専門用語として普及しそうな語の自動抽出

辻慶太 *　　芳鐘冬樹 †

Abstract

専門用語辞書の編纂・改訂の効率化，トレンド分析への応用を目的として，学術論文に現れた新語の中から，今後専門用語としてよく用いられるようになる語を抽出する手法を検討した。調査対象語は雑誌 JASIST の 17 年分の論文 1,025 個に現れた 2 名詞単位列である。調査の結果，第 1 単位の出現率が経年的に増加しつつあり，第 2 単位の出現率は一定して高く，かつ両単位の共起度が増加しつつある語は，その後よく用いられる場合が多い傾向が示された。

1　はじめに

テキストからの専門用語自動抽出に関しては様々な研究が行われてきた。それらの多くはテキストが表す分野において，＜現在既に，専門用語として普及している語＞の抽出を目指してきたように思える。本研究は，テキストに新たに出現した語の中から，そのテキストが表す分野において，＜今後，専門用語として普及する語＞の抽出を目指すものである。新たに出現した語は，まだその分野の専門用語とは言い難いという点で，本研究は従来の専門用語抽出研究とは，抽出対象が異なっている。また新語の中から，今後専門用語として普及する語を抽出するという作業は,「その語はそれ以前にはなかった（テキスト中には現れていなかった）」という条件下で行うことから，従来の専門用語抽出手法のいくつかはそのままでは適用できず，技術的にも異なってくる可能性が高い。

上記のような予測が可能になると，まず専門用語辞書の編纂・改訂の際に，ある語を加えるべきか否かを判断する場面で有用である。現代では，専門用語辞書が全くない分野は稀であり，多くの場合，既存の辞書に，新たに生まれた語を追加する形で編纂作業が行われている。新語に対象を絞った予測研究は，そうした現実状況において求められている情報を，効果的に提供し得ると思われる。また，今後よく普及する語が分かれば，ある分野で今後注目を集める研究・トピックが把握しやすくなり，トレンド分析的な面でも有用であろう。

さて上記の抽出手法を開発するには，(1) 経年的に並べられたテキストを用意し，(2) 現代からある程度離れた過去の時点を，調査対象新語出現期間に設定し，(3) その期間に現れた新語のうち，そこから現代にかけて頻出した語を「よく普及した語」，頻出しなかった語を「あまり普及しなかった語」とみなし，調査対象出現期間において両者を区別し得る特徴を探すというのが，１つの明解な方向と思われる。本研究もその方向で行う。そして以下の３点を仮定する。即ち，(a) 調査対象出現期間から現代にかけて，出現した時点は異なっていても，表記が同じ語は同語である，(b) 得られる知見・調査結果は，ある程度過去に関するものとなるが，それらは現代にも通じ，先述の応用場面で有効に機能する，(c) 個々の一般語の出現率は経年的に大きく変化せず，逆に，新語として現れ頻出するようになった語は，一般語でなくその分野の専門用語となった語である，の３点である。

新語の中から専門用語として普及する語を特定する手法の開発という研究が，先述のような有効性・重要性にも拘わらず，これまでほとんど行われなかった理由には，経年的情報を伴った大量の専門分野テキストが少なかったことも挙げられるであろう。だが現代では，多くの学術論文が PDF などの形式で入手でき，処理しやすいテキスト形式に変換して調査することが出来る。本研究では *Journal of the American Society for Information Science (and Technology)* (以下‘JASIST’) の論文を，そのように処理して調査対象コーパスとした。

本研究では複合語を抽出対象とする。また普及するか否かを予測する手がかりには様々なものが考えられるが，今回は主に，語構成要素のテキスト中での言語学的特徴に焦点を絞り，言語外的要因や他分野テキストなどは扱わないことにした。

2　専門用語自動抽出研究の概観

本研究が目指す自動抽出は，従来の専門用語自動抽出手法によって達成出来るかを検証したい。これまでに提案された用語抽出手法としては，まずテキスト中に出現した回数を基本にする手法（Damerau (1993)，Daille ら (1994)，Justeson & Katz (1995)，

