語基の連接情報を用いた専門語抽出

湯本　紘彰*1　大畑　博一*2　森　辰則*3　中川　裕志*4
横浜国立大学*1,2,3　東京大学*4

1.はじめに

専門用語の自動抽出は、ある学術的、技術的な分野のコーパスから特定の語を抽出することが目的である。主要な専門語はたいてい複合名詞であり、複合語の統計に基いた専門語の自動抽出の手法はいろいろ成されてきた。専門語の重要な面は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という“ターム性”の観念（Kageura & Umino 96）である。よって我々が最も追求すべき用語抽出の手法はターム性の観念を用いなければならない。前に述べたとおり、専門語の主要なものは複合名詞からなっており、85%の専門語が複合名詞であるといわれている。残りの 15%は単名詞である。語基をそれ以上分割できない最も基本的な単語と定義すると、そのうちの名詞を単名詞と呼ぶ。すなわち、複合名詞とはこの 15%の単名詞から構成されているといってよい。我々は単名詞と複合名詞の関係、つまり複合語を構成する単名詞に注目した。また、複合語を作るこのような手法は、ターム性の観念と密に関係していた。理由として以下のようなことが挙げられる。まず、ターム性はたいてい語の頻度とその傾向によって計算される。しかしながらこの計算値は近似値であり表面上の統計であるため、直接なターム性ではない。次にある文書を扱う際、その分野のキーとなる概念を表す単名詞 N があったとしよう。著者は単名詞 N をそのまま使わず、N を含んだ複合語をたくさん使うだろう。よって、我々は新しい専門語の自動抽出手法を導くために単名詞と複合名詞の関係に着目する。この関係を利用した最初の試みはある単名詞の左、もしくは右側に連接する名詞の数を区別するという手法で（Nakagawa & Mori 98）によって成された。Nakagawa は語候補のスコア付け関数を定義し、NTCIR1 TMREC タスクにおいて高い再現率、適合率を示した。しかしながら、この専門語自動抽出手法は単名詞と複合名詞の関係を利用したひとつの例にすぎない。また、この手法は本質的に名詞バイグラムに基いている。本稿では名詞バイグラムに基いたこの関係を統合し、NTCIR1 TMREC のテストコレクションを使用していくつかのスコア付け関数を実験的に評価する。この結果から複合語を構成する単名詞を利用することが、専門語自動抽出において有効であることが示された。以下、2 章で専門語自動抽出の背景について述べ、3 章で名詞バイグラムによる抽出する語のスコア付け関数について記述し、4 章で実験的な評価を示す。

2.背景

2.1 専門語自動抽出の標準的手順

用語抽出は一般的に 2 つの手順からなっている。1 つめはコーパスから語の候補を抽出することである。2 つめは最初の手順によって抽出された語の候補にもっともらしくスコア付けし、それに従ってランク付けを行う。しかしながら、ランク付けされた語候補のうちどれくらいを選択するかという問題は、いまだ未解決である。実際にどれくらい語を抽出したいのか、誰が何の目的で抽出された語を使用するのか、というような要因のためこの問題を明確に定義することはできない。

2.2 候補語抽出

専門語自動抽出においてまず、すべきことは与えられたコーパスから候補となる語を抽出することであるが、これには主に 2 つの方法が挙げられる。NTCIR1 TMREC タスクにおいては、すでに part of speech(POS) tagging を含む形態素解析済みのコーパスを使用し単名詞、複合名詞が抽出すべき語の候補として選択される。

2.3 スコア付け

コーパスから候補語が抽出されると、次にターム性の観念からスコアを割り当て、順にランクを付けねばならない。しかし、与えられた語候補から直接ターム性を図ることは難しい。なぜなら与えられた文章においてその語が特別重要であるということは、著者のみが知っていることであるからである。多くの研究者たちは、ターム性を近似するスコア付けの定義を行い事実、これらの多くは td･idf 法のような表面上の統計に基く計算がされている。例として、(Frantzi & Ananiadou 96, 99) のコーパスから複合語の独立性をカウントする C-value,NC-value 法や(Hisamitsu & Niwa 99)の専門語とそれ以外の語の距離を図る方法などが挙げられる。

3.語基の連接情報を用いた用語抽出法

3.1 語基バイグラム

1 章で述べたとおり単名詞と単名詞で構成される複合語の関係は重要である。しかし、この関係は今まであまり注目されていない。(Nakagawa & Mori 98) らの手によってこれを利用したスコア付けの手法を、本稿によって包括的に拡張する。技術的な文書において、その分野を特徴付けるよう

