音声言語情報処理 22−15
(1998. 7. 25)

ニュース音声の記事分類における キーワード選択法の比較

鷹尾 誠一　緒方 淳　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷1-5
Tel: 077-543-7427
E-mail: {tail,ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

あらまし ビデオ・オン・デマンドを目指したニュースデータベースを構築するには、ニュース記事を話題(トピック)毎に分類する必要がある。本研究ではニュースを分類する際に必要となるキーワードについて、代表的なキーワード選択法である χ^2 値、相互情報量、TF-IDF 等について特質を比較した。また、得られたキーワードを分類の際にどう使えば分類率が良くなるかについても比較、検討を行った。

キーワード：キーワード選択、音声ディクテーション、記事分類

Comparison of Keyword Selection for Classification of News Speech Articles

Seiichi Takao　Jun Ogata　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: {tail,ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

Abstract In order to construct a news database with a function of video on demand (VOD), it is required to classify a news articles into topics. In this study, we implemented and compared keyword selection methods such as χ^2, mutual information and TF-IDF. These selected keywords are used to classify the articles after news speech dictation. Further more we compared the classification methods which use the selected keywords.

Key words : keyword selection, speech dictation, article classification

1 はじめに

近年、放送の多チャンネル化により、多くのニュース番組が放映されるようになった。これを受けて視聴者には、知りたいニュースだけを見たいという要求が生じている。この要求に対応するには、ニュース記事を分類してデータベースを構築しておく必要がある。しかし、人手でニュース記事を分類することは不可能であり、機械によるニュース記事の自動分類が望まれる。この点から、ニュース音声に対する話題分類、話題同定の研究が行われている。[1] - [11]。

ニュース記事を分類するには、記事中のキーワードを抽出し、キーワードと分類分野との関係を用いて分類するのが一般的である。ここで問題となるのは、ニュース記事の分類性能が記事、特にニュース音声記事からキーワードを抽出する精度だけではなく、キーワードと分類分野との関係をどのように設計するかに依存している点である。

我々はこれまで、人手で与えたキーワードを基にニュース記事を分類する研究を行ってきた。今回、幾つかのキーワード選択法によりキーワードを自動選択し、選択されたキーワードと分類分野との関連度を計算した。この関連度を用いて新聞記事テキスト、ニュース音声記事テキスト、ニュース音声記事のディクテーションに対して分類実験を行い、キーワード選択法を比較した。

また、ニュース記事を分類する方法においても、キーワードと分類分野との関連度をどう使うかによって、分類精度が変わることに着目し、関連度の用い方を幾つか変えて実験を行った。

2 記事分類の概要

RWC から提供されている形態素解析された93年の記事11,000件を、朝日新聞92年度の分類表索引[12]を用いて分類した。朝日新聞記事データベース分類表索引には、約1万2千のキーワードが載っており、キーワード毎に分類づけがなされている。この分類表索引では、大分類として、総類、政治、経済、労働、文化、科学、社会、事件、スポーツ、国際の10分野が示されている。実験では、この10分野に対して、記事の分類を行った。

11,000記事の正解分野は文献[3]に述べる従来法で求めた。この分類結果を用いて、キーワード選択、キーワードと分類分野の関連度計算、記事分類の評価を行った。11,000記事の正解分野の索引付けを人手で行うことも可能であるが、その場合においても、11,000件の正解分野が少し異なるだけで、キーワード選択と関連度の計算法、分類評価はそのまま用いることができる。

分類した11,000記事のうち10,000記事を学習データ、1,000記事を評価用データとした。まず、学習データ10,000記事から名詞だけを抜き出し、この名詞からキーワードを選択するとともに、キーワードと分野間の関連度を求めた。次に、選択したキーワードと関連度に基づいて、残り1,000記事を分類した。さらに、選択したキーワードを用いて、ニュース音声記事テキストとニュース音声記事のディクテーション結果に対して分類実験を行った。

3 新聞記事からのキーワード選択法

新聞記事中の単語w_iと、その記事が属する分野t_jとの間で関連度γ_{ij}を求め、閾値処理を行って、関連度γ_{ij}の大きいものをキーワードとして抽出する。以下には、単語w_iと分野t_jとの関連度γ_{ij}を求める幾つかの方法を示す。