*国立情報学研究所　E-mail: keita@nii.ac.jp
†大学評価・学位授与機構　E-mail: fuyuki@niad.ac.jp

Pantel & Lin (2001)，及び Frantzi ら (1998)・Mima & Ananiadou (2000) の C-Value に基づく手法）がある。これらは先述のように，本研究が対象とする，これまでのテキスト中の頻度がゼロであった語にはあまり有効でないように思われる。用語抽出手法としては他に，候補語の出現箇所の周辺に共起する語の性質に基づく手法がある（Hisamitsu ら (2000)，Maynard & Ananiadou (2000)，合原ら (2000)，山田ら (2000)，竹内 & コリアー (2002)，及び Frantzi ら (1998)・Mima & Ananiadou (2000) の NC-Value に基づく手法）が，本研究が対象とする新語は，本来的に出現箇所が少ないので，得られる共起語のサンプルが少ないという問題がある。この手法の検証は今後の課題としたい。また語の表記に注目する手法もあるが（福田 (1997)，山田ら (2000)），大量の正解専門用語の入手・精査を必要とするので，この検証も今後の課題とする。最後に，用語抽出手法として，語構成要素の前後に連接する語の異なり数に注目する手法がある（Nakagawa (2000)，湯本ら (2001)）。今回は Nakagawa (2000) の手法を比較対象として取り上げたい。

3 調査・実験

以下ではまず本研究が分析対象としたデータを説明し，次に予備調査と抽出実験の結果を述べる。

3.1 データ

本研究では JASIST において，ある期間に発表された論文中の語の集合は，ある 1 分野のある期間の語の集合であるとみなして，分析の対象とした。期間は 1986 年〜 2002 年の計 17 年で，1,025 論文の本文部分を対象とした。テキストは Brill tagger で品詞付けを行い，‘NN’，‘NNS’，‘NNP’ と判定された語を名詞として扱った。語・名詞の具体的な数量は表 1 の通りである（名詞はすべて単数形に統一した）。

全延べ語数	8,760,627
全異なり語数	226,353
名詞の延べ語数	2,455,928
名詞の異なり数	132,545

表 1: JASIST 中の語・名詞に関する基本データ

3.2 予備調査

本研究では，上記 1,025 論文を，発表時期の観点から 2 つに分け，後半（現代に近い方）の 513 論文のみに偏って現れる語を，「JASIST に表される分野に新たに現れた専門用語」とみなし，それらが初出時点において，どのような言語学的特徴を持っていたかを分析する。より具体的には，513 論文のうち，最初の 72 論文（1997 年 6 号から 1998 年 5 号までの 1 年分）に初めて現れた 2 名詞列 S_1S_2 を調査対象とし，それらのうち，後半部分で 10 論文以上に現れた 2 名詞列を「専門用語としてよく普及した語」，即ち抽出すべき正解語とし，9 論文以下しか現れなかった 2 名詞列を「普及しなかった語」，即ち抽出すべきでない不正解語とし，両者を区別し得る特徴がないかを調べる。[1]

ここで，「全論文に占める語 w の出現論文の割合」を $p(w)$ と定義し，「全論文に占める語 w_1,w_2 の共起論文の割合」を $p(w_1, w_2)$ と定義する。さらに $D(w_1, w_2) = p(w_1, w_2) * 2/(p(w_1) + p(w_2))$ とする。初出時から 3 年毎に過去に遡って，2 名詞 S_1，S_2 の，これら p と D の平均値を調べたところ，表 2 のようになった。表で「比」とは「9–6 年前の p，D の平均」に対する「3–0 年前の p，D の平均」の比を表している。表 2 から例えば，今回抽出すべき正解語となったのは 16 語であり，それらの語構成要素 S_2 の平均出現率は 3 期間に渡って 0.40 以上を保ち，また S_1，S_2 の 3–0 年前の平均共起度は，9–6 年前に比べて 1.87 倍になっていること，それに対して先述の 72 論文に 1 回現れただけで以後全く現れなかった 2 名詞列は 8,523 個あり，それらの語構成要素の平均共起度の「比」は 1.17 倍にとどまっていること，などが分かる。

