

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

恒川俊克¹ 山下洋一² 溝口理一郎¹

¹ 大阪大学産業科学研究所知識システム研究分野 ² 立命館大学理工学部情報工学科
〒567 茨木市美穂ヶ丘 8-1 〒525-77 滋賀県草津市野路東 1-1-1
E-mail: {tsune,miz}@ei.sanken.osaka-u.ac.jp,yama@cs.ritsumeikai.ac.jp

あらまし: 大量のニュース音声データベースからある特定の話題に関するニュースのみを検索しその情報を得るようになるにはニュース音声に対する話題インデキシングの技術が必要になると考えられる。そこで本研究ではニュース音声に対しキーワードスポッティングに基づいた話題のインデキシングを行う。まず話題との相互情報量、単語長を選出基準として、話題の同定に貢献する名詞 3000 単語をキーワードとして選出した。このキーワードの話題に関する条件付き確率を基にテキストベース上で記事ごとに話題同定の実験を行ったところ約 76%の記事を正しく同定することができた。さらにニュース音声に対しスポッティングを行い、抽出されたキーワードの音韻スコアと条件付き確率を基にして話題ごとに「話題の確からしさ」の計算を行い、話題同定を行った。正しい話題が話題同定結果の上位 3 位以内に含まれていれば正解分類とした場合、約 66.5%を正しく同定することができた。

和文キーワード: キーワードスポッティング, 話題同定, キーワード選択, ニュース音声

Topic Classification of News Speech based on Keyword Spotting

Toshikatsu Tsunekawa¹ Yoichi Yamashita² Riichiro Mizoguchi¹

¹ I.S.I.R.Osaka University ² Dep. of Computer Science,Ritsumeikan University
8-1,Mihogaoka,Ibaraki-shi,Osaka,567 Japan 1-1-1,Noji-Higashi,Kusatsu-shi,Shiga,525-77 Japan
E-mail: {tsune,miz}@ei.sanken.osaka-u.ac.jp,yama@cs.ritsumeikai.ac.jp

Abstract: For retrieving news data related to a specific topic from a great amount of news data base, we need a technique of indexing speech data with topics. In this report, we try the topic identification for news speech based on keyword spotting. To begin with, we selected three thousands of nouns as keywords which contribute to topic identification, based on criterion of mutual information and a length of word. This set of keywords identified correct topics of 76 percent of text article data from newspaper data base. Further, we performed keyword spotting for TV news speech and identified a topic by computing possibilities of all topics based on phonetic scores of spotted words and topic probability of the words. Topic identification rate is 66.5 percent assuming that identification is correct if the correct topic is included in the first three places of the result of topic identification.

Keywords: keyword spotting, topic identification, keyword selection, news speech

1 はじめに

近年では情報化の発展に伴い、膨大な音声データやテキストデータから必要なデータを効率よく検索し取り出したいという要求が高まりつつある。この要求を実現するためにはシステムが自動的にそのデータベース中のそれぞれの記事に話題情報を付与し、話題ごとに分類を行うことが必要であり、音声データに対する話題同定の研究がさかんに行われている [2-10]。

人間があるデータが特定の話題に関する情報であると判断できるのはそのデータ中に含まれる特定の話題を反映させるいくつかのキーワードをピックアップしその情報を基に話題の同定を行っているからであると考えられる。つまり話題を決定するには特定の話題との結び付きの強い単語が重要な要素であるといえる。

そこで本研究ではキーワードスポッティングを行い、抽出された単語の持つ話題との関連性の情報と音韻スコアを利用して話題同定を行う。今回は評価する音声データとしてニュース音声を選んだ。

まず2章ではキーワードスポッティングに基づいた話題のインデキシングの手法について述べる。3章では話題同定に必要なキーワードの選出手法について述べる。4章では選出したキーワードが話題同定に妥当であるか、またキーワードに基づいた話題のインデキシングが有効であるかを確認するため、まず新聞記事のテキストデータに対して話題の同定実験を行った。5章ではキーワードスポッティングの手法について述べる。その後6章でニュース音声に対してキーワードスポッティングを行いその結果を基に話題同定の実験を行う。

