音声言語情報処理 6－3
（1995. 5. 25）

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

横井 謙太朗　河原 達也　堂下 修司
京都大学 工学部 情報工学教室
〒606-01 京都市 左京区 吉田本町

あらまし

大量の音声メディアの情報に効率的にアクセスするための処理として、音声データに対する話題のインデキシングが有効であると考えられる。本研究では、キーワードのスポッティングによる話題同定と、そのために必要なキーワード集合の選定について検討し、ニュース音声の話題同定を試みた。大語彙ワードスポッティングのために、単語辞書を、話題同定に貢献するキーワード辞書と入力音声を近似する基礎単語辞書とに分割した。キーワード選択の基準として、話題との相互情報量、音素列類似度、音素列長の3つの基準を用いた。そして得られた単語辞書によって、原稿テキストと音声の両方に対して話題同定実験を行なった。

和文キーワード　音声メディア処理, 話題同定, キーワード選択, 大語彙ワードスポッティング

Topic Identification of News Speech based on Keyword Spotting

Kentaro Yokoi　Tatsuya Kawahara　Shuji Doshita
Department of Information Science, Kyoto University
Sakyo-ku, Kyoto 606-01, Japan
e-mail: yokoi@kuis.kyoto-u.ac.jp

Abstract

For efficient access to desired information from a large quantity of speech media, it is useful to index it with its topics. In this report, we study topic identification with keyword spotting and keyword selection, and apply them to news speech. For large vocabulary word spotting, we prepare two dictionaries. One is for keywords that contribute to the topic identification, and the other is for basic words that are used to approximate the input speech. For keyword selection, we adopt three criteria: mutual information with topics, similarity of the phone sequence, and its length. With the obtained dictionary sets, we have performed topic identification experiments both on news texts and speech.

英文 keywords　speech media processing, topic identification, keyword selection, large vocabulary word spotting

1 はじめに

社会の情報化がすすみ、我々は様々なメディアの膨大な情報を享受できるようになった一方、データの膨大さゆえにかえって自らの望む情報が捜し出せない状況にある。特に音声メディアにおいては、テキスト環境のような検索ツール [1] に恵まれておらず、検索手段は極めて貧弱である。そのため、音声メディアにおいて自分の望む情報に効率良くアクセスするための処理として、音声データの話題の同定・インデキシングを行なう技術が求められる [2][3]。本研究ではそれを実現するためのキーワード集合を選択し、キーワードスポッティングの手法によってその評価を行なう。

対象データとしてはニュース音声を選んだ。これは、第一に、今後その情報量がますます増えることが予想されること、第二に、発話内容にインデキシングするに充分な話題依存性があると考えられることによる。

2章でキーワードによる話題同定の方法とそのキーワードの選択方法について述べる。3章で、ヒューリスティックな言語モデルを用いたワードスポッティングの手法について述べる。4章では、最初にキーワード辞書の適切性を評価するためにテキストベースで話題同定の予備実験を行ない、さらに、実際の音声としてニュース原稿の朗読音声に対する同定実験を行なう。

2 キーワードによる話題の同定

2.1 キーワードによる話題の活性化

キーワードスポッティングにより発話から話題を同定する試みとしては、BBN の P.Jeanrenaud らの研究 [4][5] がある。彼らは、話題を「フレーズまたは節の集合」として文法的構造でとらえようとした。このアプローチは極めて限定されたタスクについては効果的であるが、より一般的な意味で言う「話題」を文法的構造で記述し尽くすのは非現実的である。そこで我々は、統計的手法による、一般性のあるモデルを考える。

2.1.1 話題の活性化機構

ニュース音声における各々の文は、その時点のニューストピック（＝話題）に対する解説文であると考えられる。そして、人間がそのニュース音声を聞いた時に容易にその発話話題を判断することができるのは、発話内からいくつかのキーワードを抽出し、そのキーワードのもつ話題依存性の情報をもとに話題内容を正しく同定しているためだと考えられる。

このキーワードからの話題同定をシミュレートする機構として、図1のようなモデルを考える。単語 i は、「単語 i が文内に現れた時にその文がある話題 n であると考えることの有意性の情報」（以後これを「活性度」と呼ぶ）を持っているとする。そして単語 i が文内に現れる毎にこの活性度を各話題ごとに加算していき、最終的に最も活性度が高くなった話題をその発話の話題とみなす。

