2008－NL－187 (6)
2008／9／24

翻訳知識を用いた英語論文表題の構造解析

近藤 友樹　難波 英嗣　竹澤 寿幸

広島市立大学大学院 情報科学研究科

〒731-3194 広島市安佐南区大塚東 3-4-1

E-mail: kondo@nlp.its.hiroshima-cu.ac.jp, {nanba,takezawa}@hiroshima-cu.ac.jp

あらまし 我々は，研究者が技術動向を分析する作業を支援するシステムの構築を目指している．本稿では，その要素技術となる英語論文表題の構造解析手法を提案する．一般に英語論文表題の構造は日本語のものより複雑であるため，英語論文表題の構造解析は日本語の場合よりも解析精度が低い，という問題があった．そこで，日本語論文表題の構造解析結果と翻訳知識を組み合わせた，英語論文表題の構造解析手法を提案する．実験の結果，精度 77.98%，再現率 75.15%が得られ，提案手法の有効性が確認された．

キーワード 技術動向分析，表題構造，機械学習，機械翻訳

Analysis of Research Papers' Titles Written in English using Translation Information

Tomoki KONDO　Hidetsugu NANBA　Toshiyuki TAKEZAWA

Graduate School of Information Sciences, Hiroshima City University

3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

E-mail: kondo@nlp.its.hiroshima-cu.ac.jp, {nanba,takezawa}@hiroshima-cu.ac.jp

Abstract We have been studying towards construction of a support system for technical trend analysis. In this paper, we propose a method for analyzing research papers' titles written in English, which is an elemental technology for the support system. Generally, structures of research papers written in English are more complicated than those in Japanese, and this degraded the performance of the analysis method for English titles in comparison with that for Japanese. To solve this problem, we propose a method for the English titles using the analysis results of Japanese titles and Translation Information. We conducted an experiment and found that our method obtained a precision of 77.98% and a recall of 75.15%

Keyword technical trend analysis, structures of research papers' titles, machine learning, machine translation

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，このような技術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の論文表題の構造を解析し，その結果を用いて技術動向分析を可能にするシステムを提案する．

学術論文データベースから動向情報の抽出を行うこれまでの試みとして，近藤ら[近藤 2007]の研究がある．多くの日本語論文の表題には，「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 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と考えられる．そこで，近藤らは，論文表題を解析して A や B に相当する個所(用語)を抽出し，その論文の著作年を 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自動的に作成するシステムを構築している．本研究では，日本語論文を対象にした近藤らの手法に基づき，対象を英語論文にも広げることで，よ

り網羅性の高い技術動向分析を可能にするシステムの実現を目指す.

ここで，英語論文表題の構造解析には，次の2つの問題点がある.

(1) 英語論文表題の構造は一般的に日本語のものよりも複雑である.
(2) 構造解析に用いる手がかり語が日本語ほど有効に機能しない.

このため，詳細は後述するが，近藤らの日本語論文表題の構造解析と同等の手法で英語論文表題を解析すると，日本語論文表題よりも解析精度が 5%程度低いことが実験により明らかになっている．この問題を解決するため，本研究では，解析精度の高い日本語表題解析技術と翻訳知識を組み合わせることにより，英語論文表題の解析精度の向上を試みる.

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，日本語表題解析技術と翻訳知識を組み合わせた英語論文表題の構造解析手法を提案する．提案手法の有効性を調べるために行った実験について 4 節で報告し，実験結果を 5 節で考察する．最後に 6 節で本稿をまとめる.

2. 関連研究

今井[今井 1999]は日本語論文表題の構造を解析し，その結果を用いて論文を自動分類する手法を提案している．この手法は，「標準化」と「コード割当」という 2 つの処理から構成される．まず，「標準化」処理では，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．次に，「コード割当」処理では，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，論文を分類する．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扱っているが，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と，日本語だけでなく英語論文も対象にすることで，網羅的な動向分析を目指している点が異なる.

3. 英語論文表題の構造解析

本節では，まず，近藤ら[近藤 2007]の日本語論文表題の構造解析手法を 3.1 節で説明する．次に，3.2 節で英語論文表題の構造解析の基本方針，3.3 節で英語論文表題を対象とした場合の問題点について述べ，その改善方法を 3.4 節で説明する.

3.1 日本語論文表題の構造解析

近藤らは，日本語論文表題の構造を表現するための 4 種類のタグ:HEAD, METHOD, GOAL, OTHER を定義している.

