

連接異なり語数による専門用語抽出

大畑博一 中川裕志 †
横浜国立大学 工学部 † 東京大学 情報基盤センター

1999年3月にNACSISにより主催されたNTCIRコンテストの用語抽出タスクはNACSIS論文データベースから取り出された1,870の抄録を対象に用語を抽出するというものである。本稿で目的とするのは用語抽出タスクにおいてより高い精度の結果を出せる抽出方法の提案と実験の評価である。ここで提案する用語抽出における基本単語の順位付けのアイデアは単語がその前後に何種類の単語を接続して複合語を作るかという尺度である。この方法を複合語の順位付けに拡張した専門用語を抽出する方法を用いて、「事例の予測」「常識的な知識」のような「AのB」「形容詞+名詞」のものを用語候補とする方法、しない方法および種々のチューンアップを行なった方法について実験しNTCIRデータでの評価したところ、良好な結果を得られたので報告する。

Automatic Term Recognition by the Relation between Compound Nouns and Basic Nouns

Hirokazu Ohata, Hiroshi Nakagawa†
Faculty of Engineering, Yokohama National University
†Information Technology Center, The University of Tokyo
ohata@naklab.dnj.ynu.ac.jp
† nakagawa@r.dl.itc.u-tokyo.ac.jp

The term recognition task initiated by NTCIR TMREC group, held in March, 1999, have provided us with the test collection of term recognition task. The goal of this task is to automatically recognize terms from data collection which consists of 1,870 abstracts extracted from the NACSIS Academic Conference Database. This paper describes the term extraction method we have developed to extract compound nouns in a very flexible manner. The basic idea of scoring a basic noun:N of our term extraction system is to count how many nouns attached the noun:N to make compound nouns. In this scoring, we have variations such as including a patters of A no B (A の B), or not, preferring longer compound noun or shorter compound nouns. By tuning up these variations, we get the almost highest score for NTCIR TMREC results.

1 はじめに

情報検索の世界では自動索引(キーワード抽出)の研究が長く行なわれてきたが、

近年のコーパス利用の流れにのって、専門用語の自動抽出というテーマが自然言語処理の分野で研究されるようになってきた。自動索引と自動用語抽出の違いは

(影浦, 1997) で以下のように述べられている、

索引語 : 想定するデータベースにおいて、個々の文献あるいは文献集合を特徴づける語

専門用語 : 想定する「分野」を特徴付ける語彙および／あるいはその要素。
よって、

専門用語抽出 想定する「分野」を特徴付ける語彙を想定母集団として、そこからの何らかのサンプルに相当する語彙を抽出すること。

と定義できるが、「索引語」と「専門用語」は、実際の要素としてはしばしば一致するため両者の抽出技法はあまり区別されていない。

日本語文章において用語を抽出する際には、

1. 単語と単語に英語のスペースのような明示的な区切りがない
2. 複合語が無限に造語されるため、複合語とそれらを構成する単語の関係の処理が複雑である
3. 字種が多い

などの問題がある。

これまでに提案されている複合語キーワード抽出法には、単語毎にキーワード性を判定し、その連続部分を複合語として抽出する方法(会森, 依田, 高原, 1988; 山口 杉山, 1988) や、キーワードパターンにマッチングする単語の連続部分を複合語として抽出する手法(小川, 望主, 別所, 1997) などがある。キーワード抽出法により取り出されるキーワードのスコア付けは、(Kageura, 1996) によれば

1. ある表現 (*Collocation* や複合語など) がテキストデータベース中で安定して使用される度合を表す *unithood* 基準

2. ある表現が対象分野固有の概念をどれだけ高い関連性をもって表現するかを表す *termhood* 基準

の2つの基準に分ける事ができる。

既存の *unithood* 基準の手法としては、コロケーションの安定度をまとまった形で高頻度で出現していることにより評価する *Nested Collocation* 方式 (Frantzi & Ananiadou, 1996) が挙げられる。

termhood 基準の方法としては、固有の概念というものを局所的に出現していることで表している *tf・idf* 法が挙げられる。NTCIR TMREC タスクのように分野を特徴付ける語彙を抽出する方法としては NTCIR コンテストには異分野コーパスを利用したもの(内元, 関根, 村田, 小作, 井佐原, 1999) や、電子辞書を用いた方式 (Fukushige & Noguchi, 1999) 等がある。また、世界的に見れば、2か国語の並行コーパスを用いる方法 (Daille, Gaussier, & Lange, 1994) など研究されている。