3.1 χ^2 法

χ^2検定で用いられるχ^2値は、分野における単語の偏りを示す指標として用いることができる。単語w_iの出現確率は全分野を通じて等しいという仮説を設定し、この仮説に基づいて単語w_iの分野t_jにおける予測頻度m_{ij}を計算する。また、各単語w_iについて分野t_jにおける頻度x_{ij}を求める。このx_{ij}とm_{ij}を基に、式(1)に従って、${\chi_{ij}}^2$値を求める。もし、この${\chi_{ij}}^2$値が十分大きな値になれば特定の分野に偏って表われる単語という事になり、分野の識別に有効な単語と見なせる。別の見方をすれば分野t_jとの共起が高い単語w_iを、キーワードとして選んでいると解釈できる。

$$ {\chi_{ij}}^2 = \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \tag{1} $$

$$ m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij} \tag{2} $$

m : 異なり単語数

n : 分野数

x_{ij} : 単語w_iの分野t_jにおける頻度

m_{ij} : 単語w_iの分野t_jにおける予測頻度

3.2 相互情報量

単語w_iと分野t_jとの相互情報量$i(t_j;w_i)$とは、単語w_iを知ることによって、分野t_jに関して得られる情報量のことであり、次式で表される。

$$
\begin{aligned}
i(t_j;w_i) &= i(t_j) - i(t_j|w_i) \qquad (3)\\
&= -\log P(t_j) + \log P(t_j|w_i)\\
&= \log \frac{P(t_j,w_i)}{P(t_j)P(w_i)}
\end{aligned}
$$

この情報量は分野t_jがそもそも持っていた情報量$i(t_j)$から、単語w_iを知った後でもまだ分野t_jが有している情報量$i(t_j|w_i)$との差として定義される。この差が単語w_iによって運ばれる分野t_jに関する情報量となる。式(3)より相互情報量は、分野t_jの生起確率$P(t_j)$と単語w_iの生起確率$P(w_i)$が独立していれば小さくなり、依存していれば大きな値となる。すなわち、単語w_iが分野t_jに偏っていれば、相互情報量は高くなる。

なお、$P(t_j,w_i)$、$P(w_i)$、$P(t_j)$は次のように定義できる。

$$
\begin{aligned}
P(t_j,w_i) &= \frac{x_{ij}}{\sum_i \sum_j x_{ij}} \qquad (4)\\
P(w_i) &= \frac{\sum_j x_{ij}}{\sum_i \sum_j x_{ij}}\\
P(t_j) &= \frac{\sum_i x_{ij}}{\sum_i \sum_j x_{ij}}
\end{aligned}
$$

この式を相互情報量の式(3)に代入すると、対数の中の式は次のようになる。

$$
\begin{aligned}
\frac{\frac{x_{ij}}{\sum_i \sum_j x_{ij}}}{\frac{\sum_i x_{ij}}{\sum_i \sum_j x_{ij}} \cdot \frac{\sum_j x_{ij}}{\sum_i \sum_j x_{ij}}} &= \frac{\frac{x_{ij}}{\sum_i \sum_j x_{ij}}}{\frac{m_{ij}}{\sum_i \sum_j x_{ij}}} \qquad (5)\\
&= \frac{x_{ij}}{m_{ij}}
\end{aligned}
$$

x_{ij}/m_{ij}が大きい単語w_iをキーワードとして選ぶ場合には、x_{ij}/m_{ij}の代わりに$(x_{ij}/m_{ij}-1)^2$を用いてもよい。これは$(x_{ij}-m_{ij})^2/(m_{ij})^2$になる。