○ **HEAD**：論文の主題(研究分野)を示す.
○ **METHOD**:論文中で用いる要素技術を示す.
○ **GOAL**：論文の目的や最終目標を示す.
○ **OTHER**：その他.

例えば「SVM を用いた重要文抽出」という論文表題は，「重要文抽出」という主題において，「SVM」という要素技術が使われている，と解釈できるが，この場合，上記のタグを用いて次のように表現できる.

<METHOD>SVM</METHOD><OTHER> を 用いた</OTHER><HEAD>重要文抽出</HEAD>

近藤ら[近藤 2007]は，論文表題の構造解析を「論文表題を形態素解析し，各単語に上記の 4 種類のタグのいずれかを付与」する系列ラベリング問題と考え，機械学習を用いて上記のタグの自動付与を行っている．学習には，「表題中の各単語」，「品詞」，および以下に述べる「手がかり語の有無」と「不要語の有無」を素性として用い，F 値 82.0%の解析精度を得ている.

● 手がかり語

論文表題中の「を用いた」や「に基づく」の直前には要素技術を表す用語が，また，「のための」の直前には論文の目的や最終目標等を表す用語が出現すると考えられる．このような手がかり語のリストを作成しておき，手がかり語の有無を機械学習の素性に用いる.

● 不要語リスト

多くの日本語論文表題では，表題末尾の名詞句が主題となる場合が多い．例えば，上述の「SVM

を用いた重要文抽出」という表題では，末尾の名詞句「重要文抽出」が論文の主題となる．しかし,「汎用キャッシュメモリコントローラの開発」という論文表題では，末尾の名詞句が「開発」となるが，実際には「汎用キャッシュメモリコントローラ」が論文の主題となる.そこで，論文の末尾に出現しやすいが主題にはなり得ない名詞句の集合を半自動的に収集して不要語リストを作成し，論文表題中の各単語がこのリストに含まれるか否かを素性とする．

3.2 英語論文表題の構造解析の基本方針

英語論文表題の構造解析も，日本語論文表題と同様の手順により行うことを基本方針とする．すなわち,「表題中の各単語」,「品詞」,「手がかり語の有無」「不要語の有無」を機械学習の素性に用いる．

不要語リストの作成方法を説明する．以下の英語論文表題の例は，下線部が論文の主題であり，その直前の名詞句"An Analysis"が不要語となっている．

> An Analysis of Push-Pull Type Parametric Transformer using Orthogonal-Cores

不要語の使用状況に関して実際の英語論文表題データを用いて調査したところ，1000 論文中 585 論文において，いずれも論文表題の先頭部で"An Analysis"や"A Study"といった不要語が使われていることが分かった．そこで，以下に述べる手順で不要語リストの作成を行った．まず，NTCIR-1, 2 の言語横断検索タスク[Kando 1999][Kando 2001]で使われたテストコレクションの英語論文表題 33 万件を，TreeTagger[1]を用いて品詞タグ付けし，表題先頭部の名詞句を抽出した．次に，名詞句を頻度に並べ，上位 1000 件について不要語か否かを人手で判定した．その結果，924 個の不要語が得られた．

3.3 英語論文表題の構造解析の問題点

3.2 節で述べた素性を用い，3.1 節で説明した近藤ら[近藤 2007]と同様の手順で機械学習を行い，英語論文表題の構造を解析したところ，英語論文表題では，次の 2 つの問題があることが分かった．

(1) 英語論文表題の構造の複雑さ

次に示す 3 つの論文表題は，下線部が主題を示している．多くの日本語論文では，論文の主題は不要語を除く表題末尾の名詞句となるが，英語では，この例からも分かるとおり，表題の先頭部，中間，および表題全体が主題となる，など様々なケースがあるため，主題部の特定が容易ではない．

- [例 1] Electric Field Distribution of Helix LCX on the Ground
- [例 2] The Result of Propagation Test on Transmission Line Parallel Sorting on Multi-stage Network
- [例 3] GaInAsP/InP High-speed Optical Intensity Modulator

(2) 構造解析に用いる手がかり語の問題

日本語論文表題では，「を用いた」や「に基づく」といった手がかり語の直前の名詞句は要素技術に関する用語である．英語論文表題の場合，"using"や"based on"の直後の名詞句は，ほぼ要素技術に関する用語と考えて差し支えないが，"with"や"by"のような前置詞は，直後の名詞句が必ず要素技術となるとは限らない．

