

# ニュース映像検索システムのための索引語の自動抽出\*

高 橋 伸 弥\*\*  
 高 井 大 介\*\*\*  
 森 元 逞\*\*

## Automatic Extraction of Index Terms for Retrieval of Broadcast News

Shin-ya TAKAHASHI, Daisuke TAKAI and Tsuyoshi MORIMOTO

In this paper we describe a new method for extracting index terms for retrieval of broadcast news automatically and accurately using speech recognition technique and language model adaptation. Based on the idea that the broadcast news have similar Web documents on the Internet news site, we propose a system that makes a language model adapt to news documents using similar Web documents retrieved with candidate words of index terms included in recognition results. The performance of speech recognition is improved by executing this process iteratively. To show the effectiveness of this approach, we demonstrate some experimental results.

**Key Words:** Information Retrieval, Speech Recognition, Broadcast News, Index Terms

### 1. はじめに

近年の電気通信技術の飛躍的な発展に伴って、放送媒体の多様化・多チャンネル化が進んでいる。視聴者に提供される映像量は着実に増加しているため、映像を蓄積するだけでなく、視聴者自身による検索を容易にするための技術が求められている。特にニュース映像は、その内容の重要性と利用価値の観点から、索引付きのデータベースとして保存する価値が高いと考えられており、テレビ局を中心に既に多くの試みがなされている。しかし、このようなニュース映像に対する索引付け作業は、放送台本や取材メモを頼りに人手で行われているのが現状であり、日々大量に作り出されているニュース映像全てに人手で索引付けするのは非常に膨大なコストがかかってしまう。

この問題に対し、ニュース映像の音声データを音声認

識し、その認識結果から索引語として適切な語を抽出する方法が提案されている<sup>(1)</sup>。この手法は比較的簡単に実装でき、高速に索引語を抽出できる点で実用的ではあるが、既存の汎用言語モデルを用いてニュース音声を認識しているため、認識結果から抽出した索引語の信頼性の点で改善の余地がある。

そこで本研究では、ニュース映像のトピックに合わせて言語モデルを動的に更新させることを考える。具体的には、配信されたニュース映像と同一の情報源から作成された Web 上のニュース記事からトピックに適応した言語モデルを作成し、これを用いてより高精度な認識を行うことで、信頼できる索引語を得ようというものである。本論文では、以上のアイデアに基づいた索引語自動抽出システムを提案し、評価実験により有効性を示す。

### 2. 信頼性の高い索引語の自動抽出

#### 2.1 索引語自動抽出システム

索引語自動抽出システムの処理の流れを図 1 に示す。このシステムは、

\* 平成18年1月20日受付

\*\* 電子情報工学科

\*\*\* 電子工学専攻(現NEC通信システム)

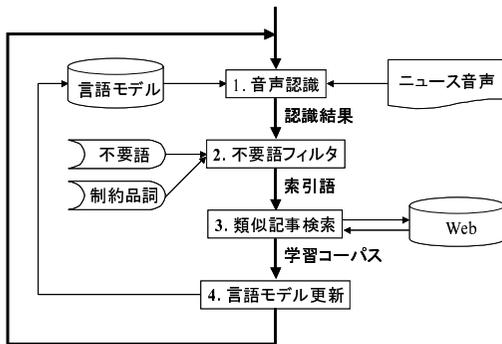


図1 索引語自動抽出システム

1. 索引語抽出対象となる単一トピックのニュース映像の音声を用言言語モデルを使用して音声認識器<sup>(注1)</sup>で認識する。
2. 認識結果から索引語として適切でない語(不要語<sup>(3)</sup>)や品詞(制約品詞)を不要語フィルタで除去し、索引語を得る。
3. その索引語を検索質問として、ニュース映像のトピックに類似した記事を Web 上から検索、収集する。
4. 収集した記事を学習コーパスとし、汎用言語モデルをトピックに適応した言語モデルへと更新させる。
5. トピックに適応した言語モデルを用いて、再び同一のニュース音声を認識する。