2 キーワードに基づいた話題同定

本報告で提案する話題同定手法の処理の流れを図1に示す。

ニュース音声の話題同定では話題の境界が未知であるため、ある一定の窓長 M の分析区間を考え、この区間に対してそれぞれの話題らしさを計算する。ニュース音声の開始から分析区間をずらしながらこのような分析を繰り返し行い、それぞ

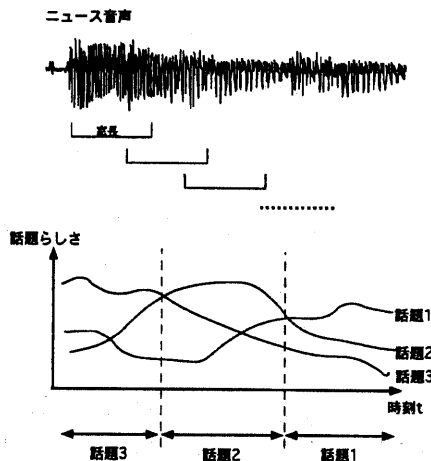


図1: キーワードスポッティングに基づいた話題同定

れの話題らしさの時間変化パターンを得る。話題の境界は図1に示すように、事後的に決定されることになる。各分析区間ではあらかじめ設定される数千程度のキーワードのセットに関してスポッティングを行い、キーワードを検出し、得られた単語列から単語と話題との関連性に基づいて話題らしさを計算する。

3 キーワードの選択手法

3.1 相互情報量によるキーワード選択

ニュース音声から検出すべきキーワードはある話題と関連性の強い単語でなければならない。本報告では単語 w と話題 T の相互情報量に基づいてキーワードを選択する。

T を記事の話題を表す確率変数、 W を記事中のある単語を表す確率変数とすると、話題の決定において単語が何であるかが分かることによって得られる情報量、すなわち相互情報量 $I(T; W)$ は、

$$\begin{aligned} I(T; W) &= I(W; T) \\ &= H(W) - H(W|T) \\ &= - \sum_w p(w) \log_2 p(w) \end{aligned}$$

$$\begin{aligned}
& + \sum_T p(T) \sum_w p(w|T) \log_2 p(w|T) \\
= & - \sum_w p(w) \log_2 p(w) \\
& + \sum_w \sum_T p(T) p(w|T) \log_2 p(w|T) \\
= & \sum_w [-p(w) \log_2 p(w) \\
& + \sum_T p(T) p(w|T) \log_2 p(w|T)] \quad (1)
\end{aligned}$$

となり、式(1)に示すように、様々な単語からの寄与を合計したものとなる。従って、式(1)の[]内の値

$$G(w) = -p(w) \log_2 p(w) + \sum_T p(T) p(w|T) \log_2 p(w|T)$$

を「話題 T の決定のために単語 w が持つ寄与度」とし、 $G(w)$ が大きい単語ほど特定の話題同定により多くの情報を持っていると考え、 $G(w)$ の大きい単語をキーワードとして選択する。

$p(T)$: 話題 T の生起確率

$p(w)$: $\frac{\text{単語 } w \text{ の出現回数}}{\text{単語総数}}$

$p(w|T)$: $\frac{\text{話題 } T \text{ における単語 } w \text{ の出現回数}}{\text{話題 } T \text{ における全単語の出現回数}}$

である。

ここで $p(w)$, $p(w|T)$ の学習のテキストベースとして「毎日新聞 CD-ROM」(1991 年から 1994 年 9 月)を用いた。総記事数約 36 万、総単語約 9600、単語の種類約 35 万である。このデータでは

1. 新情報処理開発機構 (Real world Computing Partnership) の RWC データベースワークショップにより、記事の見出しと本文に自動的に行われた形態素解析の結果データが公開されている。

表 1: 毎日新聞掲載面種別コード

01	1面	05	社説	12	総合	16	科学
02	2面	07	国際	13	家庭	18	芸能
03	3面	08	経済	14	文化	35	スポーツ
04	解説	10	特集	15	読書	41	社会

