ＮＨＫニュース記事からのキーワードの抽出と記事の分野判別実験

浦谷 則好　　畑田 のぶ子
NHK 放送技術研究所
{uratani,hatada}@strl.nhk.or.jp

1．はじめに

文書を検索する場合、分野やキーワードで指示することが通常行なわれている。キーワード抽出や文書分類に関して研究が種々なされてきた。しかし、文書を代表するキーワードをどのように選んだら良いのか、そのキーワードを用いて文書がどの分野に判別できるのかについて、効果的な基準や手法がいまだに存在しない。

ＮＨＫではニュース記事を、「政治」、「経済」、「社会」、「国際」、「スポーツ」、「各局」、「首都圏」の７つに分けて管理している。そこで、記事中からキーワードを選び出し、それらのキーワードが分野分類にどの程度効力があるのかを調べたので、それについて報告する。

2．キーワードの抽出

通常、キーワードは対象とする文章を形態素解析して、自立語の中から選ばれる。しかし、その場合、形態素解析の精度が問題となり、単語の長さを短単位にとるか長単位にするかは形態素解析プログラムに依存することになる。しかも、自立語、付属語の区別は必ずしも明確ではない。そこで、形態素解析を行なわず、字面処理でキーワードを選定することにした。

われわれは、以前字面から表現パターンを抽出するために３つの基準を比較している[1)]。その結果、翻訳ユニット（機械翻訳のための表現パターン）の抽出のためにはエントロピー基準が一番適していることが判明した。（このエントロピー基準は下畑らのエントロピー基準[2)]とは違って文字列全体の分布のみから計算されるので計算コストは著しく小さい。）

しかし、エントロピー基準は長いパターンが優先される傾向があるので頻度が小さくなり、できるだけ少数のキーワードで分野分類したいという観点から見て、キーワード抽出にはあまり適していないように思われる。仕事量基準は短いパターンが優先される傾向があり、分野に特有の複合的な語や言い回しを捉まえられない恐れが強い。そこで、今回は平均情報量基準を用いてキーワードの抽出を行なった。（手法の詳細は文献１を参照）対象とした文書は９６年９月のＮＨＫニュース記事１カ月分で、約９０万文字（記事数4,038、文数21,078）である。その結果、頻度５以上で40,175個の文字列を抽出した。

●仕事量基準
$n \times f_\alpha$
●平均情報量基準
$\log (N / f_\alpha) / I_n$
●エントロピー基準
$\log (N / f_\alpha) / H_n$
ここで、
α：長さｎの文字列（例えば defgh）
f_α：文字列αの頻度
A_n：長さｎの文字列全体の集合
N：A_nに含まれる文字列の頻度の総和
f_n：A_nに含まれる文字列の頻度の平均
I_n：A_n中で平均頻度を持つ文字列の情報量
$= -\log (f_n / N)$
H_n：Ａｎのエントロピー（$= -\Sigma p \log p$）

3．分類のためのキーワードの選定

２．で抽出した文字列から分野分類に有効と思われるキーワードを以下のようにして選定した。

どの分野にも同様に出現するものであれば、その文字列には分野を特定する能力がないと断定できる。逆にもしある文字列が特定の分野にだけ出

現するならば、その文字列を含んでいるだけで分野を特定できる。そこで、分野毎にどのような出現の偏りがあるかを調べるため、各々の文字列の分野毎の出現頻度を調べた。つまり、

$p = M / N$　　M：ある分野での出現頻度

N：全分野での出現頻度の総和

としたとき、pが大きいほど分野分類のキーワードとして適切だということができる。しかし、pが大きくてもその出現頻度が小さければ実際上はあまり分野分類に寄与するとは思われない。また、pが0.5のもの2つが出現するよりpが1.0のもの1つ出現する方が明らかに分野を確信を持って限定できると考えられるので、分別力がpに比例すると考えるのは不適切だと思われる。そこで、

$(1.0 + p \log_2 p + (1.0 - p) \log_2 (1.0 - p))$

をキーワードの信頼度（Q）とし、M×Qが大きいものを分類のためのキーワードとして採用することにした。（ただし、p<0.5のときは0とする）M×Qの大きいものを表1に上げる。

文字列は特定の1分野にだけ偏っていると言えなくても、特定の2分野を取れば明らかに偏った出現をしている可能性がある。こうした文字列は単独では分別能力が低いと考えられるが、他のキーワードと一緒に現れるならば分別に寄与することが想像できる。そこで、2分野を取ったときにM×Qの大きなものも調べた。この上位を表2に上げる。