という処理を索引語が収束するまで繰り返し、収束語の索引語をニュース映像の索引語として抽出するものである。

### 2.2 Web 検索による学習コーパスの作成

索引付け対象のニューストピックに適応した言語モデルを作成するためには、トピックに適した多量の記事を収集しなくてはならない。しかし、ニュースのトピックは日々変化するため、過去のニュース記事を用いては最新のニュース映像に適応した言語モデルを作成することはできない。そこで、この問題を解決するために、Web 上に散在するニュース記事の利用を考える。すなわち、あるトピックのニュース映像が放送されると、そのニュース映像と同一情報源をもつニュース記事がほぼ同時刻に Web 上で配信されることから、その記事を検索・収集して、言語モデル適応のための学習コーパスとしようというものである。

### 2.3 Web 検索のための検索質問

ニュース音声の認識結果から自動生成された検索質問には誤認識語が含まれていることが予想される。そのため、検索質問を論理式で表現する検索モデルを採用した既存の検索エンジン(Google<sup>(注2)</sup>や Alta Vista<sup>(注3)</sup>など)を用いた場合には、誤認識語を含んだ記事をも収集して

しまい、高精度な言語モデルを作成することは難しいと考えられる。そこで、本システムでは、以下に説明するベクトル類似度に基づく類似記事検索を行うことにする。この検索方法を用いれば、単語を含むか否かではなく全体として類似した記事を収集できるので、元々のニュース音声の内容に近い記事を収集することができ、汎用言語モデルを個々のニューストピックに特化したモデルへと適応させ、認識結果から信頼性の高い索引語を抽出することができると思われる。

## 3. 索引語自動抽出システムの実装

### 3.1 ニュース記事の自動収集

前節までに述べた類似記事検索は、図1に示した言語モデル更新のプロセスを繰り返す度に、認識結果に応じて実行するのが望ましいが、現状ではこの処理を高速に行うことは難しいため、本論文では、類似記事を Web 上から検索するのではなく、あらかじめ新聞社の Web サイト上の全ニュース記事をローカルマシンに保存しておき、その記事集合の中からニューストピックに類似した記事を検索・収集することとする。

#### 3.1.1 自動収集アルゴリズム

ニュース記事自動収集の流れを図2に示す。まず始めに、巡回の対象となる新聞社 Web サイトのトップページを調査し、リンク先 URL を抽出する(1)。このとき、抽出した URL が相対パスであれば絶対パスに変更する。次に抽出した URL のサーバドメイン名をチェックする。一般的な URL は、プロトコル名、サーバドメイン名、資源へのパス、ファイル名から構成される。ここでは、抽出 URL のサーバドメイン名が新聞社のトップページのサーバドメイン名と一致しない場合、その URL を破棄することとする。そして、このチェックを終えた URL を未調査 URL キューへと格納する(2)。次に、未調査 URL キューから調査対象となる URL を取り出し、調査済み URL ハッシュへと登録する(3)。この調査済み URL ハッシュにより、既に調査済の URL の再調査を回避する。その後、調査対象となる URL はニュース記事を含むか否かのチェックを受ける(4)。この処理によって、ニュース記事を含まない URL であると判断された場合には(5)、トップページと同様に、リンク先 URL のみを抽出する(6)。ニュース記事を含むと判断された場合には(7)、リンク先 URL の抽出と並行して、ニュース記事のみをページから抽出し、保存する(8)。このとき、抽出されたリンク先 URL は既に調査済みか否かの判断を受け(9)、サーバドメインチェック(10)の後、未調査 URL キューへと保存される(2)。

この一連の処理を未調査 URL キューが空になるまで繰り返すことで、新聞社 Web サイトを効率良く巡回し、

ニュース記事のみを収集することができる。

### 3.1.2 ニュース記事を含む Web ページの判別

新聞社 Web ページの構成は図3のようになっている。トップページには主に、各ジャンル名と各ジャンルごとの最新の見出しが掲載され、各ジャンル名には、そのジャンルの全見出しが掲載されたジャンル別一覧ページへのリンクが張られている(a)。また見出しには記事全文ページへのリンクが張られており(b)、記事全文ページには抽出すべき記事の他に、不要情報(新聞社名、広告や他の記事へのリンク)が記載されている。この図の(b)の矢印で指されているリンク先が、収集の対象となるニュース記事を含む Web ページである。したがって、このニュース記事を含む Web ページをサイト巡回の際に判別し、ニュース記事と不要情報の混在するページから、ニュース記事のみを抽出する必要がある。