2. 記事ごとに表 1 に示す掲載面種別コードが付与されており、話題の分類がされている。

となっており、単語の話題に対する条件付き確率の学習に適している。なお本研究では話題の同定には表 1 に示すコードのうち国際、経済、家庭、文化、科学、芸能、スポーツ、社会の 8 つの話題のみを使うことにした。

また $p(T)$ に関しては

1. 新聞記事のデータから求めた値を使う。
2. ニュースの中で話題がどんな頻度で出現するかわからないため、等確率としておく。

が考えられるがここでは 2 とした。さらに今回は選択する単語を名詞に限定した。

3.2 音素長による単語選択

音素長の短い単語はスポッティングにおいて湧き出し誤りを増大させる。このため音素長が 5 以下の単語を除去した。なお音素単位には母音 (a, i, u, e, o) に加え、sh, ts, k などの子音を用い合計で 26 音素である。

以上相互情報量、音素長の 2 つを基準として、キーワードを 3000 個選択した。

4 テキストデータに対する話題分類

3 節で述べた手法に基づき選択したキーワードが適切であるかどうかの評価を行うため、テキストデータに対する話題分類の実験を行った。実験対象データとして毎日新聞 CD-ROM のうちキーワード選択に用いなかった 94 年 10 月から 12 月までの新聞記事を用いた。なお評価にはその中で 3 節で示した 8 つの話題の記事のみを対象とした。また分類は記事の境界は既知で、記事単位で行った。記事中に検出されたキーワードを $w = w_1, w_2, \dots, w_i$ とするとその記事の話題が T_j である確率 $P(T_j|w)$ は

$$\log P(T_j|w) = \log \frac{P(w|T_j)P(T_j)}{P(w)}$$

表 2: テキストベースの話題分類結果

分類結果 話題	国際	経済	家庭	文化	科学	芸能	スポーツ	社会	NK	合計	正解%
国際	1356	52	9	0	1	1	3	46	0	1468	92.4
経済	137	1901	51	0	8	14	6	123	1	2241	84.8
家庭	33	74	505	7	20	16	16	232	2	905	55.8
文化	37	40	22	42	5	33	36	240	0	455	9.2
科学	4	14	5	0	16	0	3	18	1	61	26.2
芸能	30	22	12	13	0	129	42	102	1	351	36.8
スポーツ	125	39	0	0	0	8	2274	127	11	2584	88.0
社会	506	383	161	9	87	49	253	4428	12	5888	75.2
合計									28	13953	76.3

であるが、本報告では

$$\log P(T_j|w) \approx \sum_i \log P(w_i|T)$$

による近似計算を行い、この $P(T_j|w)$ の値が最も大きくなる話題 T_j をその記事の話題とした。テキストデータに対する話題分類の結果を表 2 に示す。表中の NK はキーワードが一つも出現しない記事の数を示している。なおここでは一つの記事が 50 単語以上の記事だけを評価に用いた。

結果をジャンル別に見ると「国際」「経済」「スポーツ」「社会」に関しては高確率で正しく話題の同定ができていることがわかる。一方残りの 4 話題についてはかなり正解率が下がっている。中でも「文化」の話題に関しては正解率がかなり低く「社会」をはじめ、他の話題に誤って話題同定された記事が多く見られる。「文化」の話題の条件付き確率が大きい単語を調べてみると、他の話題の条件付き確率も大きくなっている傾向があり「文化」のみの話題同定に多くの情報を持つ単語が少ないということが正解率の低さにつながっていると考えられる。またすべての話題において「社会」と誤って分類された記事が多い。話題との条件付き確率 $P(w_i|T)$ の学習に用いた新聞記事データでは「政治」「事件, 事故」「行事」「天気」などの記事についても「社会」とタグづけされており、その影響で「社会」の条件付き確率が比較的大きくなっている単語が多いということが推測される。また記事中の単語数に占めるキーワードの割合の平均値は 8.76% であった。

