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

2007－NL－182 (12)
2007／11／20

統計的特徴を利用した機能語の自動認定実験

木下 明徳† 後藤 功雄† 熊野 正† 加藤 直人† 田中 英輝†

†NHK 放送技術研究所
E-mail: †{kinoshita.a-ek, goto.i-es, kumano.t-eq, katou.n-ga, tanaka.h-ja}@nhk.or.jp

NHK の国際放送では 18 ヶ国語が使われており，それらの放送用原稿は，日本語の原稿やその英訳原稿が翻訳され作成されている．このような翻訳作業を支援するために，我々は過去の翻訳用例を検索する「多言語用例提示システム」の開発を行っている．精度の良い検索を実現するためには，検索キーワードとなりうる単語，すなわち，内容語の認定が重要である．しかしながら，内容語（あるいは機能語）を認定するには辞書が必要となるが，様々な言語に対して辞書を用意することは困難である．そこで，本稿では，言語が持つ統計的特徴を利用し辞書を使わない手法について述べる．また，８つの言語（日本語，英語，フランス語，スペイン語，ロシア語，イタリア語，インドネシア語，マレー語）に対して行った，機能語認定の実験結果について報告する．

Automatic Identification of Function Words by using statistic features common to many languages

Akinori Kinoshita† Isao Goto† Tadashi Kumano† Naoto Kato† Hideki Tanaka

†NHK Science & Technical Research Laboratories
E-mail: †{kinoshita.a-ek, goto.i-es, kumano.t-eq, katou.n-ga, tanaka.h-ja}@nhk.or.jp

NHK provides news services in 18 languages, translating Japanese news articles into English and those ones into other languages. To aid such translation work, we have been developing a translation example browser that retrieves examples similar to inputs from multi-lingual news corpora. The browser has to identify function words(or content words) in inputs by using machine-readable dictionaries to retrieve appropriate examples. However those dictionaries are difficult to be prepared for the browser in various languages. This paper proposes automatic identification methods of function words using statistic features common to many languages. We conduct a series of experiments in 8 languages, such as Japanese, English, French, Spanish, Russian, Italian, Indonesian language and Murray language.

1. はじめに

NHK の国際放送では 18 ヶ国語が使われており，それらの原稿は日本語の原稿やその英訳原稿から，それぞれの言語に翻訳されている．このような翻訳作業を支援するために，我々は過去の翻訳用例を検索する「多言語用例提示システム」を開発している[1],[2]．精度のよい検索を実現するためには，検索キーワードとなりうる単語の選択が重要である．選択すべき単語

としては専門用語も考えられるが(その抽出方法は、例えば[3][4])，最も基本となるのは内容語である．すなわち，検索キーワードの中で機能語をストップワードとし内容語のみで検索したほうが検索精度の向上が期待できる．しかしながら、内容語（あるいは機能語）を認定するには辞書が必要となるが，様々な言語に対して辞書を用意することは困難である．そこで，言語が持つ統計的特徴を利用し辞書を使わない方法で、内容語（実際にはその対である機能語）を自動認定する研究を行っている[5]。本稿では、さらに言語の種類を増やして実験を行ったので報告する．本手法により，抽出された機能語をストップワードとすることで、多言語用例提示システムでの柔軟な検索が可能となる．