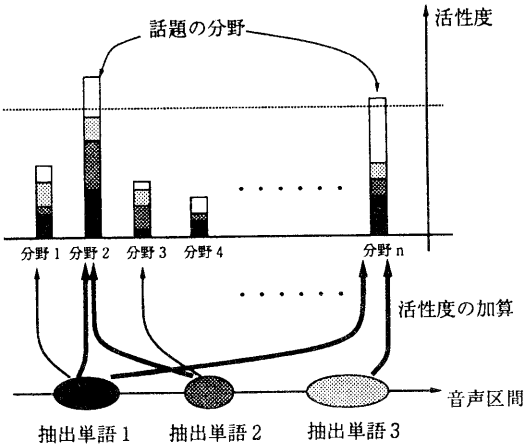

図 1: キーワードから話題の同定を行なう機構

また、音声認識においては湧きだし誤りは不可避のものであり、これに対する対策が必要になる。そこで、キーワードの出現の偏りを用いる方法が考えられる。キーワードは一般に、ある特定の話題に頻繁に出現するという性質を持つ。一方、湧きだし誤りの出現には分野の偏りがないと考えられる。つまり、認識された単語の分野の偏りを見ることで、湧きだし誤りの影響を取り除くことができると考えられる。さらに、音響スコアを反映させることで、より正確な話題同定ができると考えられる。

2.1.2 キーワードの活性度の推定

キーワードの持つ活性度は、その単語と話題との結び付きの度合である。つまり、ある分野 n に頻繁に出現する単語 i は、その分野 n に対して高い活性度を持っていると言える。

そこで単語 i の分野 n に対する活性度を「テキストデータベース中の分野 n に、その単語 i が出現する回数」と定義する。

具体的には、ニュース・新聞などに現れるキーワードに関する辞書「知恵蔵 1994」（©朝日新聞社 1993）の本文中の出現回数を活性度として採用する。テキ

ストデーターベースとして「知恵蔵」を選んだ理由としては、第一にこれが我々が処理の対象としているニュース音声に関連の深い大規模テキストであること、第二にこのテキストは10の分野(文化,経済,国際,産業,科学,政治,生活,社会,スポーツ,技術)によって構成されているので、それをそのまま本研究で同定する話題と設定することができること、の二点による。

キーワードの候補集合としては、

1. 日本語かな変換システム(Wnn)の基礎辞書に含まれる単語のうち、活用変化しないもの(名詞・形容動詞語幹・副詞・連体詞・接続詞)
2. 「知恵蔵」の見出し語

を用いた。これらの総数は約14000である。

これらのキーワード候補について、テキストの分野別に活性度を求めた。

2.2 単語集合の分割

スポッティング対象の単語数を増やすことは、単語間の音響的距離を狭めることを意味し、認識誤りを増やす原因となる。それゆえ、一概に辞書の語彙数を増やせばよいというわけではない。

我々がキーワード抽出の手法として用いるワードスポッティングでは、3章で述べるように、スポッティング対象キーワードのモデルの前後にヒューリスティックとして大語彙単語モデルを接続することでスコアを計算している。そのため、スポッティング対象単語部分については語彙数が増えると認識誤りが増える可能性がある一方、前後のヒューリスティック部分については、語彙数が増えるとより正解に近いヒューリスティックが得られるので好ましい。そこで、辞書登録単語を以下のように分割する。

基礎単語辞書 入力音声中の単語列を近似するためのヒューリスティック計算にのみ用い、抽出の対象とはしない。入力音声全体の認識をより正解に近いものにするため、ある程度語彙数が多い方が良い。

キーワード辞書 スポッティングの対象にし、ヒューリスティック計算にも用いる。語彙数が増え過ぎると認識誤り増加の原因となるので、必要最小限に絞る。

これにより、スポッティング対象の単語集合を必要最小限に抑えながら大語彙音声に対応できるようになる。

2.3 キーワードの選択基準

2.1.2節で述べたような活性度計算の結果を基に、辞書登録単語の選択を行なう。

まず基礎単語辞書については、ヒューリスティック計算のための単語であるから、実際の音声に現れる単語を出来るだけ多く網羅する必要がある。つまり、出現頻度が高く、かつどの分野にも頻繁に出現する単語を登録する必要がある。そこで、