Web ページがニュース記事を含むか否かは、Web ページの内容ではなく、その Web ページのファイル名によって判別する。このアプローチは、「ニュース記事を含む Web ページはファイル名にニュース記事掲載日が用いられている」という傾向を利用したものである。

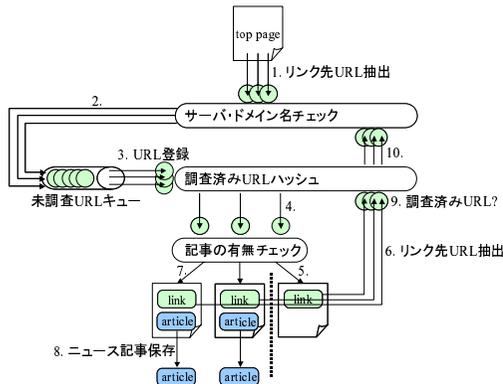


図2 ニュース記事自動収集の流れ

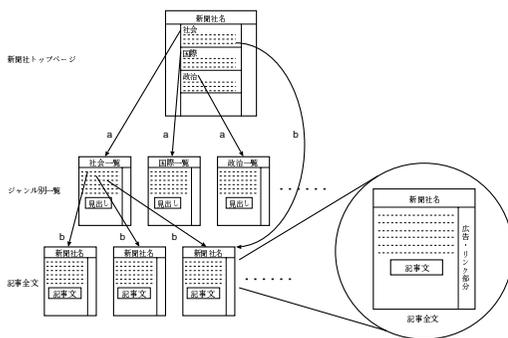


図3 新聞社 Web サイト構成

### 3.1.3 ニュース記事の抽出

ニュース記事を含むページの記事部分は、そのページの主要部分であるため、広告欄やリンク先などの不要情報に比べ占有面積が大きい。そのため、占有面積の大きい部分をニュース記事とみなすことで、ニュース記事のみを抽出することができる。面積の計算には、ウェブページを構成する HTML ファイル内のテーブルレイアウトのパラメータを手がかりとした。

### 3.2 類似記事検索アルゴリズム

#### 3.2.1 ベクトル空間モデルに基づく類似記事検索

前節までに示した方法により自動収集された記事を対象として、認識結果に類似した記事を検索し、ニュース適応言語モデルを作成する。本システムでは、この類似記事検索を実現するための検索モデルとして、前述したようにベクトル空間モデル<sup>(4)</sup>を採用することとした。このモデルは類似検索の際に用いられる代表的な検索モデルであり、その最大の特徴は文書と検索質問をそれらから抽出された索引語の重みを要素とする多次元ベクトルで表現し、文書と検索質問の適合性判断をベクトル間の類似度計算に帰着させる点にある。

まず始めに、学習コーパス生成のためにローカルマシンに保存した全記事集合をベクトル化し、ニュース記事ベクトルを得る。このニュース記事ベクトルの要素は文書から抽出された索引語の重みであるので、ベクトル化にあたっては、個々のニュース記事から記事を特徴付ける索引語を抽出しなければならない。この索引語抽出処理では形態素解析システム茶筌<sup>(5)</sup>を用いてニュース記事を形態素解析し、その結果から、索引語として適切な語を抽出する。次にこれらの索引語に対して後述する TF・IDF 重み付けを行い、この重みを要素としてニュース記事ベクトルをベクトル化する。このベクトル化が完了した後、入力された検索質問に対しても同様にベクトル化処理を施し、作成した検索質問ベクトルとニュース記事ベクトルとの類似度を算出し、類似度算出結果を得る。そして最後に、類似度がある閾値以上のニュース記事を収集し、学習コーパスを生成する。

#### 3.2.2 索引語に対する TF・IDF 重み付け<sup>(3),(6)</sup>

TF・IDF 重み付けとは、索引語の重み付けで用いられる代表的なアプローチである。この重み付けは、文書  $d$  における索引語  $t$  の重みを  $w_t^d$  とした場合、以下のような式で定義される。

$$w_t^d = tf(t, d) \cdot idf(t) \quad (1)$$

ここで、 $tf(t, d)$  は、ある文書  $d$  中に出現する索引語  $t$  の頻度を表しており、また、 $idf(t)$  は、索引語  $t$  の全文書に対する索引語  $t$  の頻度を表している。TF (term frequency) と IDF (Inverse Document Frequency)