相互情報量基準によるキーワード選択法は、χ^2法によるキーワード選択法と極めてよく似ているといえる。

3.3 TF-IDFによるキーワードの選択

TF-IDFは式(6)で表され、単語w_iが記事a_kに現れる回数が高ければ高いほど、TF(Term Frequency)が高くなり、単語w_iが現れる記事数が少なければ少ないほど、IDF(Inverse Document Frequency)が高くなる。したがって、TFは頻度の高い単語という性質を表し、IDFはその分野に偏って現れる単語という性質を表している。

$$
\begin{aligned}
TF \cdot IDF &= TF(w_i,a_k) \cdot IDF(w_i) \qquad (6)\\
TF(w_i,a_k) &= \text{単語}w_i\text{が記事}a_k\text{に現れる回数}\\
IDF(w_i) &= \log \frac{\text{索引対象の全記事数}}{\text{単語}w_i\text{が現れる記事数}}
\end{aligned}
$$

式(6)のTF-IDFでは、単語w_iが現れる記事数を求めているが、記事の中に現れる単語w_iの頻度にかかわらず、記事数がカウントされている[11]。

このため、単語w_iの頻度が低い記事も含めてしまい、IDFが小さく見積もられる可能性がある。この問題を避けるため、式(7)に示すように、単語w_iが現れる記事数の代わりに単語w_iの頻度を用いる方法が提案されている[6]。

$$
I_i = g_i \log(G_A/G_i) \qquad (7)
$$

w_i : 新聞記事中の名詞$(i=1,2,...,N)$
N : 語彙サイズ(名詞のみ)
g_i : (注目している特定の)
新聞記事中の名詞w_iの頻度
G_i : 全ての新聞記事中の名詞w_iの頻度
$G_A = \sum_i G_i$

本研究では、単語w_iと分野t_jとの関連度を求め、これを閾値処理することでキーワードを選択している。このため、単語w_iと分野t_jとの関連度を求める必要があるが、TF-IDFには本来、分野という概念がない。そこで、次式のようにTF-IDFを変形してキーワードを抽出した。

$$
\begin{aligned}
TF \cdot IDF &= TF(w_i,a_k) \cdot IDF(w_i) \qquad (8)\\
TF(w_i,a_k) &= \text{単語}w_i\text{が記事}a_k\text{に現れる回数}\\
IDF(w_i) &= \log \frac{\text{全分野数}(10)}{\text{単語}w_i\text{が現れる分野数}}
\end{aligned}
$$

ここで単語w_iが現れる分野数とは、分野毎に分類された学習データ10,000記事を対象に単語w_iがどの分野に現れたかを調べて決定している。

3.4 負の値を持つ χ^2 法

従来のχ^2法では、単語w_iが分野t_jで発生する頻度x_{ij}と予測頻度m_{ij}の差を2乗して求めている。このため$x_{ij} < m_{ij}$の場合であっても、χ^2値は正の値として計算され、単語w_iは分野t_jにおいて偏りがあると判断される。こ

の問題を解決するために文献[7]では、式(9)に示すように改良方法が提案されている。

$$\chi_{ij}{}^2 = \frac{(x_{ij} - m_{ij}) \cdot |x_{ij} - m_{ij}|}{m_{ij}} \tag{9}$$

3.5 重み付相互情報量

従来の相互情報量では、単語の発生頻度が小さくても、単語と分野の依存度(共起関係)が高ければ大きな値を示してしまう。この問題を解決する方法として文献[7]では式(10)に示すように、従来の相互情報量に単語w_iと分野t_jの同時確率を重みとしてかける方法が提案されている。この値は、単語w_iが分野t_jに従属すると見た場合の確率分布と、独立すると見た場合の確率分布間のダイバージェンスとなっている。

$$I(w_i; t_j) = P(w_i, t_j) \cdot \log \frac{P(w_i, t_j)}{P(w_i)P(t_j)} \tag{10}$$

3.6 相対相互情報量

式(3)に示す相互情報量は、$P(w_i|t_j) < P(w_i)$の場合に負の値となる場合がある。また、相互情報量には分野間の相対的な関係が考慮されていない。そこで式(11)に示すように、$P(w_i)$の代わりに、全ての分野の中で単語w_iが出現する確率が最小のもの、$\min_k P(w_i|t_k)$を用いる方法を提案する。こうすることで、単語w_iの分野t_jにおける出現確率のダイナミックレンジを大きくすることができる。この値は、相互情報量、あるいはχ^2法における基本表現x_{ij}/m_{ij}の分母を、$\min_k P(w_i|t_k)$として強調したものになっている。