- 全分野の総出現回数があるしきい値を上回る。
- 各分野の最低出現回数のしきい値を設定し、そのしきい値を上回る分野数が一定以上ある。

という条件を満たす単語を1000語選び出した。

次にキーワード辞書であるが、これについては以下に述べる3つの観点から選択した。

2.3.1 相互情報量を用いた単語選択

キーワード辞書に含まれる単語は、話題同定の際に活性化情報を提供する単語である。つまりその出現によって、より話題同定に関する情報が得られる単語を選ぶ必要がある。

そこでキーワード辞書の登録単語を選ぶための基準として、情報理論に基づいて単語 W と話題 T との相互情報量を考える[6]。単語Wの出現により、話題Tに関して得られる情報量は、

$$
\begin{aligned}
I(T;W) &= I(W;T) = H(W) - H(W/T) \\
&= \sum P(W) \log \frac{1}{P(W)} \\
&\quad - \sum P(W,T) \log \frac{1}{P(W/T)} \\
&= \sum P(W) \log \frac{1}{P(W)} \\
&\quad - \sum P(W) P(T/W) \log \frac{1}{P(W/T)}
\end{aligned}
$$

$$
\left(\begin{array}{lcl}
P(W) & : & \text{単語の生起確率} \\
P(T/W) & : & \frac{\text{単語}W\text{の分野}T\text{における出現回数}}{\text{単語}W\text{の全分野における出現回数}} \\
P(W/T) & : & \frac{\text{分野}T\text{における単語}W\text{の出現確率}}{\text{分野}T\text{における全単語出現回数}}
\end{array} \right)
$$

と定義できる。ここで、相互情報量 $I(T;W)$ の大きい単語は、より話題同定に関する情報が得られると考えられる。そこでこの相互情報量の大きい単語を上位から選ぶことにする。

表 1: 単語辞書の構成

単語数	相互情報量	音素類似度
基礎辞書	1000	825
キーワード辞書 S	400	376
キーワード辞書 M	600	570
キーワード辞書 L	800	736
キーワード辞書 LL	1000	902

2.3.2 音素記号列の類似度を用いた単語選択

先にも述べたように、抽出単語の大語彙化は単語間の音響的距離を小さくし、音響的に類似した単語が誤って抽出される可能性を大きくする。

そこで、音響的に似ている単語を削除することで認識誤りの可能性を減らし、より安定な話題同定を実現することを考える。音響的な類似度を完全に調べるのは難しいので、ここでは音素記号列の類似度を調べることにする。

具体的には、

- 音素がマッチしない場合、スコア −10
- 音素飛び越しを許せばマッチする場合、1 音素飛び越しでは −4、2 音素飛び越しでは −7

とする。

このような条件のもとで DP マッチングを行ない、スコアを音素列長で正規化することにより類似度を計算した。そして、類似度があるしきい値を上回るものは音素列の短い方を削除することにより、同音単語と類似単語を取り除いた。削除のしきい値を変化させて実験した結果、類似単語の削除のしきい値を 8.5 程度にするのが適当であることが分かった。

2.3.3 音素記号列長を用いた単語選択

音素列の短い単語は湧き出し誤りを生じやすいので、母音音節数 2 以下の単語を削除した。

以上のようにしてキーワード選択を行なった。まず母音数 3 以上の単語の相互情報量を計算し、上位からそれぞれ 400, 600, 800, 1000 語選んだ。さらに音素列の類似度に対してしきい値を設定してキーワードを選択した。結果を表 1に示す。

このようにして基礎単語辞書 (825 語)、キーワード辞書 S(379 語), M(571 語), L(738 語), LL(905 語) を作成した。

3 大語彙ワードスポッティング

音声データからキーワードを抽出するために、ヒューリスティックな言語モデルを用いたワードスポッティングを行なう。

ヒューリスティックワードスポッティング [7][8] においては、入力音声を、スポッティング対象単語・その他の（未知の）単語・ポーズからなる系列と仮定する。このような言語モデルを、スポッティング対象単語のモデルの前後の音声系列を近似するヒューリスティックとして用いる (図 2)。発話全体に対してこのような言語モデルを設定し、その言語的制約を満たし尤度の高い単語を抽出する。

