

論文 / 著書情報
Article / Book Information

論題(和文)	ニュース音声を対象とした連続音声認識に基づく話題抽出
Title(English)	Topic Extraction based on Continuous Speech Recognition in Broadcast-News Speech
著者(和文)	大附克年, 松岡達雄, 松永昭一, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1997年秋季講演論文集, Vol. , No. 3-1-8, pp. 101-102
Citation(English)	, Vol. , No. 3-1-8, pp. 101-102
発行日 / Pub. date	1997, 9

3-1-8 ニュース音声を対象とした連続音声認識に基づく話題抽出*

◎大附克年¹ 松岡達雄² 松永昭一¹ 古井貞熙³

(¹NTTヒューマンインタフェース研究所 ²NTTマルチメディアビジネス開発部 ³東工大)

1. はじめに

近年、文書(テキスト)だけでなく、画像・動画・音声・音楽などをデジタルデータとして大量に蓄積することが可能となってきている。莫大な量のデータが存在する場合にそれらを効率よく利用するためには、データベースとして整理されていることが必要である。つまり、データを情報として効率よく利用するためには、データの分類や索引付けなどが不可欠である。我々は、音声データの分類や検索のキーとして用いることができるような重要語(話題語: Topic word)の集合を音声から抽出することを検討している[1][8]。本稿では、音声データの内容を表す話題語を音声データ中に出現する単語と話題語との関連度に基づいて抽出する話題抽出法を検討したので、その結果について報告する。

2. 話題抽出

2.1 話題抽出モデル

従来行われてきたテキストやニュース音声のいくつかの分野や話題への分類は、キーワードの各分野(話題)への関連度に基づいて行われる[2-4]。各キーワードは、それぞれの分野に対する関連度の値をもっている。分類の際には、記事中に出現したすべてのキーワードの各分野に対する関連度を合計し、もともと関連度の高い分野へとその記事が分類される。

本研究では、複数の話題語を抽出するために、分類における分野を話題語に置き換えることにより、文に対して関連度の高い話題語の集合を抽出する。文中の単語と話題語とは別の集合になるため、文中に出現しない単語を話題語として抽出することも可能である。

関連度の計算は、単語および話題語の出現頻度と共起頻度に基づいて行う。今回は、確率変数間の依存性の度合を表す尺度として定義される相互情報量および χ^2 検定における χ^2 値[5]に基づいて関連度を計算した。文中の単語 w_i と話題語 t_j との関連度は、相互情報量と χ^2 値に基づく場合にそれぞれ式(1)、(2)のように計算する。

$$I(w_i; t_j) = \log \frac{P(w_i, t_j)}{P(w_i)P(t_j)} \quad (1)$$

$$\chi_{ij}^2 = \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \quad (2)$$

但し、

$$F_{ij} = \frac{\sum_{l=1}^M f_{il} \cdot \sum_{k=1}^N f_{kj}}{\sum_{k=1}^N \sum_{l=1}^M f_{kl}} \quad (3)$$

N : 単語の種類数
 M : 話題語の種類数
 f_{ij} : 話題語 t_j に対する単語 w_i の頻度
 F_{ij} : 話題語 t_j に対する単語 w_i の理論度数

相互情報量の計算において、学習データ中に単語 w_i と話題語 t_j の共起が観測されない場合、 $P(w_i, t_j) = 0$ となり

関連度の合計を求める際に問題が生じる。そこで、共起が観測されなかった場合には何も情報が得られなかったものとして、式(4)により相互情報量に基づく関連度を計算した。

$$I'(w_i; t_j) = \begin{cases} I(w_i; t_j), & \text{if } (P(w_i, t_j) \neq 0) \\ 0, & \text{if } (P(w_i, t_j) = 0) \end{cases} \quad (4)$$

また、相互情報量は、条件付き確率に基づく尺度であり、単語 w_i と話題語 t_j の出現頻度が考慮されていないため、式(5)のように共起確率で重み付けを行った相互情報量についても検討した。

$$I''(w_i; t_j) = P(w_i, t_j) \cdot \log \frac{P(w_i, t_j)}{P(w_i)P(t_j)} \quad (5)$$