記事中のキーワード数と正解率の関係を調べる

ため、全記事を同じような単語数からなる記事のサブセットを構成し、それぞれのサブセットに対して話題同定実験を行った。一記事あたりの単語数とその記事中の平均キーワード数ならびに正解率の関係を図 2 に示す。一記事あたりの単語数の増加に伴いその記事中の平均キーワード数も増えている。一方正解率では 1 記事中の単語が 80 以上の記事に対してほぼ 70% から 80% と高い割合を示しており、うまく話題の同定ができたといえる。このことからテキストデータに対しては 1 記事あたり 10 個程度のキーワードでもかなりの確率で話題同定ができていることが分かる。

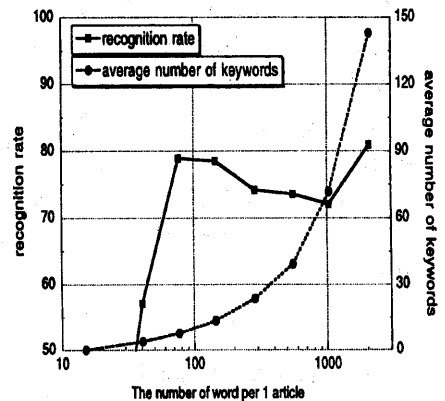


図 2: 一記事中の単語数と平均キーワード数並びに正解率の関係

5 ワードスポットティング

5.1 音韻モデル

スポットティングはHMMを用いて行った。今回用いた音韻モデルは5状態3ループ、混合数は4で、音韻モデルの数は26である。これらの音韻モデルの学習には日本音響学会の研究用連続音声データベースの中の64名分の読み上げ音声(ATR503文)を用い、連結学習を行った。音声分析条件の諸元を表3に示す。

表3: 音声分析条件

標本周波数: 16kHz
窓長: 25ms
シフト: 10ms
パラメータ数: 25次元 (12次メルケプストラム + 12次 Δ メルケプストラム+ Δ パワー)

5.2 音韻スコア

入力音声に対して図3に示す(a)(b)2種類の言語モデルを利用してそれぞれビタビアルゴリズムに基づきフレームごとの尤度を算出し、その結果を基にスポットティングを行う。

まず(a)に示す音素連結モデルを用い、発話先頭

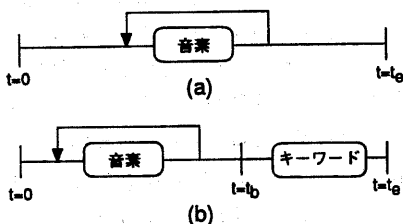


図3: 言語制約

から任意の時刻 t_e までの最適音韻系列の対数尤度 $Ls_{yl}(t_e)$ を計算する。次に(b)に示すモデルを用い、時刻 t_e でキーワードで必ず終了すると仮定したときの発話先頭からの対数尤度を $Lkw(t_e)$ を算出する。ここで(b)の制約におけるキーワードの開始時刻を t_b 、(a)および(b)の制約での認識において時刻 $t = t_1$ から時刻 $t = t_2$ までの対数尤度を

それぞれ $L_s(t_1 : t_2), L_k(t_1 : t_2)$ とすると

$$Ls_{yl}(t) = Ls(0 : t_b) + Ls(t_b : t_e) \quad (2)$$

$$Lkw(t) = Lk(0 : t_b) + Lk(t_b : t_e) \quad (3)$$

である。ここで、キーワードの区間以外では(a)(b)どちらにおいてもほぼ同じ言語制約での認識となるため、

$$Ls(0 : t_b) \simeq Lk(0 : t_b) \quad (4)$$

と近似できると考えられる。また入力音声のキーワードの候補区間 $t = t_b$ から $t = t_e$ までの特徴ベクトル系列を Y とするとこの区間がキーワード w である確率 $P(w|O)$ はベイズの定理にしたがって、

