社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－ＦＩ－ 80 (4)
2005－ＮＬ－169 (4)
2005／9／29

中文版「言選Web」の評価

前田 朗[1]　小島 浩之[2]　中川 裕志[3]

[1]東京大学理学部 [2]東京大学経済学部 [3]東京大学情報基盤センター

専門用語自動抽出システム「言選 Web」は、文章中から用語候補を抽出し、重要度のランク付けを行うシステムである。その基盤となる重要度のランク付け手法は、用語を構成する単名詞の相互の連接回数に着目したものである。我々は「言選 Web」のエンジンを中文に適用し、文字に着目した重要度計算や、ICTCLAS（中文の形態素解析器）の解析結果を元にした用語抽出、固有表現抽出手法を簡略化した用語抽出などを試みた。その結果、人民日報タグつきコーパスより、再現率の上限で 72.62%、平均適合率で 35.84%の評価結果を得ることができた。

Evaluation of Gensen-Web Chinese version

Akira Maeda[1]　Hiroyuki Kojima[2]　Hiroshi Nakagawa[3]

[1] Faculty of Science, University of Tokyo　[2] Faculty of Economics, University of Tokyo　[3] Information Technology Center, University of Tokyo

Gensen-Web is a domain term extraction system. Its main feature of calculating term weight is to employ the vocabulary level productivity which is based on how many compound terms composed of the target simple term. In this research, we apply Gensen-Web into Chinese term extraction. Several methods are experimentally evaluated, including character based weighting, word based weighting with Chinese Lexical Analysis System ICTCLAS and a simple named entity extraction. The result show that maximum recall of 72.62%, average precision of 35.84% for *ZINMIN-NIPPOU*(People's Daily News)articles.

1．はじめに

我々は、Web 上の文書から専門用語（キーワード）を抽出するサービス「言選 Web」(http://gensen.dl.itc.u-tokyo.ac.jp/)を立ち上げ、2003 年 4 月から公開した。「言選 Web」は、実用的にはメタデータベースのキーワード選定補助（東京大学経済学部 Engel）、個人のウェブログの解析、翻訳の補助など、さまざまな用途で使われている。

「言選 Web」は、次の２段階の処理からなる。

段階１：文章中から用語候補を切り出す。

段階２：用語候補の重要度を計算し、高い順に並べる。

文書を「言選 Web」で解析し、ランク上位となる用語を主要な専門用語、もしくはキーワードとみなす。

さて、「言選 Web」には、日本語版、中文[1]版、西欧言語版がある。中文について、我々は人民日報の Web 記事を対象に評価を進めてきたが、実験記事数が 30 と少ないことが難点であった [Nakagawa et al. 2004]。そこで、本稿では人民日報タグ付きコーパス(1998 年 1 月分, 全 3,055 記事)を用い、中文版「言選 Web」のより精密な評価を試みることとした。

以下、2 章では中文のテストコレクション、3 章では中文版の用語候補抽出方法、4 章では重要度計算法、5 章では実験と評価について述べる。6 章はまとめである。

[1] 中文とは中国語においては中国語で書かれたテキストを意味する。

2．人民日報のテストコレクション

2.1 人民日報タグつきコーパス

本稿では、評価のために人民日報タグつきコーパス(1998年1月分)を用いる。このコーパスは、文章が単語分割・品詞情報付与済みであるだけではなく、1部の複合語は[ ]で括られた上で品詞情報が付与されている。品詞内訳を表1に示す。組織名や地名など、新聞記事において専門用語（キーワード）と捉えられる用語である。本稿では、[ ]で括られた複合語を以後、**第1種正解語**と表記する。

表1　第1種正解語品詞内訳

品詞(タグ)	出現回数	割合
組織名 (nt)	7,381	84.5%
地名(ns)	1,015	11.6%
他の固有名(nz)	327	3.7%
成語(i)	7	0.1%未満
習用語(l)	3	0.1%未満
計	8,733	

