Vol.2009-SLP-79 No.32
2009/12/22

WWW を利用したトピック関連語推定に基づく言語モデル教師なし適応の性能評価

増村 亮[†1] 伊藤 仁[†1]
伊藤彰則[†1] 牧野正三[†1]

大語彙連続音声認識の高精度化のために，WWW 上から認識対象のトピックに関連したテキストを収集し，言語モデル適応を行う．我々は，認識対象の音声認識結果から全自動で検索クエリを生成する教師なしの方法に焦点を当てる．本稿では，WWW を利用して単語の関連性を表す特徴ベクトルを抽出することで，適切にトピック関連語およびサブトピックを推定する方法を提案した．そして，音声認識実験を行うことで提案法の有効性を確認した．

Evaluation of Unsupervised Language Model Adaptation based on Topic-related Word Estimation using WWW

Ryo Masumura ,[†1] Masashi Ito ,[†1] Akinori Ito [†1]
and Shozo Makino[†1]

To improve the accuracy of an LVCSR system, we gather topic-related documents from WWW, and adapt the language model. We focus on an unsupervised method that automatically generate search queries from an automatic transcription by a speech recognizer. In this paper, we proposed a new method to estimate topic-related word and sub-topic by extracting feature vectors from WWW, which express relevance between the words. We carried out a speech recognition experiment. The experimental result showed effectiveness of the proposed method.

†1 東北大学大学院工学研究科
Graduate School of Engineering, Tohoku University

1. はじめに

大語彙連続音声認識のための言語モデルとして，統計的言語モデルの N-gram が広く用いられている．一般的に N-gram は，様々なトピックを含むテキストコーパスから学習する．このような N-gram は，一般的な音声入力に対して高い性能をみせるが，入力の話題によっては，性能が上がらないことがある．その理由として，未知語 (Out-Of-Vocablary) や，連鎖確率が低いための誤認識の問題が挙げられる．この問題を解決する手段として，言語モデルを認識対象に適応させる方法が多く検討されている[1)2)]．我々は，言語モデル適応のアプローチとして，World Wide Web(以下 WWW) から認識対象に関連したテキストを収集する方法に焦点を当てる[3)4)]．

WWW 上に存在する大量のテキストデータから対象とするテキストを収集するには，検索エンジンによる収集が効率的である．西村らは，検索クエリを人手で準備することで WWW テキストを取得している[5)]．また，翠らは知識ベースとなるテキストからクエリを構成[6)]，根本らは副次的な情報 (講義スライド) からクエリを構成している[7)]．これらは全て人手でトピックに関する情報を与えるような教師ありの手法である．一方，人手でトピックを与えない教師なしの方法として，認識対象の音声認識結果から全自動でクエリを構成する方法がある[8)]．教師なしの枠組みにおける検索クエリ構成の問題点は，音声認識結果には多くの誤認識が存在することである．誤認識単語を含む検索クエリでは性能の上昇が望めない．したがって，我々は誤認識単語を避けながらトピックに関連した検索クエリを構成する必要がある．

本研究では，WWW テキストを利用した言語モデル教師なし適応の高精度化を目標とする．以前我々は，認識対象の未知語および重要単語を多く含むテキストを取得できるような検索クエリの構成方法について報告した[9)]．本稿では，以前提案した方法の本質を明らかにするとともに，検索クエリの誤認識単語に対する頑健性，および取得テキストの有効性について実験用テストセットを増やし再調査した．さらに，実際に言語モデル適応を行うことによる性能についても報告する．

2. WWW を利用した言語モデル適応

2.1 教師なし適応手順

WWW を利用した言語モデル教師なし適応の流れは以下の形をとる (図 1).