表1　M×Qの大きい文字列（上位15）

文字列	M×Qの値	分　野
選手	668.28	スポーツ
国際	583.49	国際
衆議院	455.04	政治
さきがけ	444.18	政治
気象庁	436.03	社会
橋本総理大臣	303.90	政治
伊豆諸島	277.59	社会
民主党	253.34	政治
リーグ	251.21	スポーツ
社民党	243.69	政治
暴風域に	242.90	社会
イラク	242.63	国際
関東地方	234.75	社会
衆議院選挙	228.59	政治
臨時国会	222.74	政治

表2　2分野でM×Qの大きい文字列

文字列	M×Qの値	分　野
地方	518.62	社会、各局
ン大統領	411.32	国際、政治
大田知事	365.77	政治、各局
千葉県	306.23	社会、首都圏
茨城県	277.42	首都圏、社会
沖縄本島	269.60	社会、各局
大阪	258.68	各局、社会
警察	258.09	各局、首都圏
エイズ	257.31	社会、各局
」と述べ	245.85	政治、国際

M×Q（2分野から求めたものは1／3として）の大きさの順に文字列をソートした。まず、以下の方法でキーワード候補を絞った。

（a）文字列をαとしたとき、それより下位にその一部となるような文字列β（例えば$\alpha = \beta\gamma$）があればそれを除く。

（b）文字列をαとしたとき、それより下位にαを含むものδ（例えば$\delta = \alpha\gamma$）があればそれを除く。

（a）の判断はβより判別に有効なより長いキーワードが存在することを意味しているからである。（b）はδが出現するならαが必ず出現するからである。

次に、M×Qの大きい順にキーワードとして次式を満足する個数まで採用することとした。

$$c S_i \leqq \Sigma (MQ_i)^2$$

S_i：分野iの記事数の二乗

MQ_i：分野iに対するM×Qの値

cは実験では1.21とした。ただし、分野毎に最低

３０個は採用し、最大は２５０個に抑えた。結果として、１００７個のキーワードが抽出された。分野毎のキーワードの個数は表３に示す。

表３　採用された分野別キーワード数

政治	経済	社会	国際	スポーツ	各局	首都圏
30	213	250	30	30	248	206

４．ニュース記事の分類

３．で選定したキーワードを用いて、キーワード選定の良否と分野判別能力を調べるために、ニュース記事の分類実験を実施した。キーワードを用いた文書分類は種々の方法が提案されている[3),4)]が、ここではｔｆ法、ｔｆ＊ｉｄｆ法[5)]、ｔｆ＊ｒｌｂ法（提案する手法）、Decision Tree法を採用した。ベクトル法である前３者は、各分野におけるキーワードの出現個数の期待値を用いて、分野の代表ベクターを求めておき、それと各記事のベクターとの内積を求めてその最大となる分野に判別するという方法をとった。（正規化は行なっていない。）

ｔｆ＊ｒｌｂ法は

$$idf = log(N/df)$$

の替わりに３．で述べた信頼度Ｑでｔｆを重み付けする方法である。Decision Tree法ではｐ（分野を特定したときのキーワードの出現確率で３．のｐと同じ）の大きい順に決定木を構成するということにした。

学習サンプル（９６年９月のニュース記事）に対するｔｆ＊ｉｄｆ法の適合率（Ｐ）、再現率（Ｒ）、Ｆ値の値を表４に示す。ただし、

$$F = 2PR/(P+R)$$

である。記事中に１つもキーワードを含まないもの（未分類となるもの）が54記事（政治：３，経済：２，社会：19，国際：13，スポーツ：６，各局：７，首都圏：４）存在した。これらは全て誤分類と見なした。表３を見ると、一番多くキーワードが採用されているにも関らず「社会」に未分類が多いことが分かる。これは、「社会」で取り上げるトピックスが幅広いことを示している。

表４．　ｔｆ＊ｉｄｆ法の分野別の判定精度（学習サンプル対象）

分野	記事数	適合率	再現率	Ｆ値
政治	652	.8137	.8773	.8443
経済	392	.7955	.8138	.8045
社会	836	.7639	.6423	.6979
国際	584	.7833	.7860	.7846
スポーツ	452	.8950	.9424	.9181
各局	537	.7635	.5773	.6575
首都圏	585	.6925	.8393	.7589
トータル	4038	.7816	.7712	.7764

「スポーツ」、「政治」、「経済」は適合率、再現率とも他と比べて高い。これらの分野はキーワードで明確に特徴付けられることを示している。詳細に見ると「社会」，「各局」，「首都圏」の間で誤判定が多い（社会→首都圏：117，各局→首都圏：80）ことが判明した。