な語は、たいてい名詞句や複合語である。その一方、数少ない単名詞が複合語を構成している。このことを基本概念とし、各々の単名詞の重要性を図る新しいスコア付けの手法を提案する。この手法は与えられた文書やコーパスに現れるある単名詞によって構成される複合語を区別することでスコア付けを行う。つまり図 1 の示すようにある単名詞 N とともに現れる語 M とのバイグラムをとるというものである。

[LN1 N] (#L1) [N RN1](#R1)
[LN2 N] (#L2) [N RN2](#R2)
: :
[LNn N](#Ln) [N RNm](#Rm)

図 1. 名詞 バイグラムとその頻度

ここで[Lni N](i=1...n)や[N RNj](j=1...m)は複合語の一部分となる名詞のバイグラムを表し、#Li(i=1...n)や#Rj(j=1...m)はそれぞれのバイグラムの頻度を表す。またバイグラムのみを扱うので文書中に実際現れるような複合名詞[Lni N RNj]は[Lni N]か[N RNj]のどちらかとして処理する。以下に“trigram”を含むような複合語を候補として抽出する例を示す。

例 1

トリグラム 統計、名詞 トリグラム、クラス トリグラム、名詞トリグラム、トリグラム 獲得、名詞 トリグラム 統計、文字 トリグラム

単名詞“トリグラム”から構成される名詞バイグラムは以下のように表される。

名詞 トリグラム(3) トリグラム 統計(2)
クラス トリグラム(1) トリグラム 獲得(1)
文字 トリグラム(1)

図 2. 名詞バイグラムの例

我々は専門語の大半が複合語であるということから、スコア付けをする関数を定義するために単名詞のバイグラムのみに焦点をおいた。

3.2 スコア付け関数

3.2.1 名詞バイグラムによる手法

前節で述べたようなスコア付けは無数の変化が考えられるので、ここで単純かつ、代表的な手法を示す。

#LDN(N) #RDN(N): これらはそれぞれ前方、後方に連接する単名詞 N の数を表している。例えば図 2 の例においては#LDN(トリグラム)=3,#RDN(トリグラム)=2 となる。

#LN(N,k) #RN(N,k): これらは図 1 で示されたような名詞バイグラムの頻度をカウントするものである。

$$\#LN(N,k) = \sum_{i=1}^{\#LDN(N)} (\#Li)^k \qquad (1)$$

$$\#RN(N,k) = \sum_{j=1}^{\#RDN(N)} (\#Rj)^k \qquad (2)$$

式 (1) (2)で示される k はパラメターである。K=1 と置いた場合、#LN(N,1) ,#RN(N,1)は前方、後方にそれぞれ連接する N の全ての頻度となる。
図 2 を例にとると#LN(トリグラム,1)=5,#RN(トリグラム,1)=3 となる。K を大きくとれば、名詞バイグラム の頻度が大きく反映され、逆に k=0 とするとその頻度は全く反映されず、N がコーパスで扱う分野のキーとなる概念を示すことになる。しかしながらこの場合は我々の専門語自動抽出の実験では良い結果は得られなかった。

3.2.2 エントロピー

名詞バイグラムの情報量を測る最も伝統的な手法としてエントロピーが挙げられる。前方に連接する N のエントロピーを $H_L(N)$,
後方に連接する N のエントロピーを $H_R(N)$と定義する。

$$M_L = \sum_{i=1}^{\#LDN(N)} \#Li, \quad M_R = \sum_{j=1}^{\#RDN(N)} \#Rj$$

$$H_L(N) = -\sum_{i=1}^{\#LDN(N)} \frac{\#Li}{M_L} \log \frac{\#Li}{M_L} \qquad (3)$$

$$H_R(N) = -\sum_{i=1}^{\#RDN(N)} \frac{\#Ri}{M_L} \log \frac{\#Ri}{M_L} \qquad (4)$$

3.2.3 Modified C-value

ベースラインの手法として強力で複合語に対して適切にスコア付けを行う C-value 法を使用する。しかしながらオリジナルの C-value 法では以下のような式のため単名詞を抽出できない。

$$C\text{-}value(a) = (length(a)-1)\left(n(a) - \frac{t(a)}{c(a)}\right) \qquad (5)$$

a は複合名詞であり、length(a)は a を構成する単名詞の数、n(a)はコーパス中に現れる a の頻度、t(a)はより長い複合語中に含まれる a の頻度、c(a)はその種類数を示している。式(5)を見て判るように a が単名詞の場合には、C-value の値は 0 となってし