*複合語の組織名や地名の全てについて、網羅的に[ ]タグが付与されているわけではない。

2.2 第2種正解語

第1種正解語は組織名と地名だけで96%を占める。しかし、組織名、地名以外でも記事において重要な用語は存在する。たとえば、ドーピングの記事であれば、その中国語に相当する「使用興奮剤」も重要語とする必要がある。第1種正解語は、この点で若干問題を抱えている。そこで、新たに文章中の内容を端的に示す語を中心に、人手によりコーパスから選定し、正解語として評価に用いる。先の第1種正解語に対し、これを**第2種正解語**と表記する。

第2種正解語は、第1種正解語に左右されないよう、コーパス内のタグ[ ]を取り払ってから選定した。この際、以下の点について留意した。

（1）原則として複合名詞（句）であること。（ただし内容によっては、単名詞や動詞を含む複合語を選定した場合もある。[2]）

（2）組織名、地名、人名は、記事のトピックとして差し支えない限りにおいて選定する。

（3）選定したキーワードが文章中で、省略語や同義語に言い換えられている場合は、これら言い換えられた語も選定する。

表2に第2種正解語の品詞内訳を示す。

表2　第2種正解語品詞内訳

品詞	出現回数	割合
組織名	392	14.9%
地名	304	11.6%
人名	246	9.3%
その他	1,286	64.2%
計	2628	

*対象記事数は229、1記事あたりの第2種正解語の割合は11.48語である。

3．用語候補抽出

段階1の用語候補抽出方法としては、ストップワードによる方法、形態素解析による方法、固有表現抽出を応用した方法の3種について検討した。

3.1 ストップワードによる用語候補抽出

ストップワードは、用語としては不適格な語である。文章中からストップワードを除去すれば、残った文字列は用語候補である可能性が高いと考えられる。

このストップワード方式による用語候補抽出法は、形態素解析ソフトに依存しない。また、中国語は孤立語であるためストップワードを適用しやすい。屈折語（英語やフランス語など）における単語の語尾変化の考慮は不要である。

中文向けのストップワードの選定は人手により行った（付録C参照）。その結果、ストップワード方式の性能向上には1～2字の語をいかに選定するかが重要であることが判明した。また、動詞に使われる漢字が他の漢字と結びついて名詞になりえるなど、妥当な解を得ることが難しいことも分かった。

3.2 形態素解析器ICTCLASによる用語抽出

段階1の用語候補抽出のもうひとつの方法として、形態素解析で得られる単語とその品詞情報によって、用語候補を抽出する方法を用いた。ここでは、形態素解析器（形態素の概念を利用して単語分割と品詞タグ付けを行うソフト）として、ICTCLAS（中国科学技術院の漢語詞法分析系統）を使用した。

[2] 例えば選挙結果のみを示す記事では"選挙"の語を選定せざるを得ない。またドーピング事件についての記事では"使用興奮剤事件"のように動詞を含んだ複合語を選定せざるを得ない。

用語候補としての形態素列の抽出には、中国語の文法書を参考にして決めた「単語の組が用語を生成するパターン」を用いる。このパターンに合致した単語の組を用語とする。詳細を次に記載する。括弧内はICTCLASの品詞タグである。

名詞に類する語(ng n nr ns nt nz nx vn an i j)　*以後「名詞」
　→　名詞,形容詞,助詞,後接成分,連詞（和,与）に結合。
　　複合語の先頭及び終端になる
形容詞(ag a)
　→　形容詞,助詞,後接成分,連詞（和,与）に結合。
　　複合語の先頭になる
助詞(u)、後接成分(k)
　→　名詞,形容詞に結合
連詞(c)
　→　和,与の場合のみ。名詞に結合。
区別詞(b)
　→　名詞, 助詞,連詞（和,与）に結合。複合語の先頭になる

3.3 固有表現抽出手法を用いた用語候補リスト

3.3.1　用語候補抽出の概要

段階1の用語候補抽出の精度をあげるために、第1種正解語の出現パターンを機械学習させ、それを元に第1種正解語に即した用語候補抽出法を試みた。

機械学習による用語候補の抽出には、決定木 C4.5による固有表現抽出手法[Sekine 1998]を簡略化して適用する。固有表現抽出の性能評価では、再現率と適合率の幾何平均であるF値が高いことが求められる。しかし、ここでの目的は用語候補の抽出であるから、適合率はさておくにしても再現率は高く保持する必要がある。