手法別の適合率、適合率を表５に示す。

表５．　手法別の精度（学習サンプル）

精度	tf	tf*idf	tf*rlb	DT
適合率	.7357	.7816	.7678	.7947
再現率	.7259	.7712	.7576	.7841

表５を見るとDecision Tree法が一番精度が良いことが分かるがｔｆ＊ｉｄｆ法とそれほどの差はない。Ｆ値で分野毎の誤分類の傾向を見てみると、誤判定の傾向はどの手法でもほぼ同じであったが、ｔｆ＊ｉｄｆ法と比べるとｔｆ＊ｒｌｂ法は「経済」で値が低く、Decision Tree法はｔｆ＊ｉｄｆ法より「社会」で優れ、「国際」で劣っていた。

テストサンプルを９６年１０月のＮＨＫニュース記事１カ月分（記事数3,728，文数20,052，文字数829,280）として、同様に適合率、再現率を測定した。結果を表６に示す。当然のことながら全て

表６．　手法別の精度（テストサンプル）

精度	tf	tf*idf	tf*rlb	DT
適合率	.5940	.6449	.6277	.6364
再現率	.5829	.6328	.6159	.6245

の手法で精度が低下している。なお、テストサンプルにおいて１つもキーワードを含まない記事は70であった。（社会：29，スポーツ：11など）

表５、表６の結果を見るとわずか７つしか分野がないにも関らず分類精度が低すぎるように思われる。この理由は前述した「政治」，「経済」などは内容の分類ではなく、実は「政治部」，「経済部」などの記事を担当した部署の分類に過ぎないからである。特に「社会」、「各局」、「首都圏」の間で誤判別が多いのは、それらが取り扱う話題が極めて似ていることに起因している。そこで、「各局」，「首都圏」を除いて５つの分野で適合率、再現率を測定し直してみた。（学習サンプル：2916，テストサンプル：2549）

結果を表７に示す。これを見ると、ｔｆ＊ｉｄｆ法が一番優れていることが分かる。Decision Tree法は学習サンプルとテストサンプルで精度の差が大きく、ロバストでないことが分かる。

先に述べたように、「分野」は内容分類になっていない。そこで、テストサンプルにおいてｔｆ＊ｉｄｆ法とｔｆ＊ｒｌｂ法で内積の大きかったもの700位まで取って、判別結果を吟味した。700位までの適合率は２つの手法で全く同じで0.9457であった。700位までに含まれる記事のうち538個が一致していた。このうち、正解と異なるものが29個（国際→経済：16，社会→政治：８など）であり、

表７．　手法別の精度（５分野）

	精度	tf	tf*idf	tf*rlb	DT
学習	適合率	.8378	.8656	.8430	.8423
	再現率	.8254	.8529	.8306	.8299
テスト	適合率	.7648	.8013	.7672	.7346
	再現率	.7474	.7831	.7497	.7179

両手法とも全て同じ結果を示していた。内容を見ると、そのうちの１つを除いて、分類としてはむしろ判別結果の方が妥当だと思われた。上位700個のうち、判別結果の方が間違っていると思われるのはｔｆ＊ｉｄｆ法で７、ｔｆ＊ｒｌｂ法で４しかなかった。つまり、700位までの真の適合率はｔｆ＊ｉｄｆ法で0.9900、ｔｆ＊ｒｌｂ法で0.9943であると思われる。この結果はキーワードの選定結果と手法の妥当性を示していると考えられる。

５．おわりに

最初に平均情報量基準を用いて、字面処理だけで記事からキーワードの候補を抽出し、文字列の包含関係と信頼度（評価尺度）を用いて、キーワードの選定を行なった。このキーワードを用いて記事の分類実験を行ない、選定されたキーワードと判別手法の妥当性を確認した。実験結果は、記事の内容検索という観点で見た場合、少なくても評価値（内積）の大きいものについてはもとの分類より判別結果の方が信頼できることを示していた。記事データベースを作成する際には、ｔｆ＊ｒｌｂ法（and/or ｔｆ＊ｉｄｆ法）が、分野の適否を半自動的に検証するのに役立ちそうである。

今後、さらに記事検索や分類に必要な要件を追求し、キーワードの抽出、キーワードの選定、判別手法の改良を図っていく予定である。

参考文献

1）浦谷則好：ニュース原稿データベースからの表現パターンの抽出，情処５０全大１Ｒ－８，(1995)

2）下畑さよりほか：隣接文字の分散値を用いた定型表現の自動抽出，情処研資NL110-11，(1995)

3）徳永健伸ほか：重み付きＩＤＦを用いた文書の自動分類について，情処研資NL100-5，(1995)

4）岩山真ほか：自動文書分類のための新しい確率モデル，情処研資H33-9，(1995)

5）G.Salton and C. S. Yang:On the specification of term vales in automatic indexing, *Journal of Documentation* , Vol.29, No.4, pp.351-372(1973)