統計的特徴を利用した内容語の自動認定実験

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 〒157-8510 東京都世田谷区砧 1－10－11
E-mail: †{kinoshita.a-ek,goto.i-es,kumano.t-eq,katou.n-ga,tanaka.h-ja}@nhk.or.jp

1. はじめに

NHK の国際放送では 22 ヶ国語が使われており，その翻訳作業を支援するために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]. 精度よく検索するためには，検索キーワードとなりうる用語の獲得が重要であり，専門用語の抽出などの研究も行われている[3],[4]. これらの検索キーワードは，内容語が基になっているが，その内容語を認定するには，辞書が必要となる．しかし，様々な言語に対して辞書を用意することは困難であるため，本稿では言語が持つ統計的特徴を利用することにより，辞書を使わずに内容語と機能語を自動分離する実験を行ったので，報告する．

2. 自動認定手法

今回，内容語と機能語を自動分離する実験を行う上で，内容語に比べて単語数に限りのある機能語に着目して評価を行うこととした．

2.1 出現頻度

機能語の統計的特徴としてまず考えられるのは，出現頻度が高いということである．そこで，コーパス内に出現する単語の集合を

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

として，単語 $w_i \in W$ に対して，それぞれに出現頻度を求め，その出現頻度 $F(w_i)$ の高い単語を機能語とする．（表 1．出現頻度の上位 20 単語）

	単語	出現頻度
1	の	1071301
2	を	586049
3	に	580025
4	た	536030
5	が	509366
6	て	507896
7	は	485383
8	で	402383
9	し	310266
10	と	270596
11	い	213690
12	まし	198013
13	ます	171482
14	十	151956
15	こと	135312
16	れ	121730
17	から	115639
18	する	103058
19	二	102990
20	など	97470

表 1 出現頻度の上位 20 単語

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，コーパス内に出現するある単語 $w_i \in W$ に対して，その直前に出現する単語の集合を

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

直後に出現する単語の集合を

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

とする．このとき，それぞれの集合の要素数の和 $|L(w_i)|+|R(w_i)|$ を求め，その要素数の和の値が大きい単語を機能語とする．

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．単語 $w_i \in W$ の直前に出現する単語 $l_j(w_i) \in L(w_i)$ が，単語 w_i の直前に出現した回数を $f_{l_j}(w_i)$ とする. 同様に, $r_j(w_i) \in R(w_i)$ が単語 w_i の直後に出現した回数を $f_{r_j}(w_i)$ とする．このとき，以下の式(1)より，エントロピーの和 $H(w_i)$ を求め，値の大きい単語を機能語とする．

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

（ただし，$F(w_i)$ ：単語 w_i の出現頻度）

2.4　エントロピーの再計算

一般的に内容語の多くは，前後に複数の異なる機能語を持ちやすいことが考えられる．そこで，その性質を用いることで，内容語のエントロピーの和の値を減少させて，相対的に機能語のエントロピーの和の値を大きくすることを考える．2.3 では，エントロピーの和の値が大きい単語を機能語とした．よって，内容語は，これらの単語の複数を前後に持ちやすいことが仮定できる．そこで，2.3 で機能語とした複数の単語を、一つのまとまりとして捉えることで，内容語のエントロピーの和の値を減少させる．すなわち，エントロピーの和の値が上位 m の単語の集合を

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

として, ある単語 w_i の集合 $L(w_i)$，$R(w_i)$ に関して，新たな集合 $L'(w_i)$，$R'(w_i)$ を以下のように，定義する．

$$L'(w_i) = L(w_i) - K \quad (2)$$

$$R'(w_i) = R(w_i) - K \quad (3)$$

この集合を用いて新たなエントロピーの和 $H'(w_i)$ を式(4)より求める．これを数回繰り返して再計算をし，$H'(w_i)$ の値が大きい単語を機能語とする．

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

2.5　頻度重み付きエントロピー再計算

頻度による結果(2.1)とエントロピー再計算による結果(2.4)を比較したところ，重複しない機能語が存在した．そこで，両方の特徴をもつ指標として，ある単語 $w_i \in W$ の出現頻度を $F(w_i)$ としたとき，

$$\log_{10} F(w_i) \times H'(w_i)$$

の値を求め，その値の大きい単語を機能語とする．

3.　実験結果と考察

前章で述べた自動認定手法を用いて, NHK 日本語ニュース，英語ニュースを対象として，次の４つの評価基準で比較実験を行った．

（ｉ）　精度 P_{I} ：異なり語数による精度

$$P_{\mathrm{I}} = \frac{\text{機能語と認定した単語数}}{\text{異なり単語数}}$$

（ⅱ）　精度 P_{II} ：出現頻度を考慮した精度

$$P_{\mathrm{II}} = \frac{\text{機能語と認定した単語の出現頻度の和}}{\text{総単語数}}$$

（ⅲ）　再現率 R ：出現頻度を考慮した再現率

$$R = \frac{\text{機能語と認定した単語 の出現頻度の和}}{\text{全ての機能語の出現頻 度の和}}$$

（ⅳ）　F 値：（ⅱ），（ⅲ）による F 値

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

3.1　日本語ニュース

NHK 日本語ニュース（約 43 万文）に対して，機能語の自動認定実験を行った．ただし，日本語では，機能語は，助詞，助動詞とした．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 回行っている．結果を図 1-1，1-2，1-3 に示す．