3.3.2　用語候補抽出

形態素が第1種正解語の開始や終了になりやすい品詞出現パターンを、人民日報半月分のデータ（1,483記事）を元に機械学習させる。

その教師データから導いたルールで人民日報の残り半月分のデータ（1,572記事）に正解語の開始と終了を判定させ、正しく開始と終了の係り受けになった形態素列を用語候補とする。

用語候補抽出の性能の評価は、次の式で示す適合率と再現率とし、適合率を犠牲にしない範囲での高再現率を目指した。

適合率 = 抽出した用語中の正解語数 / 用語抽出数
再現率 = 抽出した用語中の正解語数 / 総正解語数

教師データに使う素性は、判定対象の単語自身とその前後2単語の品詞情報とする。単語の判定は、固有表現抽出のstart/end法を簡略化して行う。start/end法のタグは以下のとおり [永田 2003]。なお、本研究では形態素を単語[3]、固有表現を正解語と読み替える。

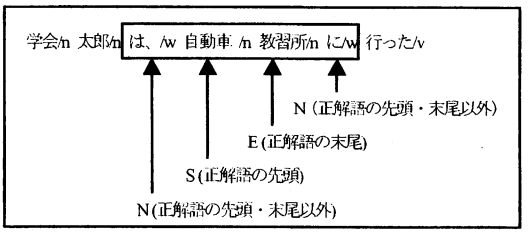
S　—　2語以上からなる固有表現の「先頭」の形態素
C　—　3語以上からなる固有表現の「中間」の形態素
E　—　2語以上からなる固有表現の「末尾」の形態素
U　—　1語で固有表現をなす形態素
O　—　固有表現の要素となる形態素以外

第1種正解語は複合語のみであるため、上記をS,C.E,Oとして適用する。しかし、決定木が判別するクラス数を減らすほうが、S、Eの検出に有利であるため、CとOを合わせたNを新たに考え、S（正解語の先頭）,E（正解語の末尾）,N（正解語の先頭・末尾以外）の3クラスとした。

人民日報タグつきコーパスは、記事に対して平均2.85語の第1種正解語が与えられている。後述の評価のために本研究では、正解となる第1正解語以上の用語数を抽出する必要がある。そこで、文章全体ではなく第1種正解語の前後のみを教師データとして選定し、より多くの第1種正解語候補を抽出できるようにする。

例えば、「学会/n 太郎/n は、/w 自動車 /n 教習所/n に/w 行った/v」との形態素解析済みの文を考える。/nが名詞、/w が助詞 /v　が動詞タグとする。「自動車教習所」が正解語としてあり、加えて「学会太郎」も正解候補として抽出したいとする。ここで、「学会」と「太郎」の出現パターンが正解語の先頭・末尾以外と学習させてしまうと、最悪の場合は「学会太郎」と「自動車教習所」の双方とも抽出できなくなる。

そこで、次の箇所に限り学習させた。



[3] 中国語の場合、形態素とは漢字一文字に相当するため

しかし、正解語とその前後だけでは教師データとして局所的過ぎ、決定木が不必要にS,Eに分類するケースが増える。つまり、最終的に不用な用語候補も多く抽出することになる。そこで、正例を第1種正解語とのその前後とし、それにランダムに選んだ一定数の負例を加える。この負例の数により、再現率と適合率を調整する。なお、負例は全てNとする。

次に生成したルールによる用語候補抽出について述べる。機械学習には決定木C4.5を用いた。C4.5で生成したルールをプログラムで実装し、評価用に残しておいた半月分の記事について正解語のS,Eのタグを付与した。

ここで、通常の固有表現抽出ではビタビアルゴリズムを使い、付与したタグの整合性をとる。本研究では、整合性をとるべきタグのパターンが単純なため、簡便さのため、以下のルールで不正なS,Eのパターンを排除した。

S…[S….E] → 後ろのS…Eを採用
(最短のStart, Endの係り受けのみを残す)
E…E…S、…E → 全て排除
(同一文節内かつS→Eの係り受けに限定する)