step1 話題非依存のベースラインコーパスから学習したベースライン言語モデルを使用し

Vol.2009-SLP-79 No.32
2009/12/22

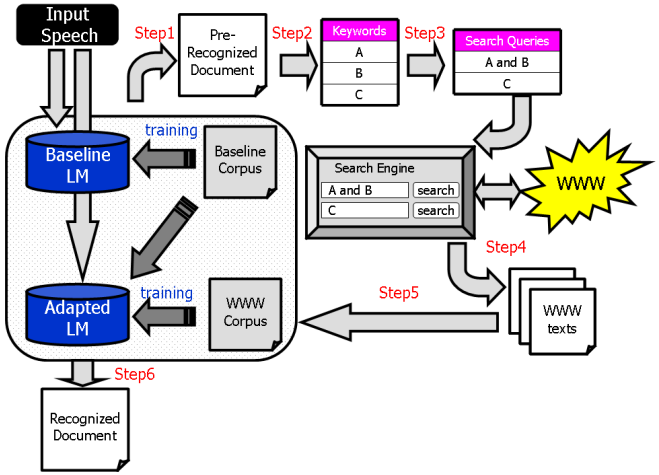

図 1 WWW を利用した言語モデル教師なし適応
Fig. 1 Unsupervised lamguage model adaptation using WWW

て，入力音声を認識．

step2 認識結果に出現した単語から，キーワードを選択．

step3 選択したキーワードから検索クエリを構成．

step4 検索エンジンを用いて WWW からテキストを取得．

step5 取得したテキストを整形し，WWW コーパスを作成．

step6 ベースラインコーパスと取得した WWW コーパスから新たに適応言語モデルを学習し，入力音声を再認識．

2.2 検索エンジンおよび WWW ページのフィルタリング

本研究では，**step4** における検索エンジンとして Yahoo Japan[10] を用いる．検索クエリをサーバに送信し，検索結果を受信するために，Yahoo API[11] を用いた．Yahoo API を用いることにより，1 つの検索クエリから最大 1000 件の URL を得ることができる．

また **step5** では，西村らによって提案されている統計ルールフィルタを用いてテキスト整形を行う[5]．

3. トピック関連語推定

step2 で音声認識結果から選択するキーワードは，認識対象のトピックに関連のある単語が望ましい．我々は，認識対象に出現する単語がどの程度トピックに関連があるのかをスコア付けする処理を「トピック関連語推定」と呼ぶことにする．

トピック関連語推定後，スコア上位単語を検索クエリのためのキーワードとして選択する．なお，我々は名詞にのみ着目する．

3.1 従来のトピック関連語推定

一般的に，文書中の単語から文書の内容にとって重要性の高い単語を抽出する際に，$tfidf$ が用いられる[8]．$tf_D(w)$ は，文書 D における単語 w の対象文書における出現頻度，$df(w)$ は，単語 w の一般的な文書における大局的な出現頻度である．我々は，単語 w を WWW 検索クエリとすることでヒットするページ数を $df(w)$ とした．また WWW 検索でヒット可能な全てのページ数を N とすることで $idf(w)$ は算出できる．

$tfidf$ は単語 w の出現頻度のみを考えるので，本稿では **Frequency-based Score** と呼ぶ．音声認識結果 D_{SR} における **Frequency-based Score** の計算式は次の (1) 式の通りである．

$$F_{Score}(w) = tfidf_{D_{SR}}(w) = tf_{D_{SR}}(w) \cdot idf(w) = tf_{D_{SR}}(w) \cdot \log \frac{N}{df(w)} \quad (1)$$

しかし，$tfidf$ で音声認識結果に出現した単語を評価する場合，誤認識単語でも $tfidf$ が高くなってしまう場合が多く存在する．したがって，認識誤りを排除するトピック関連語推定を行う必要がある．

3.2 提案するトピック関連語推定

我々は，音声認識結果に出現した単語がどのような出現分布をしているのかも考慮に入れる．正解単語，特にトピックに関連した単語の場合，音声認識結果内にその単語と関係ある単語が多く出現すると考えられる．逆に誤認識単語の場合，関係ある単語はほとんど出現しないと考えられる．したがって，音声認識結果に出現する他の単語との関係性を利用するような，**Relevance-based Score** を提案する．

そのために，我々は WWW を利用することで単語 w と関係のある単語を表現するような特徴量を抽出する．単語 w を検索クエリとして取得できる文書 D_w に含まれる単語の関係性を (2) 式のようにベクトルで表現する．

$$\boldsymbol{f}(w) = [tfidf_{D_w}(w_1), tfidf_{D_w}(w_2), ..., tfidf_{D_w}(w_k)]^\top \quad (2)$$

このベクトルは，k を日本語の全名詞の数と考えた場合，ほとんどの要素が 0 となる非常

にスパースな特徴ベクトルであると言える．つまり，0 以外の値を持つ要素は w と関係のある単語であると考えられる．なお，今回我々はそれぞれの単語 w に対して，WWW 上のテキスト文字数約 2 万 (URL 約 50 件) から特徴ベクトルを生成した．