χ^2 値における理論度数とは、全ての話題語に対して等確率でその単語が出現した場合の出現頻度である。実際の出現頻度と理論度数とのずれが大きければ、その単語はその話題語に対して偏って出現していることになる。しかし、式(2)では出現頻度が理論度数より小さい場合にも大きい場合と同じ正の値となってしまうため、 $f_{ij} - F_{ij}$ の符号を考慮して式(6)のように χ^2 値に基づく関連度を計算した。

$$\chi_{ij}'^2 = \frac{(f_{ij} - F_{ij}) \cdot |f_{ij} - F_{ij}|}{F_{ij}} \quad (6)$$

2.2 学習データ

話題抽出モデル(単語と話題語の関連度)の学習には、4年9ヶ月分の新聞記事(約90万記事)の本文および見出しを用いた。形態素解析により本文および見出しを単語(形態素)ごとに区切った。頻度の少ない単語と助詞や記号などの単語は関連度の計算には用いなかった。見出し単語(話題語)の種類数は約7万語となった。

3. 評価データ

3.1 音声データ

話題抽出評価用の音声データとして、一発声ごとに切り出した放送ニュース音声の記事ごとにまとめた142発話からなる29記事セットを用いた。1記事は2~14発話、平均5発話からなる。話者は、番組のメインのアナウンサー8名、記者などのそれ以外の話者が7名の15名ですべて男性話者である。評価用音声データには、言い直しなどの自然発話現象および背景騒音・音楽を含むものもある。

3.2 話題データ

話題抽出評価用の音声データを書き起こし、各記事ごとの書き起こしテキストに対してその記事の内容を表すキーワード(話題語)を「最低5個、平均10個程度」という条件で付与した。話題語の付与は3人の被験者がそれぞれ行った。3人の被験者が付与した話題語から、3人が共通に付与した話題語のセット(AND)および3人が付与した話題語をすべて含むセット(OR)の2種類の話題語の

* Topic Extraction based on Continuous Speech Recognition in Broadcast-News Speech.

By Katsutoshi Ohtsuki¹, Tatsuo Matsuoka², Shoichi Matsunaga¹ and Sadaoki Furui³

(¹NTT Human Interface Laboratories, ²NTT Multimedia Business Department, ³Tokyo Institute of Technology)

セットを作った。評価には、付与された話題を形態素解析したものを用いた。話題の形態素数は一記事あたりANDセットで10.4、ORセットで35.7となった。

4. 実験

4.1 大語彙連続音声認識

放送ニュース音声の大語彙連続音声認識は、音素HMMによる音響モデルと統計的n-gram言語モデルとを用いて行う[6]。音響モデルは、tree-based clustering[7]に基づく前後の音素環境に依存するtriphoneモデルであり、53名話者による13270発話(約20時間)を用いて学習した。言語モデルは、5年分のニュース原稿テキスト(約50万文)から、統計的n-gramモデルを学習した。認識システムの語彙は2万語とした。単語bigram言語モデルおよび単語trigram言語モデルをそれぞれ用いた場合の評価用データに対するパープレキシティおよび音声認識結果を表1に示す。

4.2 話題抽出実験

ニュース音声の書き起こしテキストから抽出した話題をORセットに対して評価した結果を図1に示す。関連度の最も高い話題語を抽出したものが左上のプロットであり、関連度上位50単語までをプロットしてある。ここで、再現率(Recall)とは、正解話題語のうち実際に抽出された割合であり、適合率(Precision)とは、実際に抽出された話題語全体に占める正解話題語の割合である。再現率と適合率は、それぞれ式(7)、(8)のように計算される。

$$Recall = \frac{C}{T} \cdot 100 \quad (7)$$

$$Precision = \frac{C}{H} \cdot 100 \quad (8)$$

但し、

- C: 正しく抽出された話題語数(Correct)
- T: 正解の話題語数(True)
- H: 抽出された話題語数(Hypothesized)