次にここで述べた用語候補抽出方法の性能を評価する。前述のとおり、正例に対する負例の数により用語候補の適合率と再現率を調整し、その中から適当なものを選ぶことができる。正例は19,334(内訳をクラスで示すと、S,Eがそれぞれ4,256、Nが10,842)である。その正例に負例を与えた結果を表3に示す。

表3 正例に対し負例を与える数による再現率と適合率の調整

負例を与える数	正解数	抽出数	再現率(%)	適合率(%)
0	2,066	28,266	78.76	7.30
10,000	1,962	12,601	74.80	15.57
20,000	**1,936**	**8,099**	**73.81**	**23.90**

本研究では、負例を20,000与えた機械学習結果、すなわち再現率73.81%,適合率23.90%の第1種正解語候補を用いた。

3.4 用語候補抽出法の評価

「言選Web」の段階1で抽出した用語候補を第1種正解語で評価した。抽出した用語候補に対し第1種正解語の占める割合がストップワード方式で、0.76%、ICTCLAS方式で2.31%と低い。これは、ストップワード方式とICTCLAS方式とも、組織名や地名以外の用語候補および単名詞も抽出したためと思われる。

次に本研究で独自に選定した第2種正解語を教師データとして、機械学習による用語抽出を試みた。しかし、ICTCLAS方式よりも再現率が劣ったため、本研究では評価対象としないことにした。

4 用語候補の重要度計算手法

段階1の結果、得られた用語候補の重要度計算を行う段階2の処理としては、用語を構成する単名詞の連接情報と、用語出現頻度を組み合わせたFLR[Nakagawa 2003]を採用した。単純な出現頻度ではなく、複合語の構造情報を用いるFLRは、Webの1ページ程度のテキスト量に対しても、効果を期待できる。

FLRは単名詞の連接情報による重要度*LR*に用語の出現頻度(Frequency)を掛け合わせた重要度である。

まず、*LR*について説明する。用語を構成する名詞としての最小要素を**単名詞**と呼ぶ。例えば、「情報処理学会」は*{情報 処理 学会}*と単名詞に分けることができる。ここで他の単名詞と連接することが多い単名詞ほど重要と考える。なぜならば、それは他の単名詞と連接して複合的な概念を生成する核心にある単名詞だからである。用語全体の重要度は、単名詞の重要度の相乗平均で示すことにする。

用語Wを単名詞w_iのリスト$W=\{w_1, \ldots, w_n\}$、
$L(w_i)$=単名詞w_iの左側連接回数+1,
$R(w_i)$= 単名詞w_i右側連接回数+1
とする。なお、+1しているのは、連接回数0の単名詞に対応するスムージングのためである。複合名詞$W=\{w_1, \ldots, w_n\}$の重要度LRは次式となる。

$$LR(W)=\left[\prod_{i=1}^{n} L(w_i)\times R(w_i)\right]^{1/2n}$$

複合名詞Wの出現頻度を$F(W)$とするとFLRは次式で定義される。

$$FLR = F(W) \times LR(W)$$

さて、漢字は表意文字であることが多いため、形態素解析による単語単位でのFLRの計算法(**単語FLR**)の他に、文字を単位としてFLRを計算する方法(**文字FLR**)も考えられる。例えば、前述の例「情報処理学会」を$W=\{情\ 報\ 処\ 理\ 学\ 会\}$と形態素から文字に置き換えて、FLRを計算することができる。

用語候補の重要度計算手法には、FLR 以外に、TF, Frequency, TF*IDF などもある。本研究では、これらと FLR を対比し、FLR の性質をみていくことにする。

ここで TF(Term Frequency)と Frequency は共に文章中の用語出現頻度であるが、用語の一部に他の用語が含まれている場合、TF はその含まれた用語もカウントし、Frequency はカウントしないという定義を用いた。一例として「情報システムと情報」の場合を次に示す。