の積で定義されることから、TF・IDF 重み付けと呼ぶ。索引語頻度  $tf(t, d)$  の計算方法としては様々なものが提案されているが、本システムでは計算が容易な以下の式を採用することとする。

$$tf(t, d) = 1 + \log_e f(t, d) \quad (2)$$

ここで、 $f(t, d)$  は文書  $d$  中に含まれる索引語  $t$  の出現回数である。

上記の  $tf(t, d)$  は各文書の中の頻度は考慮しているが、文書集合全体の中での索引語の分布については考慮していない。索引語がどの程度その文書を特徴付けているかという性質を考慮するためには、他の文書についても考える必要がある。なぜなら、たとえ高頻度の索引語であってもどの文書にも現れるような索引語は、その文書の特徴付ける語にはならないからである。

このため用いられる尺度として用いられているものが IDF である。IDF はある索引語が全文書中のどの程度の記事に出現するのかを表す尺度である。その定義はいくつかあるが、本システムでは以下の式を採用する。

$$idf(t) = \log_e \left( \frac{N}{df(t)} + 1 \right) \quad (3)$$

ここで、 $N$  は文書の総数、 $df(t)$  は索引語  $t$  が出現する文書数である。

### 3.2.3 余弦尺度による類似度算出<sup>(3)</sup>

以上の手順でニュース記事と検索質問をベクトル化した後、各ニュース記事がどの程度、検索質問に適しているかの適合度をベクトル間の類似度によって計算する。類似度の計算方法としては、内積、Dice 係数、Jaccard 係数、余弦を用いるものなどがあるが、本システムでは、この中で最も一般的な余弦尺度を用いることとする。

$\mathbf{x} = (x_1, x_2, \dots, x_N)$  と  $\mathbf{y} = (y_1, y_2, \dots, y_N)$  の2つのベクトル間の余弦尺度に基づく類似度  $sim(\mathbf{x}, \mathbf{y})$  は、次の式で定義される。

$$sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| |\mathbf{y}|} = \frac{\sum_{i=1}^N x_i \cdot y_i}{\sqrt{\sum_{i=1}^N x_i^2} \cdot \sqrt{\sum_{i=1}^N y_i^2}} \quad (4)$$

ここで、 $\mathbf{x} \cdot \mathbf{y}$  はベクトル  $\mathbf{x}, \mathbf{y}$  の内積であり、 $|\mathbf{x}|$  及び  $|\mathbf{y}|$  は、ベクトル  $\mathbf{x}, \mathbf{y}$  の大きさである。

ベクトル空間モデルにおける余弦尺度では、 $N$  次元のベクトル空間において、2つのベクトルのなす角度  $\theta$  が小さいほどその2つのベクトルが類似しているとする。学習コーパスを生成する際には、あらかじめこの類似度の閾値を定めておき、その閾値以上のニュース記事を収集すればよい。

## 4. 評価実験

### 4.1 実験条件

テストデータには、2004年12月8日の21時から21時15分にNHKが放送したニュース映像“NHK ニュース9”を用いた。このニュース映像の音声データをwav形式で保存し、さらにニューストピックとは関係のない挨拶部、音楽部、シャッター音等の雑音部を除去した後、ニュース音声を単一トピックへと分割し、テストデータとして利用した。表1に分割したニューストピックと放送時間を示す。

学習用テキスト収集のための記事には、NHK ニュース9の放送翌日に自動収集した2653記事を用意した。これらのニュース記事は、朝日、毎日、読売、産業経済、日経、東京、西日本、京都、中日、中国新聞社、河北新報社のWebサイト上から収集した。

言語モデルの学習は5回行い、学習0回目、すなわち学習開始時に用いた汎用言語モデルは、Julius 付属の高頻度語2万語言語モデルを用いた<sup>(7)</sup>。1回目の学習以降で作成するニュース適応言語モデルはバイグラム言語モデルである<sup>(註4)</sup>。