$$P(w|O) = \frac{P(O|w)P(w)}{P(O)}$$

となる。この式の両辺の対数を取り、キーワードの事前確率 $P(w)$ を1とすれば

$$\log P(w|O) = \log P(O|w) - \log P(O)$$

となる。ここで $\log P(O|w)$ が $Lk(t_b : t_e)$ で与えられ、 $\log P(O)$ が $Ls(t_b : t_e)$ で近似できることを考慮しながら、式(2)から(4)を使えば

$$\log P(w|O) = Lkw - Ls_{yl}$$

となる。さらに長さの異なるキーワードに対して同じ閾値でスポットティングの処理を行えるようにするため、キーワード長で正規化して音韻スコア S を

$$S = \frac{Lkw - Ls_{yl}}{\text{キーワード長}}$$

と定義する。本報告ではキーワード長としてキーワードの音素数を用いる。このようにして算出したスコアがある閾値よりも小さい区間にキーワードが含まれるとする。

6 ニュース音声に対する話題同定

今回は評価対象データとして7時のニュース約40分間を用いた。このニュースには話題同定実験に用いる8個の話題のうち「家庭」「芸能」以外の6話題が出現する。その話題の出現割合を図4

社会	スポーツ	科学	経済	文化	国際	その他
43.6	14.1	11.0	7.93	7.05	13.2	

3.08

単位は%

図 4: ニュース音声の話題割合

に示す。なお図中の「その他」は具体的には「天気」「ノイズ」「音楽」などである。また参考までにニュース音声の開始時刻から数分間の間に出現する単語から音素数が6以上となる名詞85個をピックアップしたところ、そのうち47個が3節で選出した3000個のキーワードによってカバーされていた。

話題同定を行う分析区間内でスポッティングされた単語 w_i に対して話題 T_j に対する条件付き確率 $P(w_i|T_j)$ と音韻スコア $S(w_i)$ を用いてその区間の話題らしさを算出する。分析区間が話題 T_j である確からしさを

$$ID(T_j) = \frac{\sum_i (\log P(w_i|T_j) + c * S(w_i))}{\text{キーワード数}} - Bm(T_j)$$

に基づいて話題ごとに算出した。ここで区間長 M は30秒、60秒の2通りとし、それぞれ10秒ずつシフトさせながら計算を行った。区間は合計で233であった。なお右辺第一項の c は音韻スコアの重みで0.5とした。また右辺第二項 $Bm(T_j) (m=1, 2)$ については

$$1 \quad B1(T_j) = \frac{\sum_w \log P(w|T_j)}{\sum_w 1}$$

$$2 \quad B2(T_j) = \frac{\sum_w P(w) * \log P(w|T_j)}{\sum_w P(w)}$$

の2通りを試した。ここで $Bm(T_j)$ を考慮した理由を述べる。3節で話題との関連性が高いキーワードを選択したが、話題同定に有効なキーワードがどの話題にも等しく選択されているとは限らない。実際に調べてみると特に「社会」の話題に対する話題依存性が大きいキーワードが多いことがわかった。このためキーワードの話題に対する

表 4: 話題毎の $Bm(T_j)$

T_j	$B1(T_j)$	$B2(T_j)$
国際	-5.449	-4.556
経済	-5.236	-4.500
家庭	-5.400	-4.792
文化	-5.763	-4.955
科学	-6.733	-5.443
芸能	-5.827	-5.084
スポーツ	-5.795	-5.105
社会	-4.900	-4.322

偏りを正規化する必要がある。ここで話題ごとの $B1(T_j), B2(T_j)$ の値を表4に示す。

話題分析区間の正解話題が話題ごとに求めた話題らしさ ID の中で上位 n 位以内 ($n=1, 2, 3$) になっている割合を調べそれぞれの話題同定の評価基準とした。この評価結果を表5に示す。

表5を見ると話題らしさの第1位のみを正解とした場合は正規化なし ($B(T_j) = 0$) の方がかなり良くなっている。しかし、これは「社会」の話題に対する話題依存性が大きいキーワードが多く、ニュース音声のほとんどの区間で「社会」が1位となってしまう、結果として正解率が上がったもので正しい話題同定ができているとは言い難い。このときの話題らしさ ID の時間変化 ($M = 30$) を図5に示す。また正規化を行う場合は正解ランクに関わらずすべてにおいて $B1(T)$ による正規化を用いた方がより正しい分類ができている。また区間長 M による正解率の差はそれほど見られなかった。