TF (Term Frequency)
→ 「情報」2回、「情報システム」1回
Frequency
→ 「情報」1回、「情報システム」1回

TF*IDF は次式で算出している。

$$TF*IDF = TF \times \log(\frac{\text{総文書数}}{\text{該当の用語を含む文書数}} + 1)$$

5. 実験と評価

5.1 実験方法

以下に述べる A の 3 種の用語候補抽出法によって得た用語候補を FLR、Frequency, TF*IDF で重要度計算し順位つけたものを、以下の B の正解リストを用いて評価する。

A. 正解候補の用語
- 機械学習で抽出した第 1 種正解語候補
- ストップワード方式により抽出した用語
- ICTCLAS 方式により抽出した用語

B. 正解リスト
- 第 1 種正解語（コーパス付属の正解語）
- 第 2 種正解語（記事の内容を端的に表す語）

重要度計算の評価は、再現率(*Recall*)、適合率(*Precision*)、平均適合率(*Average Precision*)で行う。実験には人民日報半月分を 1 文書とした長文（人民日報半月単位）と、1 記事を 1 文書とした短文（記事単位）とで行う。

再現率と適合率は次の式で求める。

（1）正解に適合する用語数を $|D_q|$ とする

（2）用語リストの重要度ランク上位 k 語めが正解とマッチした場合に r_k =1,マッチしない場合 → r_k =0 とする。

第 k 位までを対象にした場合の再現率と適合率は次の式で求められる。

$$recall(k) = \frac{1}{|D_q|} \sum_{1 \le i \le k} r_i$$

$$presision(k) = \frac{1}{k} \sum_{1 \le i \le k} r_i$$

平均適合率 *AvePre* は次式になる。

$$AvePre = \frac{1}{|D_q|} \sum_{1 \le k \le N} r_k \times \quad presision(k)$$

＊ N は正解が最後に現れた順位

なお、記事単位での評価では、各記事の平均適合率を、対象記事全体に対して平均をとっている。

5.2 実験結果と考察

まず、人民日報記事半月分（1,572 記事）を 1 文書として扱い、第 1 種正解語候補に対し、各重要度計算手法の評価した結果を図 1 と表 4 に示す。

第一種正解語に対しては、文字 FLR と Frequency が良好な結果を示す。それに対し、単語 LR は、上位の語が正解と適合しておらず、平均適合率が劣る。

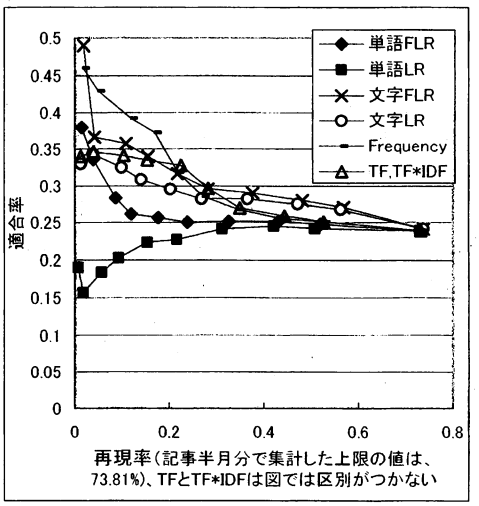


図 1　人民日報半月単位の第 1 種正解語に対する再現率 vs 適合率

表4　人民日報半月単位の第1種正解語に対する各手法評価結果

用語抽出方式	重要度計算方式	平均適合率(%)
機械学習による抽出(3.3.2参照)	単語FLR	19.39
	単語LR	17.04
	文字FLR	**22.58**
	文字LR	21.11
	Frequency	22.27
	TF	21.15
	TF*IDF	21.09

その原因として、単語LRの上位の用語が、人民日報半月分全体に対しての一般的な用語（「社会主義経済」「民族地区」など組織名・地名以外の用語）が多かったことが挙げられる。

逆に、Frequencyと文字FLRの好成績は、一般的な用語が上位に来ないためといえる。第1種正解語候補は元々第1種正解語に即して抽出を試みた用語であり、第2種正解語に対しての再現率上限も3.45%と低い。第1種正解語候補は、用語全般ではなく第1種正解語独自の傾向もつといえる。そこから考えると、Frequencyと文字LRは、第1種正解語候補において第1種正解語独自の傾向が強い用語に高い重要度を与えた可能性がある。