なお音響モデルにはJulius 付属の性別非依存モデルを用いている。

### 4.2 評価尺度

品詞制約および不要語フィルタを施したあとの索引語候補に対する再現率 (Recall) とノイズ率 (Noise) を評価尺度とした。ここで再現率は正しい索引語の抽出漏れの少なさを表す尺度であり、ノイズ率は抽出された索引語の中に含まれる誤った索引語の割合である。それぞれ次式で定義される。

表1 ニューストピックと放送時間

ニューストピック
北朝鮮、横田めぐみさんとは別人の遺骨を提出(298秒)
ETC 利用者に誤った通行料金を請求(68秒)
住宅金融公庫が抱える損失の処理(69秒)
定率減税の縮小・廃止(77秒)
血液型で性格判断をする番組について(79秒)
ビールメーカー、卸売業者へのリベートを廃止(28秒)
70歳男性、女子高校生へのストーカで逮捕(29秒)
混合診療をめぐる対立(33秒)
ナポレオンの回顧録、3500万円で落札(25秒)
為替と株の値動き(16秒)
明日の天気(35秒)

再現率 [%]

$$= \frac{(\text{抽出された正しい索引語の延べ数})}{(\text{抽出されるべき正しい索引語の延べ数})} \times 100$$

ノイズ率 [%]

$$= 1 - \frac{(\text{抽出された正しい索引語の延べ数})}{(\text{抽出された索引語の延べ数})} \times 100$$

実験で用いた不要語および制約品詞を表2に示す。

### 4.3 評価実験：トピック毎の評価値の変化

表1で示した11トピックそれぞれの評価値を求めた。ただし類似度閾値は予備実験により定めた40%に設定している。

実験結果を図4から図14に示す。ここでは、棒グラフで学習コーパスのサイズも併せて示している。実験結果をみると、全11のトピックのうち、索引語の精度が改善されたものが7、1回目の学習で改善されるものの2回目以降の学習で悪化するものが2、まったく改善されないものが2となっている。

精度が下がっているケースでの共通点は、類似度閾値を40%と低く設定しているにもかかわらず、学習コーパスの総サイズが小さくなっている点にある。これは即ち、ニュース音声の内容に類似した記事が見つからなかったことを意味している。特に、一旦改善されたのち低下するものは、類似してはいるが異なる記事を学習コーパスとしてしまったことで誤認識が増加したものと考えられる。学習コーパスのサイズと索引語の抽出精度に関しては、今後、更に検討する必要があるだろう。

### 5. おわりに

本論文では、信頼性の高い索引語をニュース音声の認識結果から自動抽出するシステムを提案した。さらに、このシステムの評価のために、11トピックを含むNHKのニュース映像を用いて、評価実験を行った。実験の結果、11トピック中2トピックだけが、学習後に索引語の精度が低下したが、他はニュース適応言語モデルの効果が見られた。また今回の実験結果では、2回目の繰り返し処理で既に索引語の収束する傾向が見られた。これは、類似記事を検索するニュース記事集合が小規模かつ固定的であったためだと考えられる。このため、ローカルマシンに記事を保存することなく、直接Web上から類似記事を収集することが今後の課題として挙げられる。

表2 不要語と制約品詞

不要語	こと, する, できる, とる, みる, やる, わかる, 見る, 言う, 考える, 行う, 行く, 人, 話, 他, ...
制約品詞	一般名詞, 固有名詞, サ変接続名詞, 自立動詞-非自立動詞の5品詞以外

さらに、今回の実験は類似した記事といった少数の学習コーパスを用いて新たな言語モデルを作成した。しかし、<sup>(9)</sup>, <sup>(10)</sup>, <sup>(11)</sup>のように、少数のデータを用いてNグラム適応を行う手法が検討されている。今後は、この手法を本システムに用いることで、索引語の精度のさらなる改善が見込まれる。

### 注

- (1) 日本語大語彙音声認識エンジン Julius<sup>(2)</sup>を使用
- (2) <http://www.google.com/>
- (3) <http://www.altavista.com/>