$$\frac{p(w_i|t_j)}{\min_k p(w_i|t_k)} \tag{11}$$

4 記事の分類方法

1つの記事が与えられると、記事中のキーワードw_iを全て抽出する。このキーワードw_iと分野t_jとの関連度γ_{ij}を基に、キーワードw_iが分野t_jの分類に寄与する割合C_{ij}を計算する。最後に、1つの記事中に含まれているキーワードの発生回数をN_iとすると、分類寄与率C_{ij}を次式のように総和して、記事xと分野t_jとの類似度を求める。

$$S(x, t_j) = \sum_i N_i \cdot C_{ij} \tag{12}$$

この類似度の大きな分野t_jに記事を分類する。以下には、キーワードw_iの分野t_jに対する分類寄与率C_{ij}の幾つかの計算方法を示す。

表1: キーワード「日米」の3つの分野に対する出現回数と関連度γ_{ij}の例

分野	出現回数	関連度
政治	3	30
経済	2	20
社会	0	0

4.1 関連度に基づく類似度

キーワードw_iと分野t_jとの関連度γ_{ij}を、分類寄与率として記事を分類する方法である。例えば、キーワード「日米」が分野「政治」、「経済」、「社会」に対して表1のような関連度を持っている場合、分類寄与率C_{ij}は次のようになる。

$$C_{\text{日米, 政治}} = 30 \qquad C_{\text{日米, 経済}} = 20$$
$$C_{\text{日米, 社会}} = 0$$

この分類寄与率に基づく、入力記事xと分野t_jの類似度を$S^1(x, t_j)$と表す。

4.2 関連度の正規化に基づく類似度

キーワードw_iと分野t_jとの関連度γ_{ij}を式(13)のように正規化して分類寄与率C_{ij}を求め、記事を分類する方法である。表1の例では次のようになる。

$$C_{\text{日米, 政治}} = \frac{30}{30+20+0} \doteqdot 0.6 \qquad C_{\text{日米, 経済}} = \frac{20}{30+20+0} \doteqdot 0.4$$
$$C_{\text{日米, 社会}} = \frac{0}{30+20+0} \doteqdot 0.0$$

$$C_{ij} = \frac{\gamma_{ij}}{\sum_j \gamma_{ij}} \tag{13}$$

この方法は、キーワードw_iの分野t_jに対する関連度を、キーワード間で比較可能にする効果がある。

この分類寄与率に基づく、入力記事xと分野t_jの類似度を$S^2(x, t_j)$と表す。

4.3 キーワードの出現回数に基づく類似度

キーワードw_iが分野t_jにおいて出現した回数N_{ij}を基に、分類寄与率C_{ij}を式(14)のようにして求め、記事を分類する方法である。

$$C_{ij} = \frac{N_{ij}}{\sum_i N_{ij}} \tag{14}$$

この方法では、分野t_jを表すベクトルを

$$v_j = (C_{1j}, C_{2j}, \cdots, C_{ij}, \cdots, C_{nj}) \tag{15}$$

と表すことができる。また、1つの記事に含まれているキーワードの出現回数N_iを基に、その記事をベクトルxとして次のように表すことができる。

$$x = (N_1, N_2, \cdots, N_i, \cdots, N_n) \tag{16}$$

したがってこの分類方法は、

$$\sum_i N_i \cdot C_{ij} = v_j \cdot x = ||x|| \cos\theta_j \tag{17}$$

であることから、単純類似度法と同じである。この分類寄与率に基づく、入力記事xと分野t_jの類似度を$S^3(x, t_j)$と表す。

4.4 キーワードの出現回数の正規化に基づく類似度

キーワードw_iが分野t_jにおいて出現した回数N_{ij}を基に、分類寄与率C_{ij}を式(18)のように事後確率として求める方法である。例えば、キーワード「日米」が分野「政治」、「経済」、「社会」において出現した回数が表1のようになっている場合、分類寄与率C_{ij}は次のようになる。

$$C_{\text{日米,政治}} = \frac{3}{5} = 0.6 \qquad C_{\text{日米,経済}} = \frac{2}{5} = 0.4$$
$$C_{\text{日米,社会}} = \frac{0}{5} = 0.0$$