これらの問題による解析精度低下の影響を軽減するため，次節で「翻訳知識」を用いた構造解析手段を提案する．

3.4 翻訳知識を用いた構造解析

論文表題中の名詞句の中には，例えば「隠れマルコフモデル(HMM)」や「サポートベクターマシン(SVM)」のように潜在的に要素技術になりやすい用語，「機械翻訳」や「自動要約」のように主題になりやすい用語などが存在すると考えられる．英語論文表題の構造を解析する際，3.2 節で述べた素性の他に，このような知識を与えれば，英語論文表題の構造解析の精度向上につながると考えられる．

1 節でも述べたとおり，日本語論文表題の構造解析精度は,英語のものよりも高い.そこで，まず，大量の日本語論文表題を解析し，METHOD タグが付与された用語(要素技術)，HEAD タグが付与された用語(主題)，GOAL タグが付与された用語(目的)を頻度順に並べ，次に，その上位 3,000 語[2]を辞書や翻訳器等を用い

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[2] 予備実験で上位 3,000 語を用いた場合と

dipole contribution	NP	0	0	0	0	0	0	B-HEAD
to	TO	1	0	0	0	0	0	B-OTHER
current-time characteristics	NP	0	0	1	0	0	0	B-GOAL
of	IN	1	0	0	0	0	0	B-OTHER
polar polymers	NP	0	0	0	0	0	0	I-OTHER

図 1　機械学習に用いる素性

て翻訳し，最後に，それらの用語の英語論文表題中での有無を素性に用いて，英語論文表題の構造解析を行う．専門用語は，以下の 3 種類の方法で翻訳する．

- 専門用語翻訳器
  要素合成法に基づく専門用語の機械翻訳器[Taniguchi 2008]を用い，専門用語を英訳する．
- NTCIR テストコレクションから自動抽出した訳語対データ
  3.2 節で述べた NTCIR テストコレクションは，約 26 万対からなる論文書誌情報(表題，著者名，出典，概要，キーワード)の日英対訳データを保持している．この中から，日本語と英語のキーワード数が一致している約 71 万語の用語対を抽出した．このデータを用いて日本語専門用語を英語に翻訳する．
- 科学技術 45 万語対訳辞典[3]
  電気・電子・情報・医学・生物学・地球科学などの科学技術用語の対訳辞典を用いて日本語専門用語を翻訳する．

なお，専門用語を翻訳する際，以下の点に注意する．専門用語は論文表題中では特定の名詞と結合して新たな名詞句となる場合が少なくない．例えば，「情報検索」という専門用語に「実験」という名詞が結合し「情報検索実験」という名詞句となったり，"light-triggered thyristor(光サイリスタ)"に"characteristic(特性)"が結合し"light-triggered thyristor characteristic (光サイリスタ特性)"となったりする場合等である．上述の 3 種類の翻訳方法は，いずれも「情報検索」や"light-triggered thyristor"を翻訳するのに適しているが，「実験」や"characteristic"といった語と結びついて作られた名詞句は翻訳することができない場合が多い．そこで，論文表題中の名詞句が概要またはキーワードに存在するか否かを調べ，もし存在しなければ，名詞句を構成する末尾の語基(名詞)を削除することでこのような不要な名詞を削除する．

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 実験手法

実験に用いるデータ

表題解析には，3.2 節で述べた 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[4]，これらのデータから無作為に抽出した英語語論文表題 1,000 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．

機械学習

機械学習には CRF++[5]を用いる．機械学習で用いる素性を表 1 に示す．表 1 において，1 列目は単語を示す．ここで，形態素の品詞が形容詞か名詞で連続する場合にはあらかじめひとつにまとめておく．2 列目は各単語の品詞を示す．ここで，品詞タグ付けには TreeTagger を用いる．3 列目および 4 列目は，METHOD タグおよび GOAL タグに関する手がかり語の有無を示している．5 列目は不要語リストで有無を示す，6-8 列目は 3.4 節で述べた提案手法の，HEAD，METHOD，GOAL 専門用語リストに論文表題中の単語が含まれるか否かを示す．9 列目は教師データである．

評価尺度

評価には，以下に示す再現率と精度を用いる．

6,000 語を用いて実験した結果，3,000 語を用いた場合の方が高い解析精度が得られたため，この実験では 3,000 語を用いる．

[3]「科学技術 45 万語対訳辞典」日外アソシエーツ株式会社，2001.