なお、TFとTF*IDFがほぼ同じ結果を示した。この理由として第1種正解語8,100語のうち、複数の記事に出現した用語が、718語と少なかったことが挙げられる。つまり、TF*IDFの計算式における「該当の用語を含む文書数」が多くの語で1になってしまい、TFとの差が現れなかったということである。

次に第1種正解語を正解として、各記事単位（対象記事は記事半月分と同じ1,572記事）で処理した場合の結果を図2と表5に示す。

第1種正解語の記事単位の結果では、記事半月分と同様に、文字FLRが優位性を示す。また、単語LRは記事半月分に対しては上位の用語が正解語と適合していなかったが、記事単位であれば上位の用語ほど多く適合する。これは、記事半月分ではなく、各記事を対象にしたことにより、記事単体についての重要な用語（正解となる組織名、地名を含む）を上位にランクしたためと思われる。

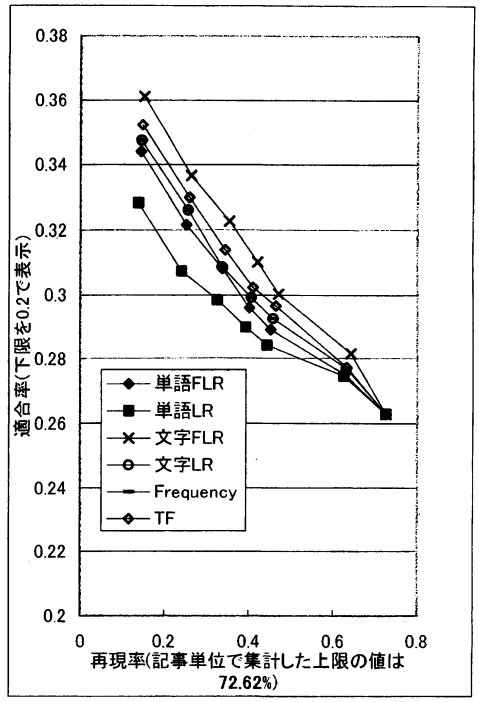


図2　記事単位の第1種正解語に対する再現率vs適合率

表5　記事単位の第1種正解語に対しての各手法評価結果

用語抽出方式	重要度計算方式	平均適合率(%)
機械学習による抽出(3.3.2参照)	単語FLR	34.16
	単語LR	33.61
	文字FLR	**35.84**
	文字LR	35.59
	Frequency	34.58
	TF	34.84

さて第2種正解語を正解とした場合の結果（図3、図4、図5、表6）について考察する。これは第2種正解語を付与した229記事が対象である。

全記事に対して、記事単位で集計した結果、FLRはFrequencyに対して、優位であった。さらに、第2種正解語中の組織名、地名のみ取り出し、第1種正解語に条件を近づけて評価した結果も単語FLRが優位性を示す。これは第1種正解語で文字FLRが有意性を示す結果とは異なる。第1種正解語と第2種正解語での実験結果の違いは、第1種正解語が組織名、地名が96%を占めていたことではなく、正解の選定基準が要因として考えられる。