$$C_{ij} = \frac{N_{ij}}{\sum_j N_{ij}} \tag{18}$$

この分類寄与率に基づく、入力記事xと分野t_jの類似度を$S^4(x, t_j)$と表す。

5 新聞記事テキストの分類実験

毎日新聞の記事テキスト10,000件を学習データとして、3節で述べた6種類のキーワード選択法によってキーワードを選択した。このキーワードを基に4節で述べた4種類の分類寄与率と類似度S^1、S^2、S^3、S^4を求め、毎日新聞記事1,000件を評価データとして分類実験を行った。実験結果を図1–4に示す。図の横軸は用いたキーワード数であり、縦軸は記事分類率である。図1–4は4種類の類似度に対する分類結果である。図中、$Chi2$はχ^2法、$Mutual$は相互情報量、$Mutual-W$は重み付相互情報量、$TF-IDF$はTF-IDF、$Mutual-R$は相対相互情報量、$Chi2-new$は負の値を持つχ^2値によるキーワード抽出法である。各図において、最大の記事分類率とそのとき用いられたキーワード選択法、並びにキーワード数を表2に示す。

表より、S^2(正規化関連度) > S^4(正規化出現回数) > S^1(関連度) > S^3(出現回数)という結果が得られた。S^2の正規化とは、関連度γ_{ij}を$\sum_i \gamma_{ij}$で正規化したものを分類寄与率として用いるものであり、S^4の正規化は出現回数N_{ij}を$\sum_j N_{ij}$で正規化したものを分類寄与率として用いるものである。ともに事後確率化することにより、キーワード間で分類寄与率が比較できるようになっており、これによりS^2、S^4の分類率が高くなったと考えられる。また、S^2、S^1といった関連度を用いる方が、S^4、S^3といった出現回数だけを用いるものより分類率が向上している。

一方、キーワード選択手法では、χ^2値、負の値を持つχ^2値、重み付相互情報量、TF-IDFが良い結果を示しているが、分類寄与率の求め方に依存しており、一般に優劣は判定しにくい。ただ、相互情報量を用いるもの(相互情報量、重み付相互情報量、相対相互情報量)は、キーワード数が少ないと分類率が低くなる傾向があり、高い分類率を得るには、キーワード数を多くしておく必要がある。これらに対して、χ^2値、負の値をもつχ^2値、TF-IDFはキーワードにあまり依存せず高い値を示している。