2. 自動認定手法

提案手法の詳細は文献[5]に譲るが，機能語が持つ統計的特徴を以下の４つの評価関数 $f(w_i)$ で表し，その値が大きい単語を機能語と認定した．

2.1 出現頻度

機能語は出現頻度 $F(w_i)$ が高いと考えられる．そこで，

評価関数 $f(w_i)=F(w_i)$

$w_i \in W=\{\ w_1, w_2, \cdots, w_n\ \}$

：コーパス中の単語の集合

2.2 前後に隣接する単語の異なり数

機能語は，その前後に隣接する単語が様々なものであると考えられる．そこで，

$f(w_i)=|L(w_i)|+|R(w_i)|$

ここで，

$L(w_i)=\{l_1(w_i), l_2(w_i), \cdots, l_j(w_i), \cdots, l_n(w_i)\}$

：直前に出現する単語の集合

$R(w_i)=\{r_1(w_i), r_2(w_i), \cdots, r_j(w_i), \cdots, r_n(w_i)\}$

：直後に出現する単語の集合

2.3 エントロピー

2.2 では，隣接する単語の異なり数を直接用いたが，ここではこれをエントロピーで表現する．

$$f(w_i)=H(w_i)$$
$$=-\sum_{l_j(w_i)\in L(w_i)}\left(\frac{f_{l_j}(w_i)}{F(w_i)}\log_2\frac{f_{l_j}(w_i)}{F(w_i)}\right) \quad (1)$$
$$-\sum_{r_j(w_i)\in R(w_i)}\left(\frac{f_{r_j}(w_i)}{F(w_i)}\log_2\frac{f_{r_j}(w_i)}{F(w_i)}\right)$$

ここで、

$$f_{l_j}(w_i)=F\left(l_j(w_i)\right),\quad l_j(w_i)\in L(w_i)$$

$$f_{r_j}(w_i)=F\left(r_j(w_i)\right),\quad r_j(w_i)\in R(w_i)$$

2.4 エントロピーの再計算

2.3 の手法で機能語と認定した複数の単語を一つにまとめる(以下，「仮機能語」と呼ぶ)．仮機能語を１単語のように捉えると隣接する単語の種類が減少するため，一般にエントロピーの値も減少する．

ところが，機能語や内容語には，

- 機能語は，連続しにくい．
- 内容語は，隣に機能語を持ちやすい．

という特徴があるので，隣接する単語を駆り機能語でまとめると，

- 機能語は，隣接する単語の種類はあまり減少しないので，エントロピーの減少が小さい．
- 内容語は，隣接する単語の種類が大きく減少するので，エントロピーの減少が大きい．

と考えられる．そこで，以下のように，仮機能語を用いて，2.3 で求めたエントロピーの値の再計算を行う．

今，2.3 でエントロピーの値が上位 m の単語の集合を

$$K=\{k_1,k_2,\cdots,k_i,\cdots,k_m\}$$

とする．これが仮機能語となる．そして，隣接する単語の中でこれらの仮機能語を 1 単語として扱い，エントロピーを再計算する．すなわち，

$$\begin{aligned} f(w_i) &= H'(w_i) \\ &= -\sum_{l_j(w_i)\in L'(w_i)}\left(\frac{f_{l_j}(w_i)}{F(w_i)}\log_2\frac{f_{l_j}(w_i)}{F(w_i)}\right) \\ &\quad -\sum_{r_j(w_i)\in R'(w_i)}\left(\frac{f_{r_j}(w_i)}{F(w_i)}\log_2\frac{f_{r_j}(w_i)}{F(w_i)}\right) \\ &\quad -\frac{\sum_{l_j\in K} f_{l_j}(w_i)}{F(w_i)}\log_2\frac{\sum_{l_j\in K} f_{l_j}(w_i)}{F(w_i)} \\ &\quad -\frac{\sum_{r_j\in K} f_{r_j}(w_i)}{F(w_i)}\log_2\frac{\sum_{r_j\in K} f_{r_j}(w_i)}{F(w_i)} \end{aligned} \tag{4}$$

ここで、

$$L'(w_i)=L(w_i)-K \tag{2}$$

：直前に出現する単語から仮機能語を除いた集合

$$R'(w_i)=R(w_i)-K \tag{3}$$

：直後に出現する単語から仮機能語を除いた集合

これを数回繰り返す．

3. 実験結果と考察

3.1 実験対象

本稿の実験では，NHK のニュースコーパスの中から，日本語，英語，フランス語，スペイン語，ロシア語，インドネシア語，イタリア語，マレー語の 8 つの言語に対して、機能語の認定を行った．正解（ある単語が機能語であるか否か）の作成は、日本語，英語，フランス語，スペイン語，ロシア語，イタリア語に関しては，mecab，TreeTager[5]などの形態素解析結果を用い，インドネシア語，マレー語に関しては，GSK の辞書[6]を利用した．ただし，形態素解析器及び辞書によって未知語とされた単語は，すべて内容語とした．また，インドネシア語に関しては，省略形がうまく判定できず，明らかな機能語も未知語となってしまったが，それらすべてをチェックすることが困難であったため，今回は，内容語とした．