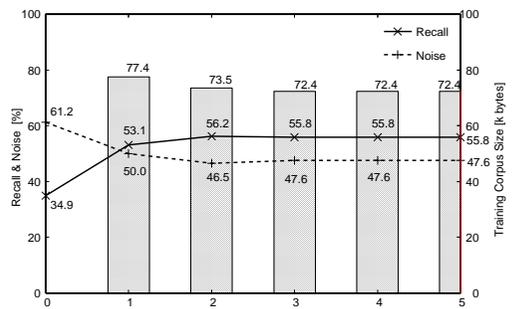


図4 北朝鮮

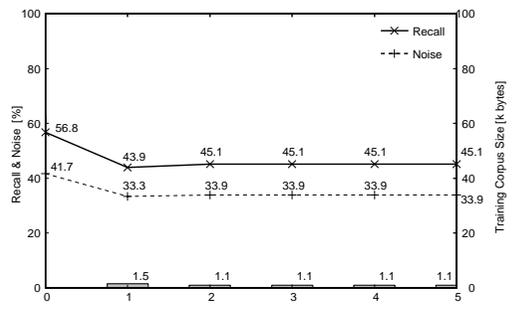


図5 ETC

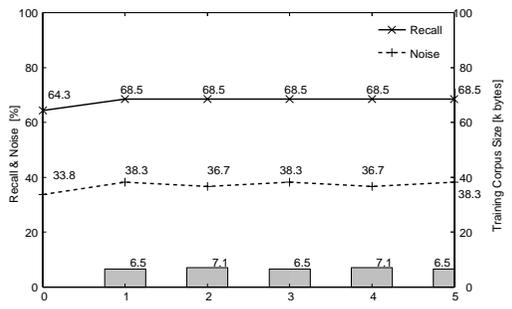


図6 住宅金融公庫

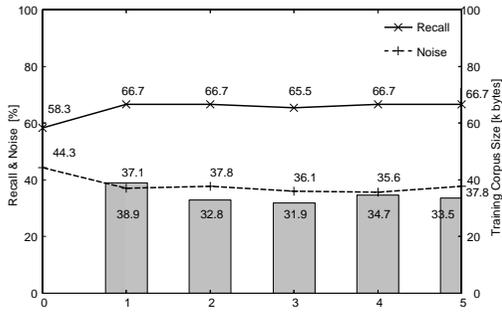


図7 定率減税

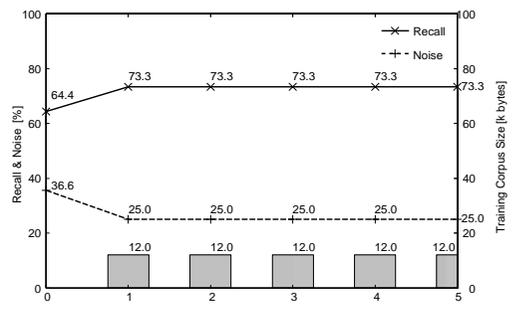


図11 混合診療

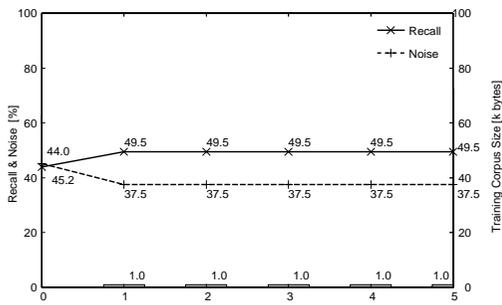


図8 血液型性格判断

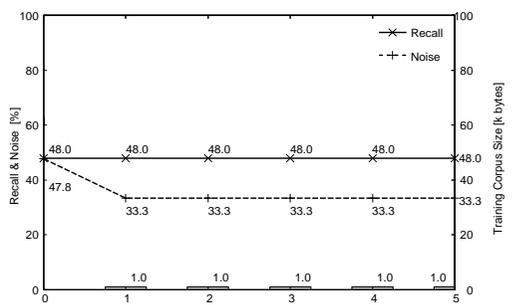


図12 ナポレオンの回顧録

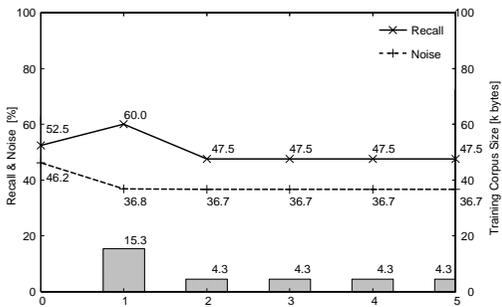


図9 リベート廃止

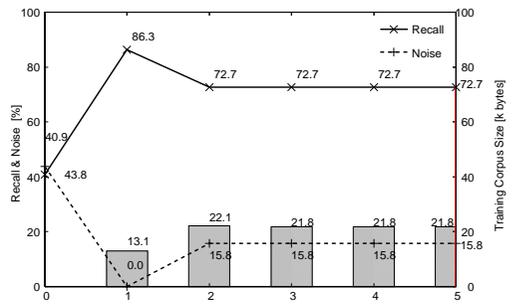


図13 為替と株

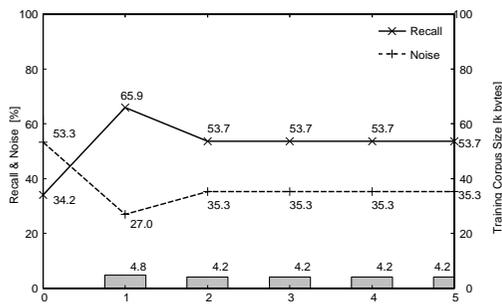


図10 ストーカ

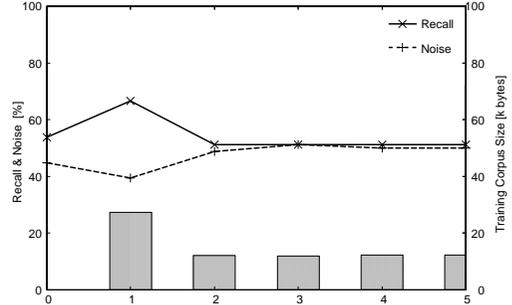


図14 明日の天気

(4) 作成には統計的言語モデルの作成キットである“CMU-Cambridge SLM Toolkit”<sup>(6)</sup>を用いた。また今回の実験では繰り返しの度に新たな言語モデルと置き換えることで言語モデルの更新を行っている。

#### 参 考 文 献

- (1) 西崎, 中川: “音声入力によるニュース音声検索システム”, 信学技報, SP99-108, pp.91-96(1999).
- (2) 李, 河原, 堂下: “単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識”, 信学論, Vol.J82-DII, No.1, pp.1-9(1999).
- (3) 徳永: “情報検索と言語処理”, 東京大学出版会(1999).