さらに，音声認識結果 D_{SR} の $tfidf$ による各単語のスコアを (3) 式のようにベクトルで表現する．

$$\boldsymbol{S}(D_{SR}) = [tfidf_{D_{SR}}(w_1), tfidf_{D_{SR}}(w_2), ..., tfidf_{D_{SR}}(w_k)]^\top \quad (3)$$

先ほどの単語 w の関係性を表す特徴ベクトル $\boldsymbol{f}(w)$ を用いて，スコアを表現するベクトル $\boldsymbol{S}(D_{SR})$ を重み付けすることにより，単語 w と関係ある単語に依存したスコアを獲得できる．**Relevance-based Score** は以下の (4) 式のように求める．

$$R_{Score}(w) = \frac{\boldsymbol{f}(w)^\top \boldsymbol{S}(D_{SR})}{|\boldsymbol{f}(w)|} = \frac{\sum_{i=1}^{k} tfidf_{D_w}(w_i) \cdot tfidf_{D_{SR}}(w_i)}{\sqrt{\sum_{i=1}^{k} tfidf^2_{D_w}(w_i)}} \quad (4)$$

Relevance-based Score は，w と関係のある単語が認識結果内で高いスコアである程高い値となる．したがって，誤認識単語の場合，関係ある単語は認識結果内にほとんど出現しないと考えられるので，従来のトピック関連語推定と比較して誤認識単語に頑健なトピック関連語推定が期待できる．

3.3 キーワード選択の評価

提案法と従来法，それぞれのトピック関連語推定により選択されるキーワードの評価を行う．実験条件の詳細を表 1 に示す．

(1) キーワードに含まれる誤認識単語の割合

最初に，提案法と従来法でそれぞれ選択するキーワードの数を 1 から 15 まで変化させ，選ばれたキーワード中に含まれる誤認識単語の割合を調べた．その結果を図 2 に示す．

この結果から，提案法は従来法と比較して誤認識単語の割合が低下していることが分かる．上位 10 単語をキーワードに選択した場合では，約 12 ポイントの改善が見られた．

(2) 未知語および重要単語カバー率

次に，キーワードの各単語を単独クエリとしてテキストをダウンロードした場合に，そのテキストが認識対象の未知語及び重要単語をどの程度カバーするかを調査した．

検索クエリを q，その検索クエリによってダウンロードしたテキストに含まれる単語の集合を $V(q)$ とする．ここで，ある単語集合 W に対するクエリ q のカバー率を以下の (5) 式のように定義する．

$$c(q, W) = \frac{|V(q) \cap W|}{|W|} \quad (5)$$

選択したキーワードをキーワード集合 $S_k = \{w_1, w_2, w_3, ..., w_k\}$ とする．また，キーワード集合内の各単語でダウンロードしたテキストに含まれる単語集合の和集合を $V(S_k) =$

表 1 実験条件の詳細
Table 1 Detail of experimental condition

音声認識デコーダ	Julius 4.1.2
音響モデル	CSJ 付属状態共有 triphone モデル
言語モデル	単語 2-gram，逆向き単語 3-gram
バックオフスムージング	witten-bell
学習コーパス	CSJ よりテストセット以外の 2536 講演
総形態素数	7652534
ユニグラムエントリ数	41695 (カットオフ:1)
実験用 テストセット	CSJ より 40 講演 「あなたがよく知っていること興味関心のあることへの客観的説明」
平均単語認識精度	62.45%
補正パープレキシティ	226.71
未知語率	1.85%