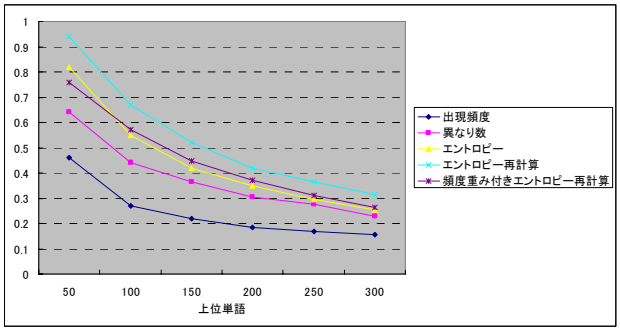

図 1-1　異なり語数による精度 P_{I}

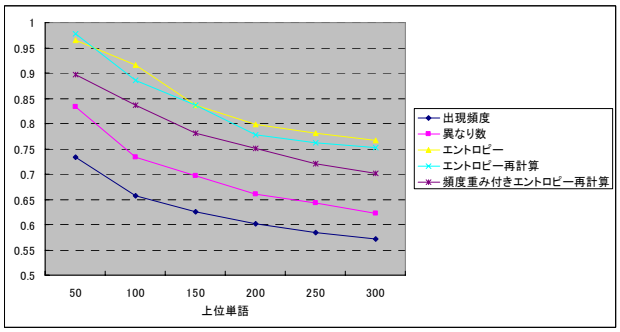

図 1-2　出現頻度を考慮した精度 P_{II}

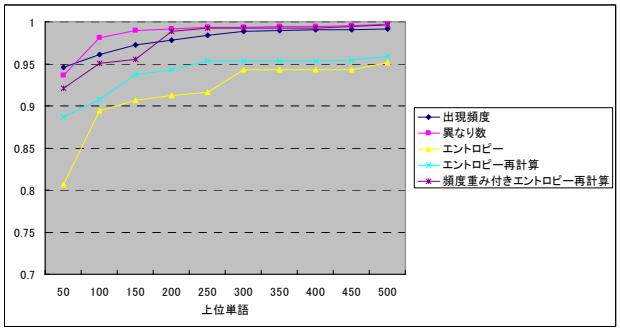

図 1-3 出現頻度を考慮した再現率 R

出現頻度，前後に隣接する単語の異なり数，エントロピーの手法を比べた場合，図 1-1，1-2 から明らかなように，エントロピーの手法を用いたときに最良の精度が得られ，出現頻度を考慮した精度 P_{II} が，上位 50 単語で，0.96 であった．また，今回用いたコーパスにおいて機能語と判定された単語の異なり総数は 255 単語であったが，上位 250 単語においても，精度 P_{II} が 0.78 であった．ただし，エントロピーの手法で上位となった機能語は，必ずしも出現頻度の高い順に並んでいるわけではなく，頻度が低くても上位になった単語，頻度が高くても下位になった単語が存在した．そのため，図 1-3 からもわかるように，出現頻度を考慮した再現率 R で上位 50 単語を比較した場合，エントロピーの手法で 0.80，出現頻度の手法で 0.94 の再現率となり，出現頻度の手法を用いた場合の方が良い結果となった．また，異なり語数による精度 P_{I} の場合では，3 手法のうち最良のエントロピーの手法でも，上位 100 単語で，実際に機能語であったものは 55 単語であり，十分に抜き出せているとは言い難い結果となった．

そこで，前章で述べたエントロピーの再計算(2.4)，頻度重み付きエントロピー再計算(2.5)の手法による実験も行った．その結果，エントロピー再計算の手法では，出現頻度を考慮した精度 P_{II} に関しては大きな変化がみられなかったが，図 1-1，1-3 からわかるように，異なり語数による精度 P_{I} と出現頻度を考慮した再現率 R では，一定の値の上昇が得られた．また，頻度重み付きエントロピー再計算の手法では，エントロピーの手法やエントロピーの再計算の手法の場合より，再現率を上げることができたが，精度に関しては，エントロピーの手法を下回る結果となった．

3.2　英語ニュース

NHK 英語ニュース（約 44 万文）に対しても同様の実験を行った．ただし，英語では機能語は，冠詞，前置詞，be 動詞とした．結果を図 2-1，2-2，2-3 に示す．