3.3 自動抽出実験

以下では Nakagawa (2000) の手法による自動抽出実験結果を示し，次に前節の予備調査に基づく本研究の手法による結果を示す。まず Nakagawa (2000) は次の尺度の値が高い名詞列 $S_1S_2...S_k$ を専門用語として抽出することを提案している。

$$Imp_1(S_1S_2...S_k) = (\prod_{i=1}^{k}((Pre(S_i)+1)(Post(S_i)+1)))^{\frac{1}{k^a}}$$

ここで $Pre(S_i)$，$Post(S_i)$ は S_i の前後にそれぞれ連接した名詞の異なり数を表す。本研究では Nakagawa (2000) の実験で最も結果が良かった a=1 を

[1] 3 名詞以上の連続列で 10 論文以上に現れたのは ‘Web search engine’ のみとサンプルが少なかったため今回は取り上げなかった。また単名詞や名詞以外については，後述のように今後の課題とし，今回は取り上げない。

論文数	語数		9–6 年前	6–3 年前	3–0 年前	比
10+	16	$p(S_1)$	0.28	0.32	0.37	1.30
		$p(S_2)$	0.40	0.45	0.47	1.16
		$D(S_1, S_2)$	0.026	0.041	0.048	1.87
7–9	20	$p(S_1)$	0.30	0.34	0.38	1.29
		$p(S_2)$	0.39	0.42	0.44	1.12
		$D(S_1, S_2)$	0.133	0.148	0.139	1.05
4–6	173	$p(S_1)$	0.37	0.40	0.43	1.16
		$p(S_2)$	0.40	0.44	0.47	1.16
		$D(S_1, S_2)$	0.057	0.065	0.066	1.16
3	227	$p(S_1)$	0.31	0.35	0.37	1.16
		$p(S_2)$	0.40	0.44	0.46	1.14
		$D(S_1, S_2)$	0.033	0.041	0.041	1.22
2	867	$p(S_1)$	0.30	0.33	0.35	1.13
		$p(S_2)$	0.36	0.39	0.41	1.13
		$D(S_1, S_2)$	0.035	0.040	0.040	1.14
1	8,523	$p(S_1)$	0.20	0.22	0.23	1.12
		$p(S_2)$	0.27	0.29	0.30	1.12
		$D(S_1, S_2)$	0.013	0.014	0.015	1.17

表 2: 出現論文数毎の名詞 S_1,S_2 の出現率・共起度

抽出条件	精度	再現率	F 値
$Imp1 \geq 100$	0.22	87.50	0.43
$Imp1 \geq 1,000$	0.40	75.00	0.79
$Imp1 \geq 5,000$	0.37	25.00	0.72
$Imp1 \geq 10,000$	0.19	6.25	0.36

表 3: Nakagawa (2000) の手法の結果

採用した。結果は表 3 のようになり，それほど良くはなかった。[2]

前節の予備調査で，後半 10 論文以上に現れていた 2 名詞列の，6 〜 9 年前の共起度に対する 0 〜 3 年前の共起度の比（以下 'gco' と呼ぶ）は，他の 2 名詞列のそれに比べて高かった。そこで gco が 1，2，3，5 以上の 2 名詞列を抽出してみた。だが結果は表 4 のようになり，Nakagawa (2000) の手法より F 値は高いものの，満足できる値ではなかった。

2 名詞が文中で共起する度合いでは単純過ぎるこ

抽出条件	精度	再現率	F 値
$gco \geq 1$	0.52	68.75	1.03
$gco \geq 2$	0.56	50.00	1.10
$gco \geq 3$	0.55	43.75	1.08
$gco \geq 5$	0.59	43.75	1.17

表 4: 共起度の増加率に基づく抽出結果

[2] ちなみに，何の条件も設けずに，今回調査対象とした 2 名詞列をすべて抽出した場合は，精度 0.16%（再現率 100%），F 値 0.33 である。

抽出条件	精度	再現率	F 値
$giv \geq 1$	0.25	75.00	0.49
$giv \geq 2$	0.39	68.75	0.78
$giv \geq 3$	0.46	62.50	0.92
$giv \geq 5$	0.43	43.75	0.84

