# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | ニュース音声からの話題抽出法の検討 |
| Title(English) | |
| 著者(和文) | 岩崎淳, 古井貞熙 |
| Authors(English) | SADAOKI FURUI |
| 出典(和文) | 日本音響学会 1998年秋季講演論文集, Vol. , No. 1-1-14, pp. 27-28 |
| Citation(English) | , Vol. , No. 1-1-14, pp. 27-28 |
| 発行日 / Pub. date | 1998, 9 |

# 1—1—14 ニュース音声からの話題抽出法の検討 *

◎岩崎 淳　古井 貞熙 (東工大)

## 1. はじめに

音声からその話題を表す語を自動的に抽出することができれば，放送ニュースのデータベース化などに有用であると期待される．

我々はこれまで，ディクテーション結果としての単語のセットから，単語の相対出現頻度に基づいて話題語を直接選択する方法を提案した [1][2]．人がマニュアルで付与した話題語の集合を正解とし，放送ニュース音声のディクテーション結果から自動的に5つの話題語を選択したとき，その 82.8% は正解であることが確認されている．本稿では，実験の際に用いた種々の尺度の有効性を比較し，その背景について考察した．

## 2. 話題抽出

### 2.1 話題抽出尺度

ニュースの話題を表す単語 (話題語と呼ぶ) は，ほとんどが名詞であるので，音声のディクテーション結果から名詞のみを抽出し，その中からそのニュースの話題語を抽出する方法を検討した．抽出する尺度としては，従来情報検索のための単語重みづけ法として検討された種々の式の中から，以下の式を選んで検討した [5]．

$$I_i = g_i \cdot \log \frac{G_A}{G_i} \quad (1)$$

$$I_i = \frac{g_i}{\log G_i} \quad (2)$$

$$I_i = \frac{g_i}{\log (g_A \cdot G_i)} \quad (3)$$

$$I_i = \frac{g_i}{g_A \cdot G_i} \quad (4)$$

$$I_i = \frac{g_i^2}{g_A \cdot G_i} \quad (5)$$

$$I_i = \hat{g}_i - \hat{G}_i \quad (6)$$

$$I_i = \frac{\hat{g}_i - \hat{G}_i}{\sqrt{\hat{G}_i}} \quad (7)$$

ただし，

$w_i$: ニュース音声中の名詞 ($i = 1, 2, \cdots, N$)

$N$: 語彙サイズ (名詞のみ)

$g_i$: (注目している特定の) ニュース音声中の名詞 $w_i$ の頻度

$G_i$: 全データベース中の名詞 $w_i$ の頻度

$G_A = \sum_i G_i$: 全データベース中の総名詞数

$g_A = \sum_i g_i$: (特定の) ニュース音声中の総名詞数

$\hat{g}_i$: (特定の) ニュース音声中の $w_i$ の相対頻度 ($= g_i/g_A$)

$\hat{G}_i$: 全データベース中の $w_i$ の相対頻度 ($= G_i/G_A$)

$G_i$, $G_A$ の計算には，ニュース音声とは独立の，日経新聞の約5年間の記事 (約 9740 万語) [4] を用いた．

### 2.2 評価実験

各ニュース音声に対し，その中のすべての名詞に関して前述の尺度によるスコアを計算し，この値が大きい単語を選んだ．実験では，注目している各ニュースから話題語を選択するので，全データベース中の総単語数 $G_A$，特定のニュース音声中の総単語数 $g_A$ は定数になる．

評価用音声データには，男性 15 名が発声した 29 のニュース音声 (142 文) を用いた．各ニュースに対して 3 人の被験者に話題語を付与してもらい，3 人の内の少なくとも一人につけられた話題語を形態素解析したものを正解として，話題抽出の評価を行なった [3]．比較のために，音声認識 (ディクテーション) 結果ではなく，正しいテキストが与えられた場合の実験も行った．

話題語抽出性能の評価には，情報検索の分野で用いられている次の尺度を用いた．

$$\text{Recall} = \frac{C}{T} \cdot 100(\%) \quad (8)$$

$$\text{Precision} = \frac{C}{H} \cdot 100(\%) \quad (9)$$

ただし，

$C$: 正しく抽出された話題語の数

$T$: 正解の話題語の総数

$H$: 抽出された話題語の総数

### 2.3 各尺度の比較と検討

各ニュース音声から話題語を 5, 10, 15 単語抽出した際の Precision を表 1 に示す．

この結果から，式 (1) から (7) では，式 (1)(2)(3)(6) と式 (4)(5)(7) がそれぞれまとまった傾向を示し，前者，特に式 (1) が最も優れていることがわかる．式 (1) の対数項 $\log G_A/G_i$ は，単語 $w_i$ が有する情報量に相当するから，この式には，対象としているニュース音声中に含まれる単語 $w_i$ の総情報量としての意味がある．さらにこの式には，$G_i$ が正規化され，学習用データベースの絶対的な大きさに依存しないというメリットもある．図 1 に，式 (1) と式 (5) による Precision と Recall の関係を示す．

*Topic extraction from broadcast-news speech.