さて、以上のような用語抽出の研究の流れを受けて 1999 年 3 月から 8 月にかけて学術情報センターが中心になって用語抽出タスクに関するコンテストが催された。その詳細については、(Kageura K, 1999) に詳しい記述があるので参照されたい。

本論文では NTCIR の用語抽出タスクと同じく単一言語コーパスからの専門用語抽出する方法として、*termhood* 基準を基本ポリシーとして使用する方法 (Nakagawa, 1997) (Nakagawa & Mori, 1998) において、方式上のパラメーター調整などを行ない、NTCIR の用語抽出タスクに適合させる実験を行なった。このような調整はテストコレクションに過度に適合させるだけで、技術的な本質とは無関係であるというような批判もある。しかし、NTCIR の用語抽出タスクのテストコレクションにおける正解の用語群は対象分野の専門用語辞典など利用しており、これに適合するような調整は決して無意味ではない。また、我々の提案する方式は

固定された単一の方式ではなく、順位付けの方法の選択の幅が広く、柔軟性がある。NTCIRの用語抽出タスクをテストコレクションとして用いれば、この柔軟性がどの程度、強力に作用するかを確認できる。よって、我々の方式の本質を評価するために役立つ。以上のような考察により、以下の実験結果について報告する次第である。

2 用語抽出方法

本節では我々が検討対象としている用語の候補語を順位付けする方法について述べる。コーパス中には非常に多くの複合名詞や単名詞が存在するが、それら全てが必ずしも用語として適当なものではない。そこで与えられた複合名詞や単名詞がどれだけ用語性を備えているかを表す尺度を見つめる必要がある。そこで我々は以下に示す単名詞の接続数と名付けた尺度を用いる。

2.1 接続方式

ここでは、「辞書」と「ファイル」という単名詞を接続した「辞書ファイル」は複合名詞と呼ぶ。これに加えて「の」でつながった「辞書ファイルのシステム」のような名詞句も検討の対象範囲に含めて考えることにする。

実際、日本語の文書においては非常に多くの複合名詞が用いられるが、この中から必要な用語を選び出すための尺度として一つの単名詞の前後にいくつの単名詞が接続されて複合名詞を形成するかを用いた。これを前方接続数、後方接続数とし、以下のように定義する。

与えられたコーパスにおいて、単名詞 N の前方接続数 $Pre(N)$ は、さまざまな複合名詞中において、単名詞 N の直前に、もしくは「の」をはさんで前に、つながることがあった単名詞の種類数とする。同様、単名詞 N の後方接続数 $Post(N)$ は、さまざまな複合名詞中において、単名詞 N に直後に、もしくは「の」をはさんで後ろに、つながることがあった単名詞の種類数とする。但し、「AのB」の場合「の」自体の接続数は数えず「AB」とみなしてAとBの接続数数えることにする。

ここで単名詞の接続数について考えてみよう。ある学問ないし技術分野においては概念を表すために多数の複合名詞が用語として使われる。では、なぜ複合名詞なのか。概念を表すための言語表現として単名詞がまず考えられるものの、単名詞を創造することは容易なことではない¹。大量に存在する分野固有の概念は大方の場合、既存の単名詞ないし複合名詞を組み合わせで作られる複合名詞によって表現される。

そこで、ある名詞がその分野において重要かつ基本的な概念であるなら、その概念から派生する様々な概念を、その名詞を含む複合名詞として表現する。つまり、多数の複合名詞の要素になっている名詞は重要度の高い名詞であるといえる。この重要度は前に定義した前方及び後方接続数で直接に表されている。まとめると次の仮定となる。

$Pre(N)$ 、 $Post(N)$ の値が高い単名詞 N は与えられたコーパス中で用いられた主要な概念を表す。

この仮定は、本論文の中で後に、この仮定に基づいて抽出した用語がよい評価を得ることにより実証されることになる。

¹筆者の1名は約半世紀を生きているにもかかわらず、ひとつの単名詞も創造していない。

$$Imp(N_1 N_2 \cdots N_k) = \left(\prod_{i=1}^k ((Pre(N_i) + 1) \cdot (Post(N_i) + 1)) \right)^{\frac{1}{k^a}} \quad (1)$$

$$Imp(N_1 N_2 \cdots N_k) = \frac{1}{k^a} \cdot \sum_{i=1}^k (Pre(N_i) + Post(N_i)) \quad (2)$$