実験で用いた 8 言語のコーパスの特徴を表 1 に示す．表 1 からわかるように，今回，対象とした言語において，機能語の異なり語数は，100 語程度から 400 語程度，コーパス全体に占める機能語の割合は，10%強から 40%弱と言語によっては，3 倍以上の差があった．また，機能語の異なり語数とその割合も言語によって様々であり，ある程度性質の違う実験対象となっていると考えられる．

3.2 評価基準

評価に関しては，次の 4 つの基準で比較を行った．

（ｉ） 精度 P_{I} ：異なり語数による精度

$$P_{\mathrm{I}}=\frac{\text{実際に機能語であった　単語数}}{\text{機能語と認定した単語　数}}$$

（ii） 精度 P_{II} ：出現頻度を考慮した精度

$$P_{\mathrm{II}}=\frac{\text{実際に機能語であった単語の出現頻度の和}}{\text{機能語と認定した単語の出現頻度の和}}$$

（iii） 再現率 R ：出現頻度を考慮した再現率

$$R=\frac{\text{実際に機能語であった 単語の出現頻度の和}}{\text{全ての機能語の出現頻 度の和}}$$

	日本語	英語	フランス語	スペイン語	ロシア語	イタリア語	インドネシア語	マレー語
延べ語数	16,863,126	7,901,655	2,837,052	2,914,647	2,274,971	1,576,471	1,710,825	1,239,163
機能語の割合	38%	37%	35%	40%	21%	36%	11%	18%
機能語の異なり語数	360語	217 語	332語	105語	397語	164語	152語	114語

表１．各言語のコーパスの特徴

(iv)　F 値：(ii), (iii)による F 値

$$F=\frac{2P_{\mathrm{II}}R}{P_{\mathrm{II}}+R}$$

3.3　異なり語数による評価（P_{I}）

８言語に関して，各手法によって得られた上位の語を機能語と認定した．上位 50 単語と上位 200 単語に対する機能語の認定結果の精度 P_{I} を表２，表３に示す．ここで，表中のカッコ内に記された数字は，実際に機能語であった単語の数である．また，エントロピー再計算(2.4)のパラメータ m は，$m=100$ とし，再計算は，2 回行った．表２，表３において，灰色に塗りつぶされた部分の値が，４つの手法のうち各言語で最も良い値を示したものである．

表２，表３を見ると「出現頻度の手法」は，上位 50，200 単語ともに，どの言語でも一番結果が悪かった．また，「異なり数の手法」と「エントロピーの手法」を比較すると、上位 50 単語ではほとんどの言語であまり差が出なかったが，上位 200 単語では日本語やインドネシア語など一部の言語で「エントロピーの手法」の方がよかった．「エントロピーの手法」と「エントロピーの再計算の手法」を比較した場合は，上位 50 単語ではフランス語とスペイン語で，「エントロピーの手法」の方が良かったが，上位 200 単語では，全ての言語で「エントロピー再計算の手法」の方が良かった．しかし，表２，表３からもわかるように，上位 50 単語で「エントロピー再計算の手法」を用いた場合の精度 P_{I} が，日本語で 0.96，イタリア語で 0.86 と高い数値を示した以外は，あまり高い精度を得ることができなかった．また，表４より，上位 200 単語の時点での精度 P_{I} は，どの言語のどの手法においても 0.50 以下にとどまっている．

上位50単語の精度 P_{I}	出現頻度	異なり数	エントロピー	エントロピー再計算
日本語	0.46(23)	0.68(34)	0.84(42)	0.96(48)
英語	0.60(30)	0.68(34)	0.74(37)	0.74(37)
フランス語	0.44(22)	0.54(27)	0.46(23)	0.42(21)
スペイン語	0.48(24)	0.54(27)	0.58(29)	0.54(27)
ロシア語	0.46(23)	0.50(25)	0.44(22)	0.50(25)
イタリア語	0.58(29)	0.76(38)	0.84(42)	0.86(43)
インドネシア語	0.36(18)	0.50(25)	0.58(29)	0.62(31)
マレー語	0.14(14)	0.36(18)	0.36(18)	0.40(20)