表 5: 共起語集合の類似度の増加率に基づく抽出結果

抽出条件	精度	再現率	F 値
$oc2 \geq 0.7$, $go1 \geq 2$, $gco \geq 2$	4.08	12.50	6.15
$oc2 \geq 0.7$, $go1 \geq 3$, $gco \geq 3$	6.25	12.50	8.33
$oc2 \geq 0.8$, $go1 \geq 2$, $gco \geq 2$	5.41	12.50	7.55
$oc2 \geq 0.8$, $go1 \geq 3$, $gco \geq 3$	8.00	12.50	9.76

表 6: S_2 の出現率，S_1 の出現率の増加率，及び S_1S_2 の共起度の増加率に基づく抽出結果

とを考え，2 名詞がそれぞれ文中で共起する名詞の集合の類似度を，共起語ベクトルの内積という観点で測り，その値の経年的な増加率が高い 2 名詞列を抽出することを試みた。即ち，ここで N_i を全論文中の異なり名詞（$1 \leq i \leq M$）とし，ベクトル $V(w)$ を $(D(w,N_1), D(w,N_2), ...D(w,N_M))$ とする。そして 2 名詞列 S_1S_2 のうち，6 〜 9 年前の $V(S_1)\cdot V(S_2)$ に対する 0 〜 3 年前の $V(S_1)\cdot V(S_2)$ の比（以下 'giv'）が高いものを抽出した。即ち，共起する名詞の集合が，年が経つにつれ似てきた 2 名詞列というものを抽出した。だが結果は表 5 の通りで，先ほどの単純な共起度の増加率に基づくものよりも，全体に F 値は低かった。

前節の予備調査で，後半 10 論文以上に現れる 2 名詞列 S_1S_2 の性質として，第 2 単位 S_2 の出現率が一定して高く，第 1 単位 S_1 の出現率が経年的に増加している傾向を見た。そこでそれらの条件も組み合わせたところ，抽出結果は表 6 のようになった。表 6 で $oc2 \geq X$ は S_2 の出現率が 3 期間とも X 以上であることを表し，$go1 \geq X$ は S_1 の出現率の「比」が X 以上であることを表している。gco は先述の通りである。表 6 から，条件を組み合わせることで，これまでに比べれば F 値がかなり良くなることが分かる。専門用語の性質として 'termhood' と 'unithood' が挙げられることがあるが (Kageura & Umino (1996))，第 1 章の仮定 (c) に基づくならば，termhood を帯びつつある語を第 1 単位にし，かつ unithood を高めつつある語が，専門用語として普及しやすいと言うことも出来よう。

10 論文以上に現れた 2 名詞列	誤って抽出された 2 名詞列の例
Web search (71)	anchor type (1)
Alta Vista (28)	chat system (1)
Web user (28)	collaborative research (2)
data mining (25)	collaborative work (3)
Dublin Core (19)	drug problem (2)
information visualization (17)	E-R data (1)
usability study (11)	formatting information (1)
visualization system (10)	HCI problem (2)
image feature (10)	HCI information (2)
image description (10)	jasa system (1)
information architecture (10)	math problem (1)
Internet site (10)	rectangle example (1)
interaction model (10)	RUG study (1)
segmentation method (10)	server information (2)
Web interface (10)	summarization process (2)
p q (10)	TRF term (2)

表 7: 2 名詞列の例（括弧内は出現論文数）

表 7 に，今回正解抽出対象とした 16 個の 2 名詞列と，$oc2 \geq 0.8$，$go1 \geq 3$，$gco \geq 3$ という条件で誤って抽出された 2 名詞列の例を挙げた。前者には ‘Web user’ や ‘visualization system’ のように，第 1 単位が増加傾向にあり，第 2 単位は一般的な語という 2 名詞列が散見される。後者には頭字語が見られ，また ‘work’ や ‘example’ のように，一般に専門用語を構成しにくい語が第 2 単位になっているものが見られる。従って今後，表記に関するヒューリスティックスや，禁則語を導入することを検討したい。