次に、複合語 $N_1 N_2 N_3 \cdots N_K$ の重要度の尺度 $Imp(N_1 N_2 N_3 \cdots N_K)$ は Pre 、 $Post$ の関数として定義する。これには無数の定義法があるが、ここでは相乗平均(1)式もしくは相加平均(2)式を使った(1)(2)式で表される定義 Imp を用いた。ここで a はこの方式における重要なパラメータである。(1)式及び(2)式から容易に分かるように a を大きくするほど、短い複合名詞が高い Imp 値となる。つまり、 a の値を適宜変えることによって、基本名詞(すなわち 長さ=1 の複合名詞)および複合名詞の順位付け方法を変えることができる。

2.2 ナ形容詞の獲得

NTCIR 作成の正解用語には「経験的な知識」「ローカルなデータ」のような(ナ形容詞) + (名詞)のパターンからなる用語も含まれていた。名詞だけを用語候補とするシステムでは、このパターンの用語は抽出できない、今回は実験的に上記のパターンを用語候補とするシステムについての評価も報告する。なお、この場合も「経験的な」のようなナ形容詞も名詞と同様に扱うことによって接続数をカウントし、同じ計算式(1)あるいは(2)を用いてスコア付けしている。

3 実験および評価

今回実験に用いたシステムは用語候補の取り方として以下の4種類を用いた。

A: 「の」で接続型 抽出できる用語例:
「解析システム」「未知の規則」

B: 「の」で接続 + ナ形容詞型 抽出できる用語例: 「解析システム」「未知の規則」「経験的な知識」

C: 「の」で接続しない型 抽出できる用語例: 「解析システム」

D: 「の」で接続しない + ナ形容詞型 抽出できる用語例: 「解析システム」「経験的な知識」

以上の4種類の実験において(1)式、(2)式の a を変化させ順位付けを行い、上位13000語に対して評価を行ない、最適な a 値を求めた。

使用したデータ

NTCIR の用語抽出タスクに用いられた形態素タグ付きのコーパス(人工知能分野: 1,870 個の抄録集合)と NACSIS 作成の正解用語 8843 語である。

実験結果

まず、各方式とも上位13000語まで採用した場合の完全一致の F 値が最も良かった a での評価を表1に示す。また、表2においては NTCIR の上位グループとの比較を示す。この表で Z と W は NTCIR コンテストで最上位クラスのグループの結果である。

表中の記号説明

1st char - F : 完全一致数 ; I : 抽出結果が正解を含む ; P : 抽出結果が正解の一部 ; B : F+I ;

2nd char - R : Recall ; P : Precision ; F : F-value ;

$$F - value = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (3)$$

表 1: 各方式の評価

	a	Total	F	I	P	FR	FP	FF	BR	BP	BF
A相乗平均	5	13000	4767	5805	682	0.539	0.366	0.436	0.594	0.813	0.686
A相加平均	6	13000	4854	5463	699	0.548	0.373	0.444	0.597	0.793	0.681
B相乗平均	5	13000	4705	5835	666	0.532	0.361	0.430	0.590	0.810	0.683
B相加平均	6	13000	4748	5503	679	0.536	0.365	0.434	0.595	0.788	0.678
C相乗平均	1	13000	6362	4825	429	0.719	0.489	0.582	0.817	0.860	0.838
C相加平均	3	13000	6337	4441	742	0.716	0.487	0.580	0.775	0.829	0.801
D相乗平均	1	13000	6010	5400	333	0.679	0.462	0.550	0.796	0.877	0.835
D相加平均	3	13000	5977	4931	650	0.675	0.459	0.547	0.747	0.839	0.790

表 2: 他の抽出システムとの比較

	Total	F	I	P	FR	FP	FF	BP
連接方式	16000	7367	5197	923	0.833	0.460	0.593	0.785
Z	16112	6536	5464	690	0.739	0.405	0.524	0.744
W	23270	7944	9432	879	0.899	0.341	0.494	0.746

表 1によれば、まず (1) 式の相乗平均と (2) 式の相加平均では大きな差はない。また、「の」を挟む連接を含む A,B 方式は、これを含まない C,D 方式より結果が劣る。さらに C,D 方式を比べると、結局、ナ形容詞も含まず、単に名詞が連接する複合名詞だけを採用する C 方式の結果が一番よい。正解用語の中には「の」で連接する名詞句やナ形容詞を含む名詞句も存在するが、少数であり、これらを含めることによって不正解語を多数抽出してしまうことになっている。表 2によれば、上位約 16000 語を選んだ場合においては我々の用語の順位付け方法は完全マッチ