[4] このうち，約 26 万件が日英対訳データとなっている．

$$\text{精度} = \frac{\text{提案手法により正しく付与されたタグの数}}{\text{提案手法により付与されたタグの総数}}$$

$$\text{再現率} = \frac{\text{提案手法により正しく付与されたタグの数}}{\text{人手で付与したタグの総数}}$$

比較手法

以下の4種類の手法で英語論文を表題解析して，タグ付与した結果を比較する．

- BASELINE (ベースライン手法):

日本語の表題解析実験と同様の素性を使用する．表記，品詞，METHODの手がかり語の有無，GOALの手がかり語の有無，不要語リストの有無である．以下の比較手法は，上記の5つの素性を含む．

- TRANS (提案手法):

英語の専門用語を，「翻訳知識」の機械翻訳で得られた専門用語リストと照合した素性を加える．

- NTCIR (提案手法):

英語の専門用語を，「翻訳知識」のNTCIRテストコレクションから自動抽出した訳語対データで得られた専門用語リストと照合した素性を加える．

- DIC (提案手法):

英語の専門用語を，「翻訳知識」の科学技術45万語対訳辞典で得られた専門用語リストと照合した素性を加える．

4.2 実験結果

4.1節で述べた各手法の解析精度を表1に示す．提案手法の解析精度は，いずれも若干ではあるがBASELINE手法を上回る結果となった．このことから，日本語論文表題の構造解析結果と翻訳知識が英語論文表題の構造解析に有効であったと考えることができる．

提案手法の中で最も精度・再現率の値が得られたのはTRANSであり，ベースライン手法と比べ，精度が0.47%，再現率が0.53%向上した．1節でも述べたとおり，本研究の最終目標は，「どのような要素技術がいつ頃から使われているのか」をユーザに提示することであり，その点では，今回付与した4種類のタグの中でMETHODタグに関する解析精度が特に重要となる．今回最も精度の良かったTRANSは，METHODに限定すると，BASELINEと比べ，精度が1%以上上昇していることが分かる．

5. 考察

5.1節で3種類の翻訳知識の違いについて考察する．5.2節では，解析誤りについて述べる．

5.1 翻訳知識の相違点

3種類の翻訳知識TRANS, NTCIR, DICを，いくつかの側面から比較した．表2は，各提案手法によるGOAL，HEAD，METHODリスト内の専門用語数である．用語のカバレッジの面からはNTCIR手法のリストの用語数が最も高いが，解析精度の面では，TRANSの方が高い．このことから，訳語の品質はTRANSの方が高いと考えられる．表3は，各提案手法で共通するGOAL，HEAD，METHODリスト内の専門用語数を示したものである．この表から，リスト内の用語は手法間でかなり異なっていると思われる．

5.2 解析誤りの分析

TRANSによる解析誤りは，大きく次の6種類に分類できる．

(1) 統計的理由による失敗(40.6%)
(2) "of"の問題(15.0%)
(3) 並列句処理の問題(10.5%)
(4) 動名詞の問題(6.6%)
(5) 不要語の不足(1.3%)
(6) その他(26.0%)

以下に，それぞれの解析誤りについて説明する．

(1) 統計的理由による失敗(40.6%)

以下の例では，"Multiple-valued Pla"の個所にHEADタグが付与されるべきであるが，その直前にGOALタグの手がかり語である前置詞「for」があるため，誤ってGOALタグが付与され，代わりに"A Minimization Technique"の個所にHEADタグが付与されている．

[正解] A Minimization Technique for <HEAD>Multiple-valued Pla</HEAD>
[解析結果] <HEAD>A Minimization Technique </HEAD> for <GOAL> Multiple-valued Pla </GOAL>

[5] http://www.chasen.org/~taku/software/CRF++

表 1. ベースライン手法および提案手法による解析精度(%)

	BASELINE		TRANS		NTCIR		DIC	
	再現率	精度	再現率	精度	再現率	精度	再現率	精度
GOAL	82.03	81.06	82.63	80.70	82.03	80.58	82.03	80.11
HEAD	72.55	68.76	72.80	68.99	73.00	69.05	72.80	68.99
METHOD	73.57	90.44	74.09	91.66	74.09	91.66	74.61	91.13
OTHER	70.35	69.78	71.10	70.58	70.66	70.34	70.88	70.51
平均	74.62	77.51	75.15	77.98	74.94	77.90	75.08	77.68