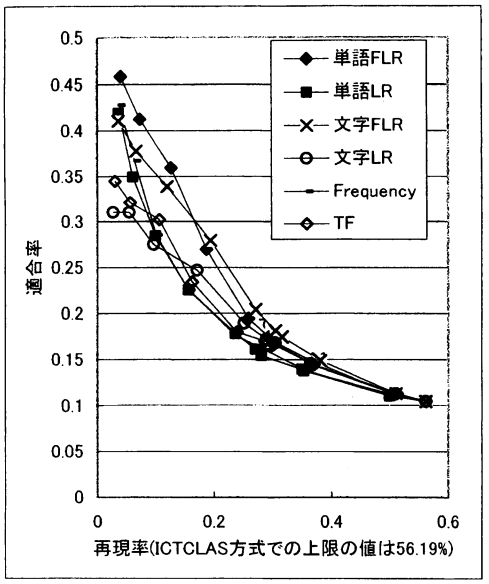


図3　記事単位の第2種正解語に対する
ICTCLAS 方式　再現率 vs 適合率（全正解語のみ）

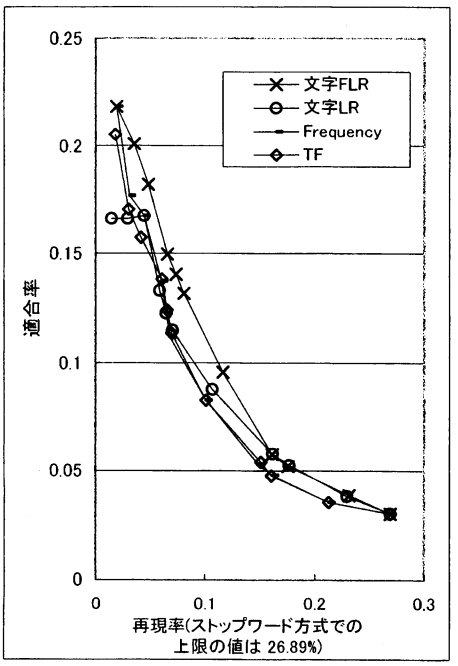


図4　記事単位の第2種正解語に対する
ストップワード方式　再現率 vs 適合率（全正解語のみ）

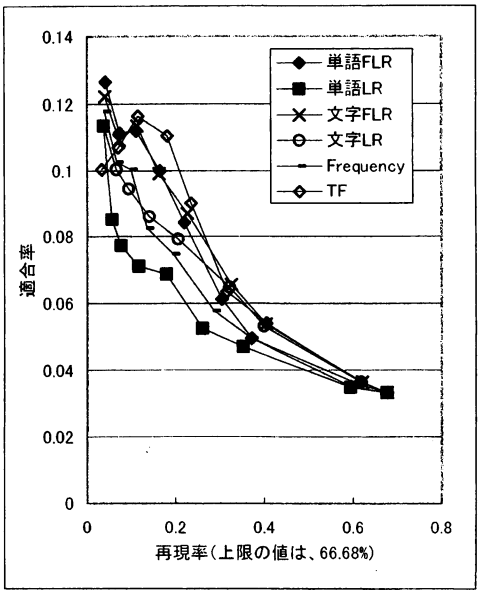


図5　記事単位の第2種正解語に対する
ICTCLAS 方式　再現率 vs 適合率（組織名・地名のみ）

表6　記事単位の第2種正解語に対する各手法評価結果

用語抽出方式	重要度計算方式	平均適合率(%)	
		全正解語	組織名、地名
ICTCLAS方式	単語FLR	**23.49**	**13.42**
	単語LR	21.26	11.62
	文字FLR	23.37	13.33
	文字LR	21.20	12.38
	Frequency	19.19	11.79
	TF	19.11	12.25
ストップワード方式	文字FLR	7.54	7.10
	文字LR	6.49	6.75
	Frequency	6.60	5.98
	TF	6.41	6.39

第2種正解語が全抽出用語に含まれる割合は、図3と図4の再現率上限に対応するが、ストップワード方式26.89%に対し、ICTCLAS 方式が 56.19%と高い。これは中文の用語抽出において、形態素解析を利用した手法が有効であることを示している。

しかし、中国語では動詞と動名詞的用法の区別がつかず、品詞依存の ICTCLAS 方式に比べ、こういった部分ではストップワード方式が威力を発揮することが検証されている[山崎 2005]。今後、ストップワードの

選定方法を再検討することで、再現率を上げることが可能かもしれない。

最後に、中文に特有の方法である、文字FLRの性能について考察する。

文字FLRは、文字LRをベースとし、Frequencyを掛け合わせた方式である。この文字LRは第1種正解語を対象に評価した場合、高い平均適合率を示した。これは単語LRと異なった結果である。この文字LRは、単語LRより連接対象数が多い。用語の文字列が長ければ「用語中に含まれる文字の出現頻度の相乗平均」に近似できる。つまり、文字の連接ではなく、頻出する漢字を含む用語を上位にしたという意味にも考えられる。