図6は窓長 $M = 30$ として $B1(T)$ による正規化を用いて話題ごとに算出した話題らしさ ID が時間とともに変化する状態と実際のニュースの正解話題ならびにその区間を示す。time はニュース開始からの時刻で単位は秒である。約1250秒から約1430秒までの区間は正解話題が「科学」であるのに対し、求めた ID (科学) では1300秒あたりから上昇し、1330秒から1350秒までは ID の中で1位になっており、約1400秒まで高位置につけている。さらにほぼすべての時刻で ID (スポーツ) は低い値をとり続けているが、正解がスポーツとなる区間に入ると1470秒付近で急激に上昇しており、スポーツである話題らしさがうまく算出できている

表 5: ニュース音声の話題分類結果

RANK	正規化なし (B(T)=0)		B1(T)		B2(T)	
	M = 30	M = 60	M = 30	M = 60	M = 30	M = 60
第 1 位まで	50.3	50.3	12.7	12.0	14.7	14.7
第 2 位まで	57.4	59.2	24.4	23.6	47.7	50.3
第 3 位まで	60.4	59.2	50.1	55.4	66.5	61.8

単位は [%]

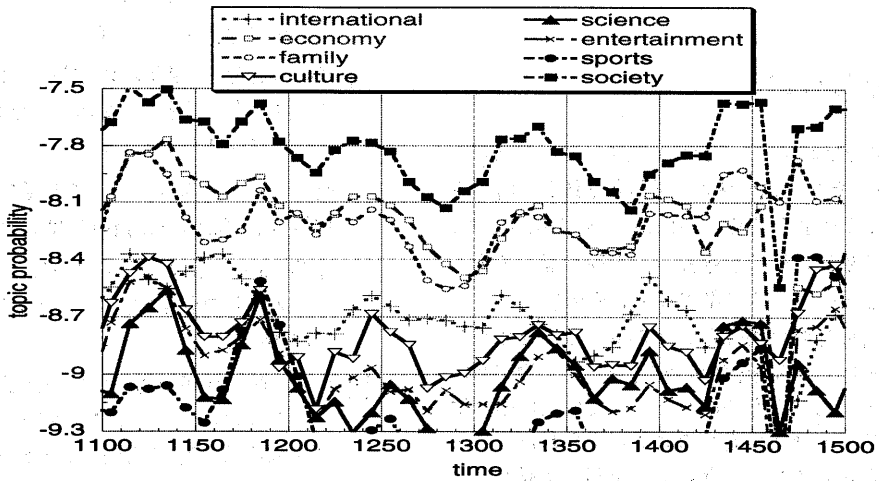


図 5: ニュース音声の話題らしさ ($M = 30, B(T) = 0$ (正規化なし))

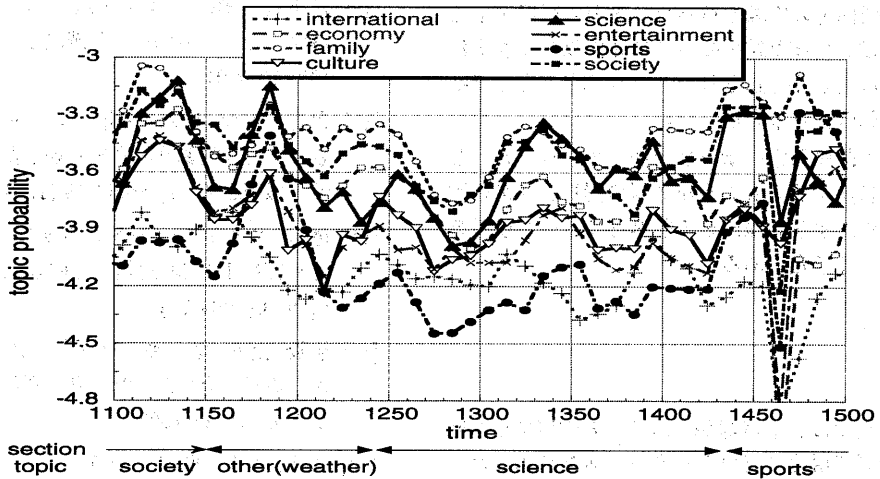


図 6: ニュース音声の話題らしさ ($M = 30, B1(T)$)