まう。そこで単名詞に対してもスコア付けできるよう式(5)を次のように改良した。

$$MC\text{-}value(a) = length(a)(n(a) - \frac{t(a)}{c(a)}) \quad (6)$$

3.2.4 複合語のスコア

次に単名詞におけるスコア付け関数を複合語にも拡張しなければならない。我々は幾何平均(Geometric Mean)を使用する。まず最初に 3.2.1 節、3.2.2 節で述べた#LDN(N)、#LN(N,k)、$H_L(N)$ を記号 FL(N)で表し、#RDN(N)、#RN(N,k)、$H_R(N)$ を同様に記号 FR(N)で表すことにする。複合語を $CN=N_1N_2....N_L$ とすると CN に対する GM は以下のように表すことができる。

$$GM(CN)=\left(\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1)\right)^{1/2L} \quad (7)$$

GM は複合語の長さには依存していない。複合語とその長さとの関係についての重要性についてはまだ明確ではないので、長さに関係なく全ての複合語を公平に扱うことにする。一方、MC-value については MC-value 値をそのまま使用することにする。

3.2.5 独立した単名詞の頻度

3.2.1 節、3.2.2 節では独立して出現する単名詞の頻度の情報を利用しなかった。この情報を新たに付加すればより良い結果が期待できる。独立して出現した単名詞にはスコアに重み付けをすることにする。つまり、式(7)において L=1 の場合のみこの式を改良した GM1(N)を以下に定義する。

$$GM1(N)=f(N)\times GM(N) \quad (8)$$

ここで f(N)とは単名詞 N が独立して現れた頻度である。

4.実験的評価

4.1 実験

我々の実験には NTCIR1 TMREC のテストコレクションを使用した。TMREC は日本語の用語抽出のテストコレクションを提供している。このタスクの目的は、NACSIS Academic Conference Database から集められた 1870 の要約を含むコーパスと 8834 語の基本的な用語から専門語を自動的に抽出することである。各々のスコア付け関数を基にした GM(NC)と 3.3.3 節の MC-value 値をそれぞれの候補語に割り当てた。評価としてそれぞれの手法でランク付けされた上位 3000,6000,9000,12000,15000 語について再現率、適合率、F 値を求め比較した。

4.2 語基バイグラムに基く手法

まず 3.2.1 節の手法においてパラメーターk の効果を知るために、それぞれパラメーターk を 0,0.25,0.5,1,2,4 とした場合の 複合語のスコア付けを行った。結果を表 1 に示す。

PN	K					
	0	0.25	0.5	1	2	4
3000	1746 .295	1743 .294	1772 .299	1784 .301	1778 .300	1769 .299
9000	4713 .529	4694 .526	4704 .527	4744 .532	4714 .528	4711 .528
15000	7036 .590	7037 .590	7035 .590	7042 .591	7048 .591	7045 .591

表 1. 各パラメター値 k における正解語完全一致数と F 値

各セルの上部に正解語とマッチした候補語、下部に F 値が示されている。PN はそれぞれ抽出された候補語を上位順に並べたものである。この結果から k=1 の場合が最も良い結果を示し、特に低い PN でその傾向が現れている。しかし、k の値を変化させてもさほど大差はないように思われる。

4.3 結果の比較

この節では、前章で述べた各々の手法に対するスコア付けの結果を比較する。5 つの手法によるスコア付けの結果を表 2,3 に示す。表 2 はセルの上から順に再現率、適合率、F 値が示されており、表 3 は NTCIR1 TMREC のテストコレクションによって与えられた正解語とマッチした候補語数を示している。

PN	LDN RDN	LN(1) RN(1)	GM1	Entro -py	MC- value
3000	.197 .582 .295	.202 .595 .301	.202 .596 .302	.198 .585 .296	.239 .704 .356
6000	.371 .545 .441	.372 .548 .443	.372 .549 .444	.374 .551 .445	.415 .612 .496
9000	.533 .524 .529	.536 .527 .532	.537 .527 .532	.527 .518 .522	.557 .548 .553
12000	.676 .498 .573	.680 .501 .577	.681 .502 .578	.676 .498 .574	.684 .504 .580
15000	.796 .469 .590	.796 .469 .591	.801 .472 .594	.796 .469 .590	.799 .471 .593

表 2. 各手法における再現率、適合率、F 値

PN	LDN RDN	LN(1) RN(1)	GM1	Entro -py	MC- value
3000	1746	1784	1789	1755	2111
6000	3270	3286	3293	3306	3671
9000	4713	4744	4746	4661	4930
12000	5974	6009	6025	5979	6046
15000	7036	7042	7082	7035	7068