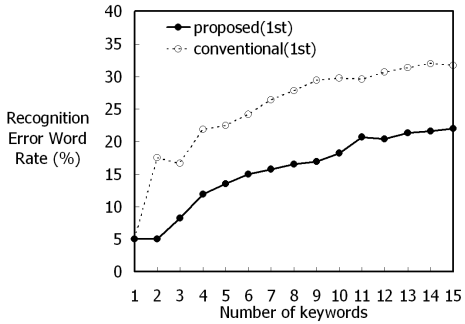

図 2 キーワード中の誤認識単語の割合
Fig. 2 Rate of mis-recognized word in keywords

Vol.2009-SLP-79 No.32
2009/12/22

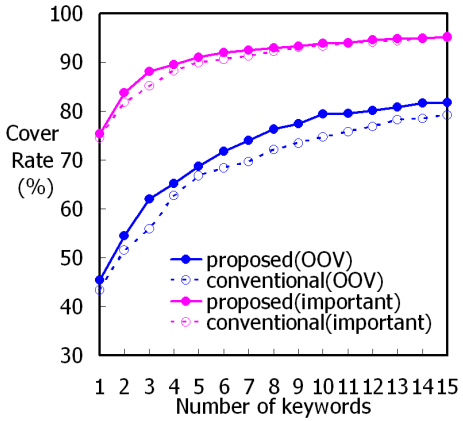

図 3 キーワード集合のカバー率
Fig. 3 Cover rate of keywords

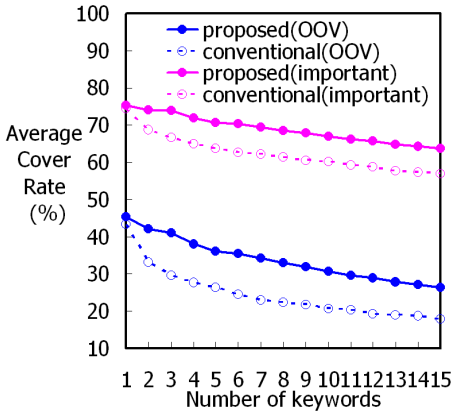

図 4 キーワード集合の平均カバー率
Fig. 4 Average cover rate of keywords

$V(w_1) \cup V(w_2) \cup ... \cup V(w_k)$ とする．ここで，ある単語集合 W に対するキーワード集合 S_k のカバー率を以下の (6) 式のように定義する．

$$c(S_k, W) = \frac{|V(S_k) \cap W|}{|W|} \tag{6}$$

さらに，ある単語集合 W に対するキーワード集合 S_k の平均カバー率を (7) 式のように定義する．

$$\bar{c}(S_k, W) = \frac{1}{|S_k|} \sum_{w_k \in S_k} c(w_k, W) \tag{7}$$

ここで，正解テキストに含まれるベースライン言語モデルの未知語の集合を U とする．さらに，正解テキストに含まれる名詞のうち，$tfidf$ の高い 50 単語を重要単語とみなし，これを V_I とする．

提案法と従来法で，それぞれ選択するキーワードの数を 1 から 15 まで変化させた場合に，各キーワードそれぞれ文字数約 50 万 (URL 約 1000 件) のテキストを取得した際の，キーワード集合の未知語カバー率 $c(V_k, U)$ および重要単語カバー率 $c(V_k, V_I)$ を調べた．その結果を図 3 に示す．また同様に未知語平均カバー率 $\bar{c}(V_k, U)$ および重要単語平均カバー率 $\bar{c}(V_k, V_I)$ を調べた．その結果を図 4 に示す．

図 3 から，各単語を検索クエリとして使用することで，提案法では，従来法よりも多くの未知語および重要単語を獲得できることが分かる．また図 4 から，提案法でキーワードを選択することによって，誤認識単語を含め，トピックに関係ない単語が大きく減少していると言える．この 2 つの結果は，従来法では下位になってしまったトピック関連語が提案法により上位になり，かつ，従来法で上位になってしまった誤認識単語が提案法により下位になったことを示唆する．上位 10 単語をキーワードに選択した場合では，キーワード集合の未知語カバー率が約 6 ポイント改善，未知語平均カバー率が約 11 ポイント改善した．

4. クラスタリングによるサブトピック推定

step3 で，選択したキーワードから検索クエリを構成する際，一般的に AND で組み合わせて検索クエリを構成するのが望ましい[4]．しかし，それとトレードオフの関係として，AND で組み合わせるほど検索でヒットするテキスト数が減ってしまう問題がある．我々はキーワードをサブトピックに分割することでこの問題に対応する．そのために，キーワード間の類似性からクラスタリングを行い，各クラスタをサブトピックとする．我々はこの処理を「サブトピック推定」と呼ぶことにする．

サブトピック推定後，各サブトピックをそれぞれ AND で組み合わせることで複数の検索クエリを構成する．

なおクラスタリングの方法は，キーワード 2 単語間の類似度を全て求め，それに従い凝集的階層化クラスタリングを行う[8]．クラスタリングの閾値として，各サブトピック内の単語を AND で組み合わせた際に，検索でヒットする URL が 1000 件以上としている．