と言える。また約1460秒ですべての話題らしさが局所的に落ち込んでいる。この時刻では相撲の中継VTRが放映されており、アナウンサーの早口の実況に加え、観客の歓声等のノイズが多く入っており窓長内にスポッティングされた単語が極端に少ないということがわかった。抽出単語数が平均約280(窓長 $M = 30$) であるのに対し、ここでは10程度の単語しか抽出されていなかった。つまり話題同定を行うだけの単語数が得られなかったということが理由として考えられる。

7 おわりに

本研究ではニュース音声に対しキーワードスポッティングを行い、抽出された単語と話題との関連性を利用して話題を同定する手法について検討した。

まず話題との相互情報量、音素長を選出基準として、話題の同定に貢献する単語を選出した。

次に選択した単語の妥当性を確かめるためテキストベースでの話題分類実験を行った。話題ごとにばらつきはあったもののトータルでは正解率は約76%を示しほぼ期待通りの結果が得られた。キーワードが正しく抽出できれば、単語と話題との関連性の情報を利用した話題の同定が有効であることを確認した。

さらにニュース音声に対してわきだし誤りも含めた話題同定の実験を行った。正解の話題が話題同定実験を行った結果の中で第3位以内に含まれていれば正しく分類されたという基準のもとでは約66.5%の正解率を示した。

今後は話題同定の新たな手法、キーワードとして複合名詞が扱えるような単語の選択、音声の雑音区間に対する話題同定の後処理について検討を進めていく予定である。

謝辞

本研究では、音声認識システムの学習に日本音響学会研究用連続音声データベースを使用した。

参考文献

- [1] 河原, 他: “ヒューリスティックな言語モデルを用いた会話音声の単語スポッティング”, 信学論, J78-DII, 7, pp.1013-1020 (1995-6).
- [2] 桜井, 有木: “キーワードスポッティングによるニュース音声の索引付けと分類”, 信学技報, SP96-66, pp.37-44 (1996-11).
- [3] 横井, 河原他: “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, 95-SLP-6, pp.15-20.1995.
- [4] 横井, 河原他: “単語の共起情報を用いたニュース朗読音声の話題同定機構”, 信学技法, SP96-105, pp.71-77 (1997-01).
- [5] 今井, 他: “放送ニュースの話題抽出モデル”, 信学技法, SP97-28, pp.75-82 (1997-06).
- [6] 今井, 他: “話題混合モデルによる日本語ニュースからの話題抽出”, 音講論, 3-1-7, pp.99-100(1997-9)
- [7] 中沢, 他: “話題要約のための話題のセグメンテーションとシソーラスを用いた話題グループ生成の検討”, 音講論, 3-1-6, pp.97-98(1997-9)
- [8] 鈴木, 福本, 関口: “ニュース文の話題セグメンテーション”, 音講論, 2-6-10, pp.63-64(1997-9)
- [9] 五十嵐, 他: “自由会話音声における話題ラベリング手法”, 音講論, 2-Q-11, pp.137-138.(1997-9)
- [10] 大附, 松岡, 松永, 古井: “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.67-73 (1997-06).
- [11] 亀山, 中里他: “ワードスポッティングに基づく意図抽出”, 信学技法, SP94-64, NCL94-33, pp.9-16(1994-12).
- [12] 岩橋, 山下, 他: “キーワードスポッティングにおけるピッチパタンの利用”, 信学技報, SP96-106, pp.79-85(1997).
- [13] 恒川, 山下, 他: “キーワードスポッティングにおける音韻スコアの検討”, 音講論, 2-1-16, pp79-80(1997-9).