表 3. 各手法における正解語完全一致数

名詞バイグラムに関連する統計情報の中では、3.2.5節で述べた単名詞のみで現れる頻度の情報を付加したGM1がもっとも良い評価となった。

4.3 考察

表2,3から見て取れるように、語基バイグラムによる手法ではGM1を除いては大差のない結果となった。他と比べエントロピーの手法は劣っていた。これは語の重要性というのは、数学的な情報量からでは直接はかれるものではなく、むしろコーパス中の語の頻度から見て取れるものであるからである。

次にMC-value法と他を比較してみる。この手法は明らかに他の手法より優れており、特に低いPNで顕著である。しかしながら、高いPNではGM1がこれを上回っている。

ここで、抽出された複合語の長さの視点から結果を見てみよう。複合語に含まれる単名詞を1とカウントしそれぞれのPNでの平均語長を表4に示す。

PN	LDN RDN	LN(1) RN(1)	GM1	Entro -py	MC-value
1~3000	2.78	2.91	2.79	2.68	2.48
3001~6000	2.88	2.88	2.88	2.76	2.79
6001~9000	2.72	2.69	2.70	2.69	3.05
9001~12000	2.46	2.45	2.46	2.75	2.42
12000~15000	1.94	2.94	1.99	2.06	2.00

表 4. 各手法における候補語の平均語長

3000語までのPNにおいて、他と比べMC-value法によって抽出された候補語の平均語長は、明らかに短くなっておりそれと同時に、高い再現率、適合率を示す結果となった。この結果から3000語までの正解語は比較的短い語が多くMC-value法はその短い語に対して強力な結果を示した。しかしPN3000語以上でのMC-value法における平均語長は、他の手法と比べ短いというよりはむしろ長かった。それと平行して、正解語との完全一致数、再現率、適合率は他の手法と近くなっていき、最終的に上位15000語ではGM1の手法のほうがよりよい結果を示した。このような結果はそれぞれの手法を特徴付けるものでどちらが良い、悪いということはできない。専門語自動抽出の意図によって使い分けるのが望ましい。もし、数少ない語の中で高質の用語や、短い用語を求めたいならMC-value法を推奨するし、可能な限り包括的に用語のリストを求めたり、長い用語を求めたいのなら語基バイグラムによる手法がより望ましい手法となる。

最後にNTCIR1との結果と我々の結果を比較する。今回は3000語から15000語までしか実験していないため直接他の研究とは比較できないが、重要な点は語基バイグラムによる我々の手法によって抽出された候補語上位16000語のうち、7367語が正解語とマッチしたという点である。この結果は他の研究と比較しても高い結果といえるだろう。これまで候補語上位16000語のうち正解語とマッチした語数は6536語が最高であった。最も正解語とマッチした語数は7944語であるがこれは候補語上位23270語からマッチしたものであり、かなり低い適合率である。

5. おわりに

本稿では、複合語を構成する単名詞に連接する語の頻度を用いた語基バイグラムによる専門語自動抽出について紹介した。NTCIR1 TMRECのテストコレクションを通して実験的評価を行い、我々の提案する単名詞のエントロピーを用いた手法が専門語自動抽出に高いパフォーマンスを示すことが示された。低いPNで優れた結果を示したC-value法を改良したMC-value法は高いPNでは正解語を充分に拾うことはできなかった。今回我々が提案した手法は今後の課題に向けて様々な可能性を秘めている。

参考文献

Nakagawa, H., 2001, "Automatic Term Recognition based on Statistics of Compound Nouns", Terminology, to appear

大畑 博一 中川 裕志. 連接異なり語数による専門用語抽出. 情報処理学会研究報告 2000-NL-136 pp119-126

Kageura, K. and Umino, B. 1996. "Methods of automatic term recognition: a review". Terminology 3(2), 259-289.

Nakagawa, H. and Mori, T. 1998. "Nested Collocation and Compound Noun for Term Recognition". In Proceedings of the First Workshop on Computational Terminology COMPTERM'98, 64-70.

Nakagawa, H, 2000, "Experimental Evaluation of Ranking and Selection Methods in Term Extraction", in "Recent Advances in Computational Terminology" D. Bourigault eds, JohnBenjamin Publishers, in printing

Hisamitsu, T. and Niwa, Y. 1999. "Term Extraction Using a New Measure of Term Representativeness". In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 475-484

K.T. Frantzi, and Sophia Ananiadou, 1996, "Extracting nested collocations", 16th COLING, 41-46