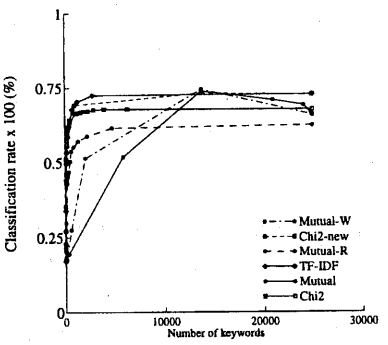

図 1: S^1による新聞記事テキストの分類率

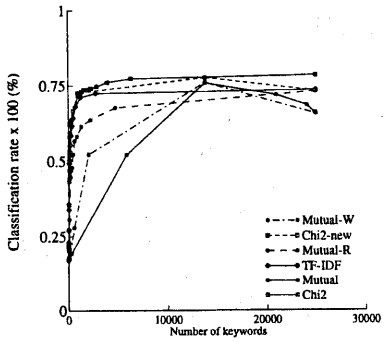

図 2: S^2による新聞記事テキストの分類率

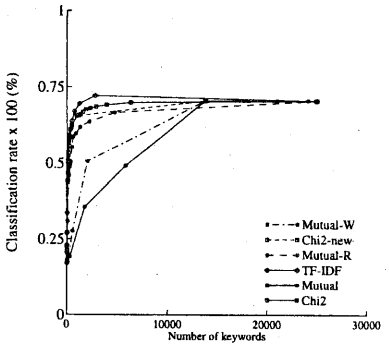

図 3: S^3 による新聞記事テキストの分類率

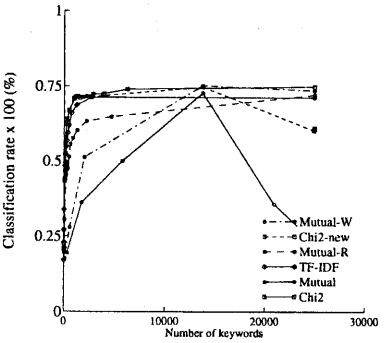

図 4: S^4 による新聞記事テキストの分類率

表 2: 新聞記事テキストに対する最大の分類率とキーワード選択法

分類法	最大分類率	キーワード選択法	キーワード数
S^1	74.5%	Mutual-W	13,755
S^2	78.4%	Chi2	24,920
S^3	72.2%	TF-IDF	2,782
S^4	75.2%	Mutual-W	13,755

6 ニュース音声記事テキストに対する分類実験

評価用データとして、新聞記事の代わりに、93年、94年のNHK1時のニュース55 記事を人手でテキスト化したものを用いた。この実験の目的は、評価用データが新聞記事からTVニュース記事になった場合の、キーワード選択法と分類率を調べることである。

4種類の類似度S^1、S^2、S^3、S^4に対する分類結果を図5–8 に示す。各図において最大の記事分類率とそのとき用いられたキーワード選択法、並びにキーワード数を表3に示す。表より、S^2(正規化関連度) > S^4(正規化出現回数) = S^1(関連度) > S^3(出現回数)という結果が得られた。全体の傾向は、新聞記事テキストの場合と、ほぼ同じである。しかし、ニュース音声記事テキストに対する分類率の方が6%以上高い分類率を示している。この理由としては、新聞記事テキスト1,000件とニュース音声記事テキスト55件中に含まれている分野バランスに差があると思われる。

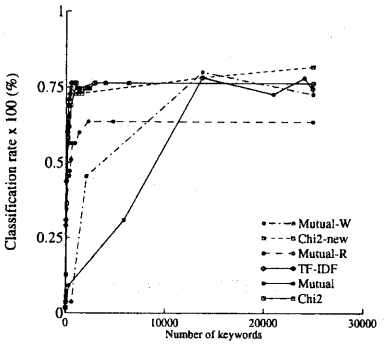

図 5: S^1 によるニュース音声記事テキストの分類率

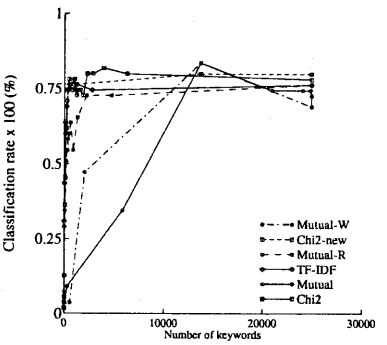

図 6: S^2 によるニュース音声記事テキストの分類率

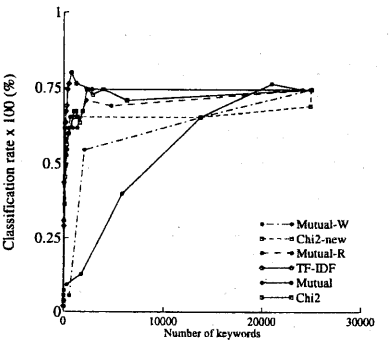

図 7: S^3 によるニュース音声記事テキストの分類率

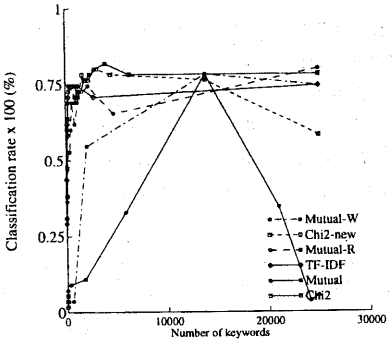

図 8: S_4 によるニュース音声記事テキストの分類率

表 3: ニュース音声記事テキストに対する最大の分類率とキーワード選択法

分類法	最大分類率	キーワード選択法	キーワード数
S^1	81.8%	Chi2-new	24,887
S^2	83.6%	Mutual,Mutual-W	13,755
S^3	80.0%	TF-IDF	745
S^4	81.8%	Chi2	3937