において最も優れている。また、完全マッチの F 値においては上位 23000 語を選ぶ W の結果よりよい。この結果、我々の方法は NTCIR の参加グループの中でもほぼ最上位に位置するといえる。

次に、表 1で最高 F 値を取った方式での 1 千語毎の評価グラフを図 1に、各方式の a 値の変化に対する F 値のグラフを図 4,5に示す。

図 1より、最高で完全一致 F 値 0.5930、部分一致を含む F 値 0.8291を得た。なお、抽出語数を増やしていった場合、完全一致 FF では 14000 語くらいでは飽和し、以後は微小な改善にとどまる。一方、部分一致 BF 場合は、むしろ 14000 語を越えるあたりからわずかに劣化する。

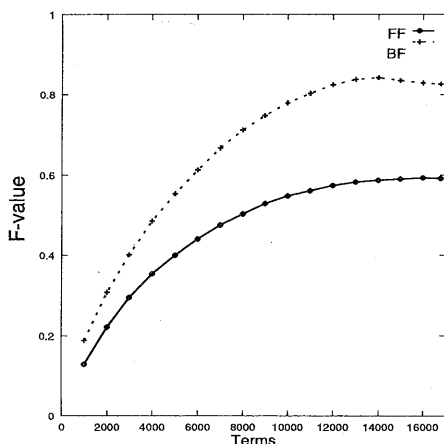


図 1: 「の」で接続しない型: 相乗平均 ($a=1$) の用語候補数千語毎の評価
FF: 完全一致 F 値
BF: 部分一致含めた F 値

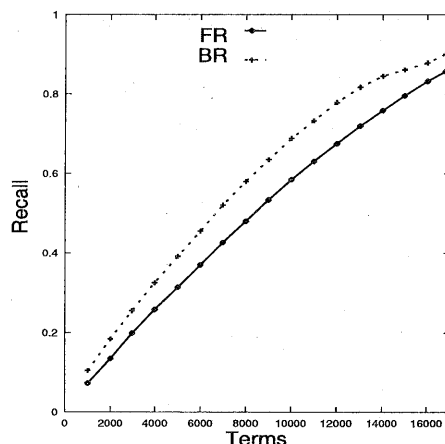


図 2: 「の」で接続しない型: 相乗平均 ($a=1$) の用語候補数千語毎の再現率
FR: 完全一致再現率
BR: 部分一致含めた再現率

図 1 に対応する再現率、適合率を示した図 2 と図 3 によれば、これは多数の語を採用すると再現率は上がるものの、不正解語が多くなり、部分一致においても適合率の劣化の割合の方が大きくなってしまいうためである。

図 4,5 からは次のことが分かる。A,B では大きい a で最良の結果がでている。これは「の」接続を含む名詞句も含む場合は、全体に長い名詞句が増えるため、むしろ短い複合名詞、名詞句を採用するほうがよい結果となるため a が大きいところで最適値が得られていることを示す。一方、C,D の場合は「の」による接続を含まないため、基本的には短い候補が多くなるため、 a を大きくして長い複合名詞を避ける必要はない。 $a = 1$ は長さで正規化した Imp 尺度であるため、接続方式本来の順位付けが十分に機能しているのが C,D の場合であるといえよう。また、F 値の絶対値を見ても、今回の実験では「の」の接続やナ形容詞を含まない場合が F 値としては最もよい結果だが、

これはある程度、使用した NTCIR のテストコレクションの性質に依存した問題であろう。すなわち、正解用語には「(名詞) の (名詞)」 「(ナ形容詞)+(名詞)」のパターンがそれほど多く含まれていない (400 語ほど) ために、新しく抽出できる正解用語よりも、パターンを増やした結果、抽出してしまう不正解語が多いことが挙げられる。ただし、「の」での接続「ナ形容詞」の獲得する方式は、NAC-SIS 作成の正解用語よりも長い単位で用語抽出したい時には、有用に活用できる。なお、全ての方式に共通して、 a 値を大きくすると、部分一致も含めた F 値が下がっていく結果となる。

4 まとめ

接続方式という複合名詞の順位付け方を提案した。この方式は NTCIR の用語抽出タスクのテストコレクションにおいては良好な結果を与えることが実証できた。今回の実験では「の」での接続語、