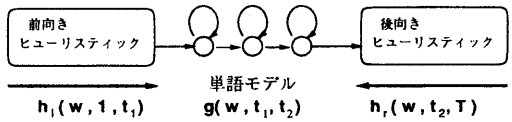

図 2: 言語モデルを用いた単語スポッティング

入力音声中の区間 $t_1 \sim t_2$ に単語 w が存在する仮説の評価値 $f(w,t_1,t_2)$ は、単語部分のスコア $g(w,t_1,t_2)$ と残りの区間 $1 \sim t_1, t_2 \sim T(T$: 入力長) の言語らしさに対するスコア $h_l(w,1,t_1), h_r(w,t_2,T)$ の和として定義する。これらの h をそれぞれ前向きヒューリスティックと後向きヒューリスティックと呼ぶ。

$$f(w,t_1,t_2) = h_l(w,1,t_1) + g(w,t_1,t_2) + h_r(w,t_2,T)$$

このようにして得られる評価値 f の高いものを抽出することでスポッティングが実現される。

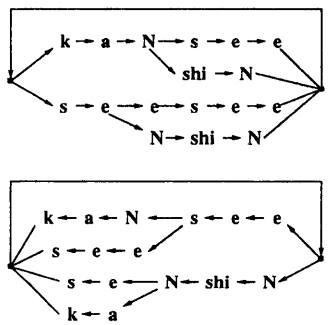

図 3: 単語連接モデル

ヒューリスティックのモデルとしては、単語連接モデルを用いる。これは想定する任意の単語の系列

表 2: ニューステキストに対する話題同定結果

概要	文番号	同定された分野												出現キーワード数			
		辞書 S(379 語)			辞書 M(571 語)			辞書 L(738 語)			辞書 LL(905 語)			S	M	L	LL
		1 位	2 位	3 位	1 位	2 位	3 位	1 位	2 位	3 位	1 位	2 位	3 位	−	−	−	−
フィリピン航空機爆発	1				国際	社会	文化	国際	社会	文化	国際	社会	文化	0	2	2	3
	2	技術	科学		技術	科学		技術	科学		技術	科学		1	1	2	2
	3				国際	社会	文化	国際	社会	文化	国際	社会	文化	0	2	2	3
	4	科学	生活	社会	科学	生活	社会	科学	生活	社会	科学	生活	社会	2	2	2	2
	5													0	0	0	0
	6	科学			科学	技術		科学	技術		科学	技術	国際	1	2	2	4
	7	技術	科学	産業	技術	科学	産業	技術	科学	産業	技術	科学	産業	2	2	2	2
ロシアのチェチェン共和国内戦	8	国際	文化		国際	文化		国際	文化		国際	文化		1	1	1	1
	9													0	0	0	0
	10	科学			科学			科学			科学	技術		1	2	2	3
	11				国際	政治	技術	国際	政治	技術	国際	政治	技術	0	1	1	1
	12				国際	産業	生活	国際	産業	生活	国際	産業	生活	0	1	1	1
	13	国際	文化		国際	文化		国際	文化		国際	文化		1	1	1	1
以下略																	
301 文合計 (juman による総語数 ・・・ 15130)														271	353	455	502

下線部分は、適切な活性化がなされていると考えられる分野を示す

を表現するモデルである (図 3)。これは語彙のみを制限するので、構文を逸脱した文章に対しても頑健である。語彙を制限するために未知語や間投語を含む入力に対する動作は保証されないが、多くの場合類似単語の系列で近似され、音節モデルを併用する場合と性能は変わらない [7]。

4 話題同定実験と考察

4.1 テキストベースによる活性度計算と語彙数に関する検討

まずテキストベースでの試行により、単語辞書の妥当性を評価した。

すべての登録単語について、その単語がテキスト中に出現するかどうかを調べ、出現すればその単語の持つ活性度情報にしたがって各分野に活性度を加算していく。ただし、部分文字列に誤ってマッチングしているものも含まれている。

確実な話題同定をするには登録語彙を必要最低限のものに絞る必要があるので、適切な語彙を調べるためにキーワード辞書の語彙数を様々に変えて実験を行なっている。

また、具体的な話題分野の同定も行なっている。話題分野の判定基準としては、

- 活性度が 90 を超えている
- 活性度 1 位の分野の活性度の $\frac{1}{2}$ 以上

のいずれかを満たすものとして、上位 3 分野までを選んだ。

テキストは、NHK「7 時のニュース」の原稿 301 文を用いた。表 2に同定結果を示す。表中、太字で示されている分野名は、人間が原稿を見ておおむね妥当と判断される活性度を出している分野を示すものである。

この結果から、キーワードがいくつか出現している文章についてはおおむね妥当な文やが活性化されていることがわかる。一方、あまり良い結果が出ていない文章については、人間が原稿を見ても話題の判断が難しい文章が多かった。

語彙数に関しては、出現単語数で見ると、辞書 L 程度の語彙が必要であると考えられる。語彙数を最小限に絞る必要性も考え合わせ、最終的に辞書 L が妥当な語彙数であるという結論に至った。