4.1 従来のサブトピック推定

伊藤らは，キーワード 2 単語間の類似度を求める際，文書共起頻度に基づいた単語間類似度を使用している[8]．この単語間類似度は，2 単語の大局的な出現頻度を見ている．よって本稿では **Frequency-based Similarity** と呼ぶ．この単語間類似度は以下の (8) 式ように求める．

$$F_{Similarity}(w_i, w_j) = \frac{2 \cdot df(w_i \cdot w_j)}{df(w_i) + df(w_j)} \tag{8}$$

しかしこのクラスタリングでは，2 単語の共起頻度しか見ていない．もし，2 単語が本当に類似した単語ならば，それぞれの単語と関係ある単語群同士も類似していることが予想される．したがって，対象とする 2 単語以外の関係性を利用できる **Relevance-based**

Simirality を提案する．

4.2 提案するサブトピック推定

我々は，WWW から取得できる (2) 式で定義した特徴ベクトルを利用して，単語 w_i の特徴ベクトル $\boldsymbol{f}(w_i)$ と単語 w_j の特徴ベクトル $\boldsymbol{f}(w_j)$ のコサイン類似度を使用する方法を提案する．この特徴ベクトルはそれぞれの単語の関係性を表すので，それぞれの単語と関係ある単語群同士の類似性を比較することができる．この単語間類似度は以下のように求める．

$$R_{Similarity}(w_i, w_j) = \frac{\boldsymbol{f}(w_i) \cdot \boldsymbol{f}(w_j)}{|\boldsymbol{f}(w_i)||\boldsymbol{f}(w_i)|} \tag{9}$$

これにより，有効なサブトピックの生成が期待できる．

4.3 検索クエリ構成の評価

クラスタリング後のサブトピックそれぞれを AND 検索クエリとした場合に，k 個のサブトピックで構成された検索クエリを $Q_k = \{q_1, q_2, ...q_k\}$ とする．また，各クエリでダウンロードしたテキストに含まれる単語集合の和集合を $V(Q_k) = V(q_1) \cup V(q_2) \cup ... \cup V(q_k)$ とする．ここで，ある単語集合 W に対する検索クエリ Q_k のカバー率を以下の (2) 式のように定義する．

$$c(Q_k, W) = \frac{|V(Q_k) \cap W|}{|W|} \tag{10}$$

提案法 (R_Score) と従来法 (F_Score) で，それぞれ上位 10 単語をキーワードに選択し，提案法 (R_Similarity) と従来法 (F_Similarity) の単語間類似度を用いてクラスタリングを行った場合の検索クエリ構成の評価を行う．

各サブトピックのクエリでそれぞれ文字数約 50 万 (URL 約 1000 件) のテキストを取得した際の，検索クエリの未知語カバー率 $c(Q_k, U)$，重要単語カバー率 $c(Q_k, V_I)$ を調べた．その結果を図 5 に示す．

この結果から，従来法によるサブトピック推定と比較して，提案法の方が，未知語および重要単語を取得できるようなサブトピックに分割していることが分かる．また従来法同士の組み合わせ (F_Score ＋ F_Similarity) と比較して，提案法同士の組み合わせ (R_Score ＋ R_Similarity) では未知語カバー率が約 7 ポイント上昇した．

5. 言語モデル適応による再認識

実際にトピック関連語推定，およびサブトピック推定に基づいて検索クエリを構成し，

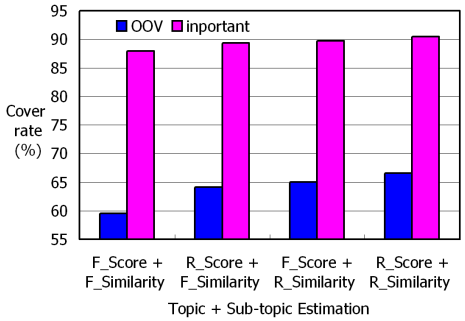

図 5 サブトピック推定によるカバー率
Fig. 5 Cover rate by estimation of sub-topic

表 2 適応言語モデルの詳細
Table 2 Detail of adapted language model

学習コーパス	CSJ よりテストセット以外の 2536 講演＋ WWW テキスト (文字数約 50 万)
ユニグラムエントリ数	WWW テキスト内の全語彙＋ 41695 (カットオフ:1)