一般に再現率と適合率とはトレードオフの関係にある。情報検索要求の種類にもよるが、情報検索システムは高い再現率と高い適合率を示すことが望ましい。

図1をみると、関連度の尺度として、 χ^2 値に基づく場合(χ^2)の方が、相互情報量に基づく場合(MI)に比べ、高い性能を示している。これは前述したように相互情報量には出現頻度(確率)が考慮されていないためだと考えられる。出現確率により重みを付けた相互情報量(wMI)を用いると、相互情報量の場合に比べて高い性能を示すが、 χ^2 値には及ばない。wMIの抽出結果をみると、一般的な話題語が多く抽出されており、それらが被験者の抽出した話題と整合せずに再現率が低くなっていると考えられる。

次に、trigram言語モデルによる音声認識結果に対して、 χ^2 値に基づく関連度を用いて話題抽出を行った結果(tg)を図2に示す。比較のために書き起こしテキストに対する結果(text)も示す。音声認識結果からの話題抽出では、認識誤りが含まれるために話題抽出性能に劣化がみられる。話題語を関連度の高いものから5単語抽出した場合(グラフのプロットで左上から5番目まで)、ORセットに対する適合率は書き起こしテキストで約86%、音声

表1: 話題抽出用音声データに対する音声認識結果

Language model	Test-set Perplexity	Word error rate
bigram	98	28.2%
trigram	56	24.6%

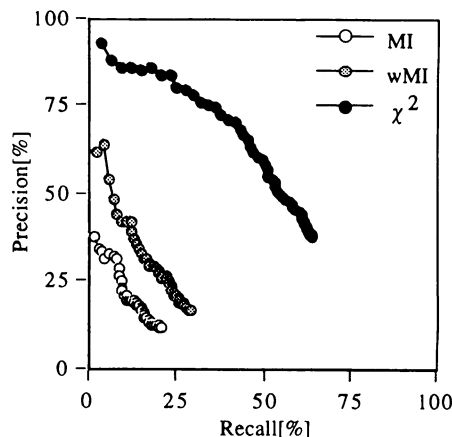


図1: 書き起こしテキストに対する話題抽出結果 (OR)

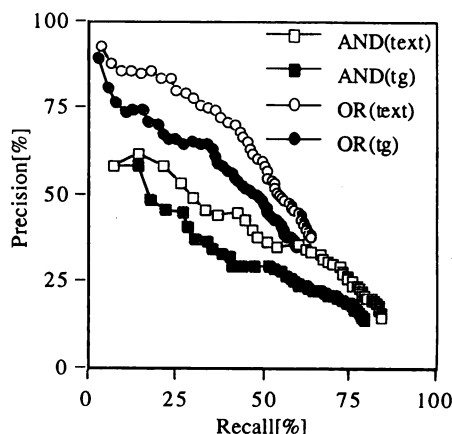


図2: 音声認識結果に対する話題抽出結果 (χ^2 値)

認識結果で約75%となった。

5. まとめ

本稿では、大語彙連続音声認識に基づくニュース音声からの話題抽出について報告した。 χ^2 値に基づく関連度を用いた場合に最も高い精度が得られた。話題語を5単語抽出した際に、誤りを含む音声認識結果に対しても、約75%の適合率が得られた。しかし、音声認識誤りによる性能の劣化がみられるため、音声認識誤りに対して頑健な話題抽出モデルを検討する必要がある。

謝辞

ニュース原稿とニュース音声を提供していただいた日本放送協会に感謝します。新聞記事テキストデータの使用を許諾していただいた日本経済新聞社に感謝します。形態素解析ツールを提供していただいたNTTヒューマンインタフェース研究所映像処理研究部の田中一男主幹研究員に感謝します。

参考文献

- [1] 大附他, 信学技報 SP97-27.
- [2] 横井他, 情処研報, SLP6-3, 1995.
- [3] 櫻井他, 信学技報 SP96-66.
- [4] Rose et al., Proc. ICASSP-91, pp.317-320.
- [5] 長尾他, 情報処理 Vol. 17, No. 2, 1976.
- [6] 田口他, 音講論, 2-6-11, 1997-3.
- [7] Young et al., Proc. ARPA Human Language Technology Workshop, pp.307-312, 1994.
- [8] 今井他, 信学技報, SP97-28.