4 おわりに

専門用語辞書の編纂・改訂や，トレンド分析での応用を想定して，新語の中から専門用語として普及する語を自動抽出するという問題設定のもと，本研究では，新たに現れてから 10 論文以上に現れた 2 名詞列と，9 論文以下にしか現れなかった 2 名詞列の性質の違いの一端を明らかにした。そうした性質の違いに基づく自動抽出実験の精度・再現率は，従来の専門用語自動抽出手法の 1 つよりは高いものであったが，まだ改善の必要がある。その方策としては次の 4 つを挙げたい。即ち，(1) 名詞の表記上の特性や初出時の文章表現の特徴といったヒューリスティックスの利用，(2) 初出させた著者の特性といった言語外的要因の考慮，(3) 他の分野のテキストコーパスの利用，(4) 用語抽出に関する他の先行研究手法の利用，の 4 つである。さらに今後の研究方向として，SIGIR や TREC の Proceedings といった，今回の JASIST と分野的に重なる部分があり，かつ新語がより早い時期に現れていると思われるテキストを調査対象としてみたい（現在，SIGIR はデータ整備を終えつつある）。また今回は取り上げなかったが，単名詞に関しても ‘XML’，‘portal’，‘applet’ といった興味深い語が後半 10 論文以上に現れていた。それらや，また名詞以外の語も対象に，今後研究を進めたい。

謝辞

本研究の一部は「科学研究費補助金若手研究 (B)」によるものであり，ここに謝意を表します。

参考文献

[1] Brill tagger http://www.cs.jhu.edu/~brill/

[2] Daille, Béatrice, Gaussier, Éric and Langé, Jean-Marc (1994) “Towards Automatic Extraction of Monolingual and Bilingual Terminology,” *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, p.515-521.

[3] Damerau, Fred J. (1993) “Generating and Evaluating Domain-oriented Multi-Word Terms from Texts,” *Information Processing & Management*, vol.29, no.4, p.433-447.

[4] Frantzi, Katerina T., Ananiadou, Sophia and Tsujii, Jun-ichi (1998) “The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms,” *Proceedings of the Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL'98*, p.585-604.

[5] Hisamitsu, Toru, Niwa, Yoshiki, Nishioka, Shingo, Sakurai, Hirofumi, Imaichi, Osamu, Iwayama, Makoto and Takano, Akihiko (2000) “Extracting Terms by a Combination of Term Frequency and a Measure of Term Representativeness,” *Terminology*, vol.6, no.2, p.211-232.

[6] Justeson, John S. and Katz, Slava M. (1995) “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text,” *Natural Language Engineering*, vol.1, no.1, p.9-27.

[7] Kageura, Kyo and Umino, Bin (1996) “Methods of Automatic Term Recognition: A Review,” *Terminology*, vol.3, no.2, p.259-289.

[8] Maynard, Diana and Ananiadou, Sophia (2000) “Identifying Terms by their Family and Friends,” *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, p.530-536.

[9] Mima, Hideki and Ananiadou, Sophia (2000) “An Application and Evaluation of the C/NC Value Approach for the Automatic Term Recognition of Multi-word Units in Japanese,” *Terminology*, vol.6, no.2, p.175-194.

[10] Nakagawa, Hiroshi (2000) “Automatic Term Recognition based on Statistics of Compound Nouns,” *Terminology*, vol.6, no.2, p.195-210.

[11] Pantel, Patrick and Lin, Dekang (2001) “A Statistical Corpus-Based Term Extractor,” *Proceedings of Advances in Artificial Intelligence, 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI 2001)*, p.36-46.

[12] 合原博, 宮田高志, 松本裕治 (2000) “医学生物学分野からの専門用語の抽出・分類,” 情報処理学会研究報告, NL-135, p.41-48.

[13] 竹内孔一, コリアー・ナイジェル (2002) “生物学文献からの専門用語抽出における機械学習モデルの検討,” 情報処理学会研究報告, NL-150, p.185-190.

[14] 福田賢一郎, 角田達彦, 田村あゆち, 高木利久 (1997) “医学生物学文献からの専門用語の抽出,” 情報処理学会研究報告, NL-121 FI-47, p.103-110.

[15] 山田寛康, 工藤拓, 松本裕治 (2000) “単語の部分文字列を考慮した専門用語抽出と分類,” 情報処理学会研究報告, NL-140, p.77-84.

[16] 湯本紘彰, 森辰則, 中川裕志 (2001) “出現頻度と連接頻度に基づく専門用語抽出,” 情報処理学会研究報告, NL-145 FI-64, p.111-118.