6 おわりに

本研究では、FLRによる用語の重要度ランク付けの効果が、正解とする語の選択により大きく左右されることを示した。

特に、文字FLRと単語FLRは、それぞれ有効に働くケースが異なることが判明した。今回は人民日報という新聞記事テキストを対象にしたが、別のジャンルのテキストで、文字FLR, 単語FLRを比較することが今後の重要な課題である。

付録A 人民日報タグつきコーパスのサンプル
（ドーピング関連記事から一部抜粋）

```
19980117-02-003-001/m 李/nr 铁映/nr 在/p
 全国/n 体委/j 主任/n 会议/n 上/f 强调/v
19980117-02-003-002/m 坚持/v 体育/n
 事业/n 发展/v 正确/a 方针/n 政策/n 同/p 使用/v
 兴奋剂/n 行为/n 进行/v 坚决/a 斗争/vn
19980117-02-003-003/m 新华社/nt 北京/ns
 1月/t 16日/t 电/n （/w 记者/n 许/nr
 基仁/nr 、/w 李/nr 贺普/nr ）/w
 [中共中央/nt 政治局/n]nt 委员/n 、/w
 国务委员/n 李/nr 铁映/nr 今天/t 在/p
 1998年/t 全国/n 体委/j 主任/n 会议/n
 上/f 说/v ，/w 体育/n 战线/n 要/v
 认真/ad 学习/v 、/w 全面/ad 贯彻/v
 党/n 的/u 十五大/j 精神/n ，/w 坚持/v
 体育/n 事业/n …
```

付録B 付録Aの記事に対するICTCLAS方式、単語FLR結果
（上位の語のみ抜粋）

```
李铁映(20.78)，体育事业(19.28)，体育战线(13.82)，
体育(9.49)，体育工作(8.11)，兴奋剂(7.94)，
全国体委主任会议(4.73)，群众体育(4.36)，
正确方针政策(4.16)，中国政府(4.12)，
竞技体育(4.05)，斗争(4.00)，
体育的必由之路(3.66)，体育健儿(3.66)，
中国体育的声誉(3.20)，整个体育战线(3.11)，
体育界(3.00)，…
```

付録C 中文版「言選Web」ストップワードリスト（一部抜粋）

```
之北 之后 之间 之南 之内 之前 之上 之外 之西
之下 之右 之中 之左 只好 只是 只要 只有
至多 至今 至少 至于 终于 逐步 逐渐 转眼
自从 自己 总共 总算 纵然 昨天 左面 阿嚏
嘣嘣 潺潺 哧溜 脆生 滴答 丁当 嘎巴 咯吱 咕咚
咕嘟 咕噜 呱呱 哈哈 哼哈 呼噜 哗啦 叽叽 嘎嘎
喳喳 本月 乒乓 扑通 比较 毕竟 必定 必然 嘻嘻
点儿 要是 一边 一面 也要 也不 别看 别说 何必
哎呀 我国 起来 来着 所谓 会得 方今 方得 啊
按 吧 把 被 比 彼 必 边 便 别 并 不 才 除 次
从 打 但 当 倒 到 得 等 点 顶 都 对 …
```

参考文献

Hiroshi Nakagawa, Hiroyuki Kojima, Akira Maeda, "Chinese Term Extraction from Web Pages Based on Compound word Productivity", 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004), Third SIGHAN Workshop on Chinese Language Processing, pp.79-85, Barcelona, Spain, July, (2004).

Nakagawa, H. and Tatsunori Mori. "Automatic term recognition based on statistics of compound words and their Components." Terminology, 9(2), pp.201-219 (2003)

S.Sekine, R.Grishman and H.Shinnou, "A Decision Tree Method for Finding and Classifying Names in Japanese Texts", 6th Workshop on Very Large Corpora, pp.148-152 (1998).

永田昌明: 「確率モデルによる自然言語解析」, 言語と心理の統計,岩波書店, 2003

山崎直樹「キーワード自動抽出システム『言選web』(中国語バージョン)を検証する」, 漢字文献情報処理研究6, 好文出版社, 2005.10 発行予定