By Atsushi Iwasaki and Sadaoki Furui (Tokyo Institute of Technology)

表1の全体的な結果から，単語 $w_i$ の全データベース中での出現頻度 $G_i$ に関して，正規化した方がよく，かつ $G_i$ の影響は大き過ぎないことが必要であることがわかる．式(2)(3)では，$G_i$ は正規化されていないが log をとっているため，$G_i$ の重みが，式(4)(5)に比べて小さく，相対的に記事中の出現頻度 $g_i$ の重みが大きい．これらのタイプでは，比較的良い結果が得られる．正規化されている式(1)(6)(7)の内，式(7)が悪いのは，相対的に $\hat{G}_i$ の重みがききすぎるためであろう．

表1には，情報検索の分野でよく用いられる次のような $tf \cdot idf$ 法[6]による結果も加えてある．

$$I_i = g_i \cdot \log \frac{M}{F_i} \qquad (10)$$

ただし，

$M$ : 全データベース中の記事数

$F_i$ : 名詞 $w_i$ が全データベース中で出現している記事数

実験結果から，$tf \cdot idf$ 法では良い結果が得られないことがわかる．これは $\log M/F_i$ の重みがききすぎるためであると考えられる．

また，正しいテキストが与えられた場合と音声認識結果を用いる場合を，式(1)(2)(3)(6)による結果で比較すると，音声認識の場合に，抽出された話題語中の誤りが約1.5倍に増えることがわかる．

表 1. 各尺度の性能比較 (Precision[%])

| 抽出数 | 5 | | 10 | | 15 | |
|---|---|---|---|---|---|---|
| 式 | text | speech | text | speech | text | speech |
| (1) | 88.3 | 82.8 | 82.1 | 73.4 | 79.5 | 66.2 |
| (3) | 85.5 | 77.9 | 78.6 | 70.0 | 78.2 | 64.1 |
| (6) | 84.8 | 77.2 | 79.0 | 70.3 | 78.4 | 65.3 |
| (2) | 82.1 | 76.6 | 76.9 | 68.6 | 75.9 | 65.5 |
| (5) | 65.5 | 55.9 | 67.9 | 59.0 | 68.3 | 58.4 |
| (7) | 69.6 | 55.9 | 70.4 | 59.0 | 70.7 | 58.4 |
| (4) | 57.2 | 49.0 | 63.8 | 52.4 | 64.6 | 53.8 |
| (10) | 43.4 | 29.7 | 45.2 | 34.1 | 46.0 | 34.0 |

### 2.4 ニュース原稿を用いた学習

学習に5年分のニュース原稿を用いて，同様の実験を行なった．尺度には，上の実験で一番結果の良かった式(1)を用いた．比較のため日経新聞データもニュース原稿の量に合わせて，1030万語程の記事(1994年6月13日～12月31日)に絞った．表2にその結果を示す．学習データの違いは，話題抽出性能の差にさほど影響を与えないことがわかる．

表 2. 日経新聞とニュース原稿による学習の比較 (Precision[%])

| 抽出数 | 5 | | 10 | | 15 | |
|---|---|---|---|---|---|---|
| 学習データ | text | speech | text | speech | text | speech |
| 日経新聞 | 89.0 | 82.1 | 82.4 | 73.8 | 77.9 | 66.4 |
| ニュース | 88.3 | 84.8 | 83.1 | 70.0 | 79.3 | 66.0 |

## 3. まとめ

本稿では，ニュース音声から話題語を抽出する方法について検討した．単語の出現頻度に基づく種々の尺度について実験的に有効性を比較し，その背景について考察した．ニュース記事中の単語の相対的出現頻度により，良い精度で話題語を抽出できることがわかった．特に情報量としての意味をもつ式(1)が優れていることがわかった．

話題抽出精度をさらに向上させるには，シソーラスや分類語彙表などを用いて，単語間の意味関係を考慮することなどが考えられる．


### 謝辞

ニュース原稿および音声データを提供して頂いた日本放送協会，新聞記事のデータベースを使わせて頂いた日経新聞社に感謝します．日頃ご支援を頂くNTT研究所の大附克年氏，そして東工大の研究室の方々に感謝します．


### 参考文献


[1] 高木，桜井，岩崎，古井: "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 (1998)

[2] S. Furui, K. Takagi, A. Iwasaki, K. Ohtsuki, T. Matsuoka and S. Matsunaga: "Japanese broadcast news transcription and topic detection", Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.144-149 (1998)

[3] 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"，音学音声研資，SP97-27 (1997)

[4] "日本経済新聞 CD-ROM 版 1990 年版～1994 年版"，日本経済新聞社，(1994-1995)

[5] 海野: "出現頻度情報に基づく単語重みづけの原理"，Library and Information Science, 26, pp. 67-88 (1988)

[6] 岩山，徳永: "重み付き IDF を用いた文書の自動分類について"，情報処理学会自然言語研究会，100-5，pp.33-40 (1995)


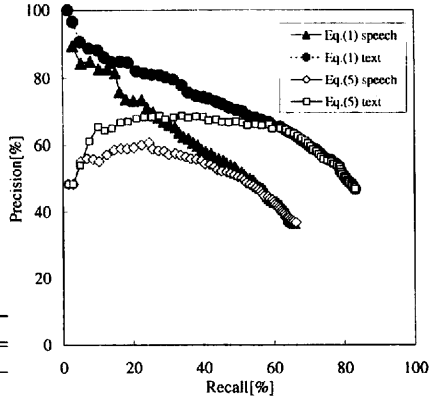


図 1. 話題語抽出の Recall と Precision