※ ()内は語数

表２．上位 50 単語での精度 P_{I}

上位200単語の精度 P_{I}	出現頻度	異なり数	エントロピー	エントロピー再計算
日本語	0.20(39)	0.33(66)	0.41(82)	0.50(100)
英語	0.15(30)	0.40(80)	0.38(75)	0.41(82)
フランス語	0.11(22)	0.25(49)	0.22(44)	0.23(46)
スペイン語	0.12(24)	0.27(53)	0.31(61)	0.33(66)
ロシア語	0.20(39)	0.25(49)	0.28(55)	0.33(65)
イタリア語	0.30(60)	0.40(80)	0.39(78)	0.41(82)
インドネシア語	0.21(42)	0.25(49)	0.30(59)	0.33(65)
マレー語	0.12(24)	0.14(27)	0.15(30)	0.17(34)

※ ()内は語数

表３．上位 200 単語での精度 P_{I}

これはもちろん手法上の問題もあるが，言語によっては，今回用いたコーパス中に出現した機能語が 200 語にも満たない場合もあることが一因となっている．例えば，表１を見ると，スペイン語，イタリア語，インドネシア語，マレー語では，機能語の異なり語数が 200 語に満たない．これは，ある言語には機能語が何語あるかという問題でもあり，これを決めるのは難しく今後の課題である．

3.4　出現頻度を考慮した評価（P_{II}）

3.3 の評価では各機能語を同等に扱った（その出現頻度を考慮しない）が，実際に利用する

上位50単語の精度 P_{II}	出現頻度	異なり数	エントロピー	エントロピー再計算
日本語	0.78	0.84	0.96	0.98
英語	0.81	0.84	0.94	0.94
フランス語	0.75	0.78	0.82	0.79
スペイン語	0.85	0.88	0.92	0.92
ロシア語	0.71	0.77	0.82	0.86
イタリア語	0.77	0.84	0.88	0.89
インドネシア語	0.52	0.60	0.72	0.84
マレー語	0.43	0.51	0.55	0.63

表4．上位50単語での精度 P_{II}

上位50単語の再現率 R	出現頻度	異なり数	エントロピー	エントロピー再計算
日本語	0.94	0.93	0.80	0.89
英語	0.87	0.89	0.81	0.82
フランス語	0.93	0.95	0.86	0.72
スペイン語	0.96	0.96	0.92	0.87
ロシア語	0.87	0.87	0.73	0.76
イタリア語	0.87	0.91	0.88	0.82
インドネシア語	0.80	0.86	0.82	0.83
マレー語	0.88	0.93	0.84	0.87

表5．上位50単語での再現率 R

上位50単語のF値	出現頻度	異なり数	エントロピー	エントロピー再計算
日本語	0.83	0.88	0.88	0.93
英語	0.84	0.86	0.88	0.88
フランス語	0.83	0.86	0.84	0.75
スペイン語	0.90	0.92	0.92	0.89
ロシア語	0.78	0.82	0.77	0.81
イタリア語	0.82	0.87	0.88	0.86
インドネシア語	0.63	0.71	0.77	0.84
マレー語	0.58	0.66	0.66	0.73

表6．上位50単語での F 値

上位200単語の精度 P_{II}	出現頻度	異なり数	エントロピー	エントロピー再計算
日本語	0.61	0.66	0.80	0.78
英語	0.67	0.70	0.77	0.81
フランス語	0.60	0.63	0.69	0.71
スペイン語	0.68	0.73	0.82	0.85
ロシア語	0.53	0.57	0.65	0.69
イタリア語	0.63	0.70	0.78	0.80
インドネシア語	0.40	0.41	0.52	0.56
マレー語	0.31	0.35	0.40	0.43

表7．上位200単語での精度 P_{II}

場合を考えると，出現頻度が高い機能語が正しく認定されているほうがよいと考えられる．これは，出現頻度が低い機能語は、正しく認定されていなくとも，実際の利用場面であまり出現しないのであるから，その影響は小さいのではないかという考えに基づく．そこで，延べ語数から見た評価（出現頻度を考慮した評価）にも意味があると考え，評価実験を行った．

表4に上位50単語における各言語の精度 P_{II} を示す．また，異なり語数による評価(3.3節)では精度が上がれば再現率も上がるが，出現頻度を考慮した場合(3.4節)では，自動認定した機能語の異なり語数が同じであっても，その単語の違いによって頻度が変わってくるため，精度が上がっても，一概に再現率も上がるとは言えない．そこで，再現率 R による評価も行った．表5，表6に再現率 R，F 値をそれぞれ示す．