7 ニュース音声ディクテーション結果の分類実験

7.1 ディクテーション

7.1.1 実験条件

今回、用いた言語モデルは、毎日新聞CD-ROM版の45ヶ月分(91年1月～94年9月)の記事から学習したものである。語彙数20Kのback-off bigramで、back-off smoothingにはwitten-bellの推定を用いている。bigramに対するcut-offは1とした。

音響モデルは、男性不特定話者HMMで、単語間の音素文脈依存も考慮したcross-word triphoneモデルである。学習には、日本音響学会新聞記事読み上げ音声コーパスのうち、男性話者137人分の21782発話を用いた。音響特徴量には39次元の特徴パラメータ(12次元のメルケプストラム係数とパワー、およびそれぞれのΔ、ΔΔ係数)を用いた。

7.1.2 連続音声認識実験

評価用音声データには、6節と同じ93年、94年のNHK1時のニュース55記事分(総計1.13時間、1記事平均71秒)を用いた。bigram言語モデルの学習データとは時期的にclosedなデータである。55記事のデータを1発話毎に区切った総発話数409のデータに対してディクテーションを行った。記事分類用音声データの諸元を表4に、ディクテーション結果を表5に示す。

表 4: 記事分類用音声データ

話者数	3名
総発話数	409
perp	78.3
20K 未知語率	0.8%

perp: test-set perplexity

表 5: ディクテーション結果

単語正解率	単語正解精度	単語誤り率
85.6%	80.3%	14.4%

7.2 分類実験

4種類の類似度S^1、S^2、S^3、S^4に対する分類結果を図9–12に示す。各図において最大の記事分類率とそのとき用いられたキーワード選択法、並びにキーワード数を表6に示す。表より、S^2(正規化関連度) > S^4(正規化出現回数) > S^3(出現回数) > S^1(関連度)という結果が得られた。全体の傾向は、新聞記事テキストやニュース音声記事テキストとおおむね同じである。6節の結果と比べて3%程度、ニュース音声記事テキストに対する分類率が低下するものの、良好な結果を示している。これは音声ディクテーションの単語正解精度が80.3%と高く、また湧き出し誤り単語列に記事分類に影響のある単語がほとんど見られなかったためである。

表 6: ニュース音声ディクテーション結果に対する最大の分類率とキーワード選択法

分類法	最大分類率	キーワード選択法	キーワード数
S^1	76.4%	Chi2,Chi2-new	13,755
S^2	83.6%	Mutual-R,Mutual-W	24,940
S^3	78.2%	TF-IDF	366
S^4	81.8%	Mutual-R	24,940