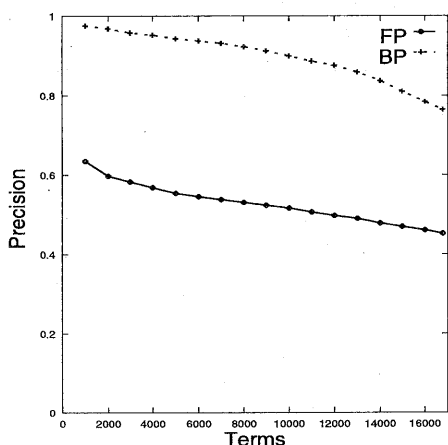


図 3: 「の」で接続しない型: 相乗平均 ($\alpha=1$) の用語候補数千語毎の適合率
FP: 完全一致適合率
BP 部分一致含めた適合率

「ナ形容詞」を含む名詞句を含ませることは抽出精度を下げる結果となったが、将来このシステムで抽出した用語を用いて英語との対訳用語辞書作成を目指している。そのためには、正解とは何かを考え直す必要があるし、その暁には「の」接続やナ形容詞を含む結果の方が望ましいかもしれない。

なお、最も評価がよかった方法において、部分一致を含めても抽出に失敗した正解語が 875 語であった。この接続情報だけでは抽出出来ない正解用語や現システムの適合率向上には、頻度情報も考慮してスコア付けする方式の検討が必要である。今後、頻度情報を用いて抽出精度が改善されれば機会を改めて発表する予定である。

参考文献

Daille, B., Gaussier, E., & Lange, J. M.

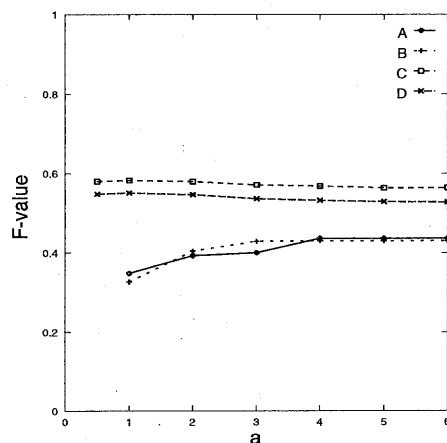


図 4: 各方式の α 値と完全一致 F 値の関係グラフ

(1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of COLING'94*, pp. 515 – 521.

Frantzi, T. & Ananiadou (1996). Extracting Nested Collocations. Tech. rep., COLING'96.

Fukushige, Y. & Noguchi, N. (1999). NT-CIR Experiments at Matsushita TMREC Task. *NTCIR Workshop 1*.

Kageura, K. B. (1996). Methods of automatic term recognition: A review. Tech. rep., Terminology 3.

Kageura K, e. a. (1999). *TMREC Task*. National Center for Science Information Systems.

Nakagawa, H. (1997). Extraction of Index Words from Manuals. In *Proceedings of RIAO'97*, pp. 598 – 611.

Nakagawa, H. & Mori, T. (1998). Nested Collocation and Compound Noun

小川泰嗣, 望主雅子, 別所礼子 (1997). “複合語キーワードの自動抽出法.” 自然言語処理, 97-15, 1993-9-17.

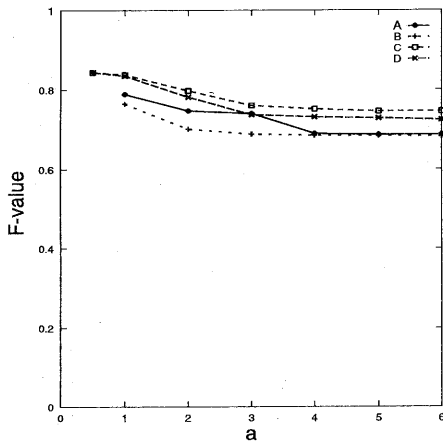


図 5: 各方式の a 値と部分一致も含めた F 値の関係グラフ

For Term Extraction. In *Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98)*, pp. 64 – 70 Montreal, Canada.

山口義一 杉山時之 (1988). “自然言語による索引語自動抽出システムの概要とその索引語の分析.” 科学技術文献サービス, No.85, 31-40.

影浦峽 (1997). “自動専門用語抽出の諸問題.” テクニカル・レポート, TPB フォーラム.

会森清, 依田透, 高原哲 (1988). “日本語キーワード抽出システムの開発および今後の課題.” ドキュメンテーション・シンポジウム予稿集, 15-19.

内元清貴, 関根聡, 村田真樹, 小作浩美, 井佐原均 (1999). “異分野コーパスを用いた用語抽出.” *NTCIR Workshop 1*.