation Models, Proceedings of Workshop on Text and Citation Analysis for Scholarly Digital Libraries, the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, 査読有, 2009, 27-35
④ 近藤友樹、難波英嗣、竹澤寿幸、翻訳知識を用いた英語論文表題の構造解析、情報処理学会自然言語処理研究会 NL-187、査読無、2008、37-43
⑤ Hidetsugu Nanba, Hiroshima City University at NTCIR-7 Patent Mining Task, Proceedings of the 7th NTCIR Workshop Meeting, 査読無, 2008, 369-372
⑥ Hidetsugu Nanba, Query Expansion using an Automatically Constructed Thesaurus, Proceedings of the 6th NTCIR Workshop Meeting, 査読無, 2007, 243-248

〔図書〕（計 0 件）
なし

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
なし

６．研究組織
(1)研究代表者
難波　英嗣　(NANBA HIDETSUGU)
広島市立大学・情報科学研究科・講師
研究者番号：50345378

(2)研究分担者
なし

(3)連携研究者
なし