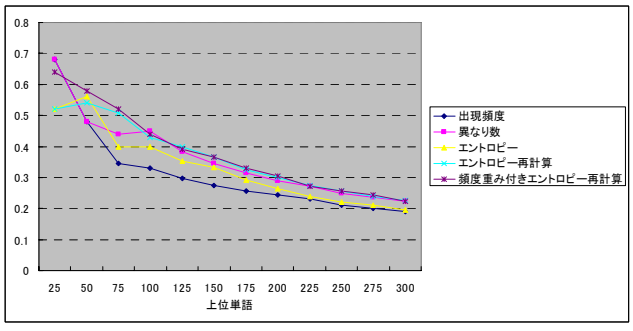

図 2-1 異なり語数による精度 P_{I}

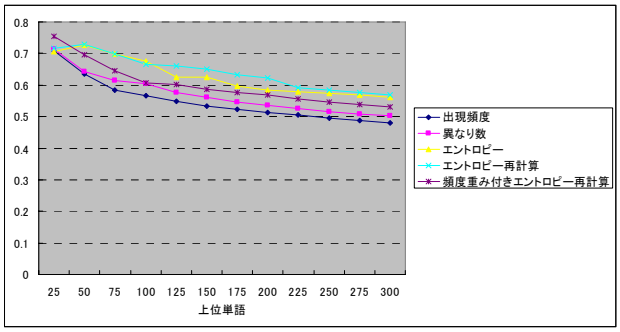

図 2-2 出現頻度を考慮した精度 P_{II}

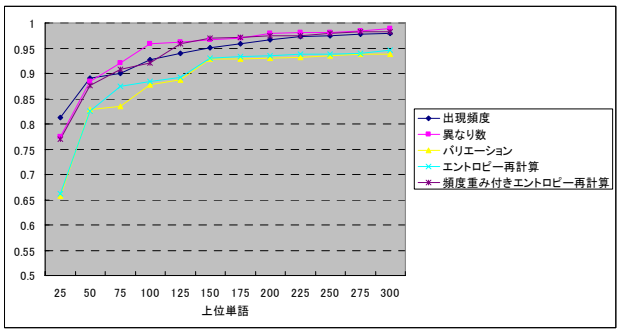

図 2-3 出現頻度を考慮した再現率 R

英語においては，出現頻度を考慮した精度 P_{II} の場合，エントロピーの手法が平均的に良い結果が得られ，上位 50 語に対して，0.72 であった．また，日本語同様，エントロピー再計算を用いることで，異なり語数による精度 P_I も，一定の値の上昇が得られた．しかし，図 2-1 からもわかるように，異なり語数による精度 P_I の場合，日本語の図 1-1 のようにエントロピー再計算の手法が最も良いという結果にはならず，手法ごとの優位性が明確には出ない結果となった．出現頻度を考慮した再現率 R に関しては，日本語と同様の傾向となり，出現頻度の手法を用いた場合に，上位 75 単語で，0.90 であった．

3.3 F 値による手法の比較

表 2，3 に，日本語，英語ニュースそれぞれに対して求めた F 値の値を示す．

	F値			
手法	上位50	上位100	上位150	上位250
出現頻度	0.826	0.781	0.761	0.733
異なり数	0.881	0.839	0.818	0.781
エントロピー	0.878	0.905	0.87	0.851
エントロピー再計算	0.93	0.897	0.884	0.847
頻度重み付きエントロピー再計算	0.909	0.889	0.86	0.847

表 2 日本語ニュースの F 値

	F値			
手法	上位25	上位50	上位75	上位150
出現頻度	0.757	0.741	0.71	0.683
異なり数	0.743	0.744	0.737	0.71
エントロピー	0.681	0.774	0.761	0.746
エントロピー再計算	0.688	0.774	0.776	0.766
重み付きエントロピー再計算	0.762	0.775	0.754	0.731

表 3 英語ニュースの F 値

この結果，日本語においては，エントロピーの再計算の手法が上位 50 単語において，最良の値 0.93 を得た．このときの精度 P_{II} が 0.98，再現率 R が 0.89 であることから，上位 50 単語を全て取り除けば，かなりの精度で，9 割近い機能語を取り除けることになる．また，重み付きエントロピー再計算の手法で，上位 50 単語の F 値が 0.909 であったが，このときの精度は 0.894，再現率は 0.925 であった．英語においては、どの手法を用いた場合も，最良の値で，0.74～0.78 であった．

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，内容語（実際は機能語）を自動認定した実験結果について述べた．日本語においては一定の分離が可能であり，内容語のキーワード認定などの前処理として利用することが期待できる結果となった．しかし，いずれの手法においても，多少の差はあるものの，接続詞や副詞などの単語が機能語と判定されてしまい，課題が残った．また，英語においては，日本語ほどの精度を得ることはできず，多言語へ展開する上では，課題が多い結果となった．

今後，より一層の機能語と内容語の分離手法を検討するとともに，他の言語でどのような結果が得られるか実験を行っていきたい．

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.