4.2 ニュース音声に対する話題同定実験

2 章で述べた単語辞書と 3 章で述べた大語彙ワードスポッティングの技術を統合して、朗読音声に対する話題同定の実験を行なった。この実験において、各分野の活性度を求めるのに抽出単語の活性度と抽出単語の音響スコアの積を加算しており、活性度と音響スコアの統合を行なっている。

音声データとしては、NHK「7 時のニュース」の原稿をアナウンサーでない 1 名の男性話者が朗読したサンプル 33 文を用いた。これは前節の実験で用いた原稿の部分集合である。話者は音素モデル学習用とは異なる。辞書は、基礎辞書 825 語、キーワード辞書 L 738 語を使用した。抽出のしきい値は最適ヒューリスティックスコアの 1.008 倍まで、1 単語につき抽出候補は 2 候補までとした。表 3に結果を示す。

表 3: ニュース音声に対する話題同定の結果

概要	正解数	辞書内数	全抽出数	同定分野
移植法案	0	5	19	技術 / 経済
携帯電話	2	4	30	技術 / 文化 / 科学
円高	2	3	34	技術 / 科学 / 経済
新進党	3	4	18	政治 / 国際
宇宙開発	3	3	40	技術 / 科学 / 経済
ボクシング	4	4	29	技術 / 科学 / 社会
株式市場	2	4	19	経済 / 技術 / 科学
国会	2	2	16	政治 / 技術 / 科学
以下略				
全 33 文	66	128	964	

辞書内の単語の約半数が抽出されており、5 割強の文章に対して人間が見て判断した分野と同じ分野が 1 位または 2 位に同定された。しかし、認識誤りが多いために、誤って認識された単語の活性度が正しく認識された単語の活性度を覆い隠してしまって、適切な話題同定が出来ていない文も多い。

誤り単語を見ると、やはり音響的に似た単語が多く抽出されていた。これは、音響モデルの改善と共に、音素間の認識誤り率などを加味した類似度尺度が必要ではないかと考えられる。さらに、単語・音節連接モデル [7] を用い、助詞なども含めたより高精度なヒューリスティックを構築することが考えられる。

5 おわりに

本研究では、大語彙ワードスポッティングによる話題同定を行なう手法と、キーワード選択の基準の検討を行なった。

まず、テキストデータベースから、キーワードの分野別活性度の統計データを作成した。また、辞書登録単語を基礎単語辞書、キーワード辞書に分割した。キーワードの選択基準としては、相互情報量を用いた情報理論からのアプローチと、音素記号列上の類似単語の削除、音素記号列長の短い単語の削除、という音素記号からのアプローチを提案した。

単語辞書の予備的評価のため、テキストベースのシミュレーションを行なった。これにより、キーワードを確実に抽出できるならば妥当な話題同定が行なえること、そして語彙数としては 800 語程度必要であることが分かった。

以上の検討をもとに、ニュース原稿の朗読音声に対する話題同定システムを作成し、同定実験を行なった。今後は、音素モデルと言語モデルの高精度化を図ると共に、キーワード選択の検討をさらに進めていく予定である。

参考文献

[1] 藤澤浩道, 絹川博之. 情報検索における自然言語処理. 情報処理 Vol.34, No.10, pp.1259-1265.

[2] 伊藤慶明, 木山次郎, 岡隆一. スポッティングに基づく TV ニュース番組の音声対話理解と音声検索. 音講論, 3-P-22 (1995-3).

[3] 杉山雅英. 音声特徴を key とする音声データベース検索の検討. 音講論, 2-8-11 (1994-10).

[4] P.Jeanrenaud, M.Siu, J.R.Rohlicek, M.Meteer, H.Gish. "Spotting Events in Continuous Speech". Proc. IEEE ICASSP'94, pp.381-384, 1994.

[5] J.McDonough, K.Ng, P.Jeanrenaud, H.Gish, J.R.Rohlicek. "Approaches to Topic Identification on the Switchboard Corpus". Proc. IEEE ICASSP'94, pp.385-388, 1994.

[6] John McDonough, Herbert Gish. "Issues in Topic Identification on the Switchboard Corpus". Proc. ICSLP 94, vol.4, S-36-10, 1994.

[7] 河原達也, 宗続敏彦, 三木清一, 堂下修司. 会話音声中の単語スポッティングのための言語モデルの検討. 信学技報 SP94-28, 1994-06.

[8] 河原達也, 北岡教英, 堂下修司. フレーズスポッティングに基づく頑健な音声理解. 情処学会研究会報告,SLP94-4-6,1994.