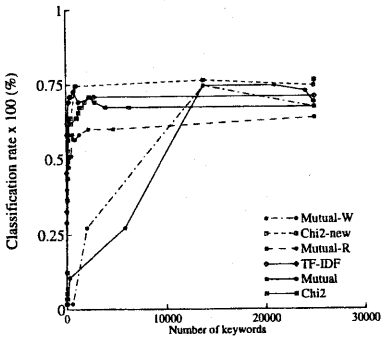

図 9: S^1 によるニュース音声記事ディクテーションの分類率

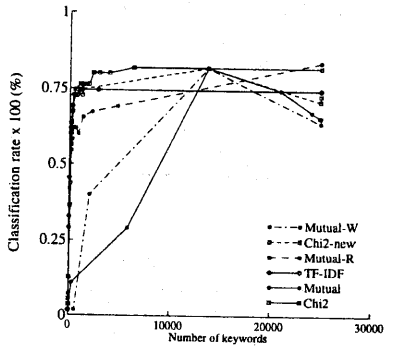

図 10: S^2 によるニュース音声記事ディクテーションの分類率

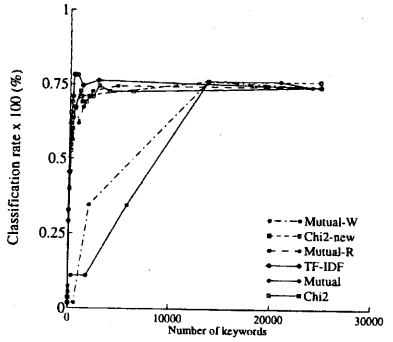

図 11: S^3 によるニュース音声記事ディクテーションの分類率

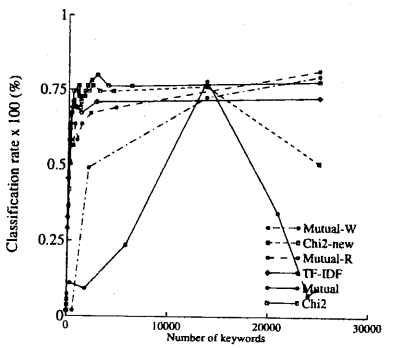

図 12: S^4 によるニュース音声記事ディクテーションの分類率

8 おわりに

新聞記事テキスト、ニュース音声記事テキスト、ニュース音声記事ディクテーション結果に対して、6種類のキーワード抽出法を適用し、キーワードを抽出するとともにキーワードと分野間の関連度を計算した。この関連度を基に、4種類の分類寄与率を計算し、記事分類を行った。χ^2 値、TF-IDF関連のキーワード選択法では、キーワード数にかかわらず高い記事分類率を示すこと、これに対して、相互情報量関連のキーワード選択法では、キーワード数が少ない場合には記事分類率が低くなることが分かった。

一方、分類手法では、キーワードの関連度、出現回数を用いるより、事後確率的に正規化すると記事分類率が高くなることがわかった。ニュース音声記事のディクテーション結果に対するキーワード選択と記事分類は、ニュース音声記事テキストに対して3%程度劣化するものの、記事分類率は最高で83.6%が得られた。

今後は、TF-IDFを分野数だけでなく、記事数、単語数を含むものに拡張する予定である。

参考文献

[1] 横井謙太郎, 河原達也, 堂下修司 : "キーワードスポッティングに基づくニュース音声の話題同定", 情処研報, **SLP95**-6-3, pp.15-20 (1995-5).

[2] 亀山晋, 中里収, 白井克彦 : "ワードスポッティングに基づく意図抽出", 信学技報, **SP94**-64, pp.9-16 (1994-12).

[3] 櫻井光康, 有木康雄 : "キーワードスポッティングによるニュース音声の分類と索引付け", 信学技法, **SP96**-66, pp.37-44 (1996-11).

[4] 大附克年, 松岡達雄, 松永昭一, 古井貞熙 : "ニュース音声を対象とした大語彙連続音声認識と話題抽出", 信学技報, **SP97**-27, pp.67-74 (1997-06).

[5] 城塚音也 桑田喜隆 小泉宣夫 : "対話音声を対象とした話題同定の検討" 日本音響学会講演論文集, pp.5-6 (平成10年3月).

[6] 高木幸一 桜井直之 岩崎淳 古井貞熙 : "ニュース音声を対象とした言語モデルと話題抽出の検討" 信学技報, **SP98**-33, pp.73-80 (1998-06).

[7] K. Ohtsuki, T. Matsuoka, S.Matsunaga, S.Furui : "TOPIC EXTRACTION MULTIPLE TOPIC-WORDS IN BROADCAST-NEWS SPEECH" ICASSP98, pp.329-332 (1998).

[8] 緒方淳 有木康雄 : "ニュース記事分類におけるディクテーションとワードスポッティングの比較" 信学技報, **SP98**-32, pp.67-72 (1998-06).

[9] 恒川俊克 山下洋一 溝口理一郎 : "キーワードスポッティングに基づくニュース音声の話題分類" 音声言語情報処理, pp.61-68 (1998.2.6).

[10] 緒方淳 森晴 有木康雄 : "単語bigramを用いた日本語ニュースディクテーションによる記事分類", 音響学会平成10年度春季大会, 2-Q-16, pp.151-152, (1998-03).

[11] 小泉敦延 奥田敬 伊藤秀一 : "文書集合における重要語の抽出", 信学技報, **DE98**-1, pp.1-6, (1998-05).

[12] "朝日新聞記事データベース分類表索引", 朝日新聞社ニューメディア本部, (1992).