- (4) G.Salton *et. al.*: “A vector space model for automatic indexing”, *Communications of the ACM*, Vol.18, No.11, pp.613-620, 1975. Reprinted in *Readings in Information Retrieval*, K. Jones and P. Willett(Eds.), Morgan Kaufmann Publishers, pp. 273-280 (1997).
- (5) 松本 他: “日本語形態素解析システム『茶筌』Version 2.0使用説明書 第二版”, *Information Science Technical Report NAIST-IS-TR99012*, Nara Institute of Science and Technology(1999).
- (6) I.H.Witten, *et. al.*: “Compressing and Indexing Documents and Images, 2nd ed.”, Morgan Kaufmann(1999)
- (7) 伊藤 他: “大語彙日本語連続音声認識研究基盤の整備-学習・評価テキストコーパスの作成”, 情処研資, SIG-SLP18-2, pp7-12 (1997).
- (8) P.R.Clarkson and R. Rosenfeld: “Statistical Language Modeling Using the CMU-Cambridge Toolkit”, *Proc. ESCA Eurospeech*, pp.2707-2710 (1997).
- (9) M.Federico: “Bayesian estimation methods for Ngram language model adaption”, *Proc. ICSLP*, pp.240-243 (1996).
- (10) 伊藤, 好田: “対話音声認識のための事前タスク適応の検討”, 信学技報 NLC96-50, SP96-81(1995).
- (11) 政瀧 他: “MAP 推定を用いた N-gram 言語モデルのタスク適応”, 信学技報 SP96-103(1997).