表 2. 各手法によるリスト内の専門用語数

	TRANS	NTCIR	DIC
GOAL	849	1430	241
HEAD	1264	2397	730
METHOD	1263	2588	880

表 3. 手法間で共通するリスト内の専門用語数

	TRANS × NTCIR	NTCIR × DIC	TRANS × DIC
GOAL	173	138	42
HEAD	342	299	101
METHOD	491	443	241

(2) "of"の問題(15.0%)

以下の例では，"Pie64"の個所に HEAD タグが付与されるべきであるが，"of"の直前にある"The Hardware Implementation"の個所に誤って HEAD タグが付与されている．訓練データの中には，"of"の直前の用語に HEAD タグが付与される場合と，"of"の直後の用語に付与される場合が存在しており，少なくとも今回機械学習に用いた素性からだけでは，正確な判断が出来ないと考えられる．

[正解] The Hardware Implementation of <HEAD>Pie64</HEAD>

[解析結果] <HEAD>The Hardware Implementation </HEAD>of Pie64

(3) 並列句処理の問題(10.5%)

以下の例では，"Superconducting Magnetic Shield"の個所に HEAD タグが付与されるべきであるが，並列句"and"の解析に失敗し，"Defects and Unisotropy"の個所に HEAD が誤って付与されている．

[正解] Study of Defects and Unisotropy of <HEAD> Superconducting Magnetic Shield</HEAD>

[解析結果] Study of <HEAD> Defects and Unisotropy </HEAD>of Superconducting Magnetic Shield

(4) 動名詞の問題(6.6%)

以下の例では，"Space Charge"の個所のみに HEAD タグが付与されるべきであるが，動名詞"induced"の後の単語"Xlpe"にまで HEAD タグが付与されている．

[正解] Measurement of <HEAD>Space Charge </HEAD>Induced in Xlpe by <METHOD>Pulsed Electroacoustic Method</METHOD>

[解析結果] Measurement of <HEAD>Space Charge Induced in Xlpe </HEAD>by <METHOD>Pulsed Electroacoustic Method</METHOD>

(5) 不要語の不足(1.3%)

以下の例では，本来ならば"Laboratory Experiment"の個所に OTHER タグが付与されるはずであるが，誤って HEAD タグが付与されている．これは，不要語リストの中に"Laboratory Experiment"という用語が含まれていなかった

ためである．この問題を解決するには，不要語リストをさらに増やす必要があるが，"Laboratory Experiment"のように論文中での出現頻度が高くない表現まですべて網羅するのは容易ではない．このため，今回用いている不要語リストとは別に，不要語を識別するための規則を作成する等の処理が必要となる．

[正解] Laboratory Experiment with <METHOD>a New Wiring Method </METHOD>of <HEAD>Distribution Transformers</HEAD>

[解析結果] <HEAD>Laboratory Experiment </HEAD>with a New Wiring Method of <HEAD>Distribution Transformers</HEAD>

6. おわりに

本研究では，英語論文表題の構造を解析する手法を提案した．英語論文表題には以下の2つの問題点があり，日本語論文表題と同程度の解析精度を得ることができなかった．

(1) 英語論文表題の構造は一般的に日本語のものよりも複雑である．
(2) 構造解析に用いる手がかり語が日本語ほど有効に機能しない．

そこで，日本語論文表題の構造解析結果と翻訳知識を組み合わせた英語論文表題の構造解析手法を提案した．実験の結果，翻訳知識として，要素合成法に基づく専門用語翻訳器を使ったTRANS手法において，精度77.98%，再現率75.15%が得られ，提案手法の有効性が確認された．

謝辞

本研究で用いた論文データは，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．

参考文献

[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.

[近藤 2007] 近藤 友樹，難波 英嗣，奥村 学，新森 昭宏，谷川 英和，鈴木 泰山:「論文データベースからの研究動向情報の抽出」．言語処理学会第13回年次大会，2007.

[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.

[Kando 1999] Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., and Hidaka, S. "Overview of IR Tasks at the First NTCIR Workshop". In Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.11-44. 1999.

[Kando 2001] Kando, N., Kuriyama, K., and Yoshioka, M.: "Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop". In Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, pp.4-37 – 4-60. 2001.

[Taniguchi 2008] Taniguchi, Y. and Nanba, H.: "Identification of Bibliographic Information Written in both Japanese and English". In Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2008. 2008. (to appear)