表4より，精度 P_{II} に関しては，フランス語とスペイン語は「エントロピーの手法」が最良の値となり，その他は，「エントロピー再計算の手法」が最良の値となった．フランス語とスペイン語に関しては，表2からわかるように，「エントロピー再計算の手法」によって得られた機能語の数が減ってしまったことによる影響によるものと考えられる．また，フランス語及びロシア語に関しては，「異なり数の手法」が最も多く機能語を獲得できていたが，出現頻

度を考慮した場合には，実際に機能語と判定された単語，間違って抽出した単語を問わず，出現頻度の高い単語が上位に出現しやすいため，間違った単語の出現頻度合計が大きくなり，「エントロピーの手法」より精度P_{II}が低い結果となった．また，マレー語に関しては，他の言語と比べて精度P_{II}の値が低くなってしまったが，これは，もともと獲得できた機能語の数が少なかった上に，コーパス中に出現する機能語の割合が最も少ない言語であったことが原因と考えられる．

表5を見ると，再現率Rは，頻度の高い機能語が上位となりやすい「異なり数の手法」が最もよく，次に「出現頻度の手法」がよかった．また，上位50単語の再現率Rは，各言語ともに最良の手法で0.85以上であった．よって，どの言語においても，上位の機能語だけで，実際に出現する機能語の大部分をカバーできることがわかる．

表6を見ると，F値で評価した場合には，各言語とも手法による精度の差があまりなかった．これは，精度P_{II}が「エントロピーの手法」や「エントロピー再計算の手法」で高い一方，再現率Rが「出現頻度の手法」や「異なり数の手法」で高いためである．ただし，「出現頻度を用いた手法」が最良の値となる言語は存在せず，どの言語においても何らかの形で前後に隣接する単語を考慮することは，有効であると言える．

表7に上位200単語の精度P_{II}を示す．上位200単語の再現率Rに関しては，各言語で0.90程度から0.98程度の高い値を示しており，手法ごとに見た場合には大きな差がないため，紙面の制約もあり割愛した．表7より，上位200単語では，日本語を除いたすべての言語で，「エントロピーの再計算の手法」が最良の結果となった．「エントロピーの再計算の手法」は，異なり語数による評価結果でも多くの言語でよい結果が得られており，ある程度有効な手法であると考えられる．

4. まとめと今後の課題

統計的特徴を利用することにより，機能語の自動認定手法を8つの言語（日本語，英語，フランス語，スペイン語，ロシア語，インドネシア語，イタリア語，マレー語）に適用し，その実験結果について述べた．今回の実験により，機能語を自動認定する上で，「エントロピー再計算の手法」が，多くの言語で有効であることがわかった．また，それぞれの手法で上位200単語における再現率Rが0.9以上であることや，自動認定された機能語の集合に多少の差があることを考慮すれば，それぞれの手法で得られた上位200単語程度の和集合を作成することで，かなりの再現率Rが得られることが期待できる．さらに，その和集合をその言語に精通している人にチェックしてもらうことで，多言語用例提示システムのストップワードを設定できると考えられる．

今後は、本手法で得られた機能語をストップワードとして多言語用例提示システムに適用していくとともに，多言語における専門用語の自動獲得についても，統計的特徴を利用した手法を検討していきたい．

参考文献

[1] I.Goto, N.Kato, N.Uratani, T.Ehara, T.Kumano, H,Tanaka, “A multi-language translation example browser,” In Proceedings of the MT Summit IX, pp. 463-466, 2003.

[2] 熊野，西脇，田中，“「翻訳パレット」を用いた翻訳支援の提案，” 言語処理学会第12回年次大会発表論文集，pp.701-704,

[3] K.Frantzi, S.Ananiadou, H.Mima, “Automatic Recognition of Multi-Word Terms : the C-value/NC-value Method,” International Journal on Digital Libraries, 2000.

[4] 山本，池野，浜口，井佐原，“検索支援に向けたWeb文章集合からの用語獲得，” 情報処理学会研究報告 2004-NL-164, pp.171-176, 2004.

[5] 木下，後藤，熊野，加藤，田中，“統計的特徴を利用した内容語の自動認定実験，” 言語処理学会第13回年次大会発表論文集

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

[7] CICC インドネシア語基本語辞書，CICC マレーシア語基本辞書 所有者 田中英輝