WWW 上からトピックに関連したテキストを取得し言語モデル適応を行う．今回は，検索クエリの構成方法を比較するため，それぞれの検索クエリで取得する WWW テキストの量を，合計で文字数約 50 万 (URL 約 1000 件) に統一して実験を行う．つまり複数の検索クエリを構成した場合，等量ずつテキストを取得し，合計で文字数約 50 万とする．実験条件は表 1 と同様である．また適応言語モデルの詳細を表 2 に示す．

今回はキーワードを 10 単語に固定して実験を行う．構成する検索クエリとして，表 3 の 6 種類の方法について検討した．各検索クエリ構成方法に対して，言語モデル適応を行った際の再認識についての詳細を表 4 に示す．

この結果から，従来法のトピック関連語推定 (F_Score) より，提案するトピック関連語推定 (R_Score) を行った場合の方が有効に言語モデル適応が行われていることが分かる．単語認識精度は約 0.5 ポイント，補正パープレキシティは約 4.5 ポイント，未知語率は約 0.08 ポイントそれぞれ改善した．

さらにサブトピック推定により，各キーワードを検索クエリとする場合と比較して，有効

表 3 検索クエリの構成方法
Table 3 Composition of search queries

F_Score	従来法によるキーワード上位 10 単語をそれぞれクエリとする
R_Score	提案法によるキーワード上位 10 単語をそれぞれクエリとする
F_Score + F_Similarity	従来法によるキーワード上位 10 単語を従来法でクラスタリングし，それぞれのサブトピックを AND クエリとする．
R_Score + F_Similarity	提案法によるキーワード上位 10 単語を従来法でクラスタリングし，それぞれのサブトピックを AND クエリとする．
F_Score + R_Similarity	従来法によるキーワード上位 10 単語を提案法でクラスタリングし，それぞれのサブトピックを AND クエリとする．
R_Score + R_Similarity	提案法によるキーワード上位 10 単語を提案法でクラスタリングし，それぞれのサブトピックを AND クエリとする．

表 4 言語モデル適応の性能評価
Table 4 Evaluation of language model adaptation

検索クエリ構成	単語認識精度 (%)	補正パープレキシティ	未知語率 (%)
適応前	62.45	226.71	1.85
F_Score	64.73	202.46	0.82
R_Score	65.19	197.92	0.72
F_Score + F_Similarity	64.92	201.13	0.74
R_Score + F_Similarity	65.42	192.09	0.66
F_Score + R_Similarity	65.43	195.18	0.66
R_Score + R_Similarity	65.57	189.19	0.64

なテキストが取得できることが分かる．従来法である (F_Score + F_Similarity) と比較して，提案法 (R_Score + R_Similarity) では，単語認識精度は約 0.7 ポイント，補正パープレキシティは約 12 ポイント，未知語率は約 0.1 ポイントそれぞれ改善した．

なお，表 4 の結果は実験用テストセット 40 講演の平均の結果であるが，単語認識精度が適応前と比較して 10 ポイント以上改善するテストセットもいくつか存在した．

6. ま と め

本稿では，WWW を利用した言語モデル教師なし適応の高精度化を目的として，音声認識結果から，トピックに関連する単語を推定し，さらにサブトピックに分割する方法について検討した．

我々は，WWW から特徴ベクトルを抽出することで，単語同士の関係性を利用できるような方法を提案した．提案法により，誤認識単語に頑健なトピック関連語推定，および適切なサブトピック推定ができていることが確認できた．

さらに，実際に検索クエリを構成し，WWW 上からテキストを取得し言語モデル適応を行った．提案したトピック関連語推定，およびサブトピック推定により，単語認識精度は従来法よりも約 0.7 ポイント，適応前と比較して約 3.1 ポイント改善した．補正パープレキシティは従来法よりも約 11 ポイント，適応前と比較して約 38 ポイント改善した．また未知語率は，従来法よりも約 0.1 ポイント，適応前と比較して約 1.2 ポイント改善した．

今後は，再認識結果を利用して，反復適応を行う枠組みについて詳細を調べる予定である．

参 考 文 献

1) 伊藤 彰則，好田 正紀，“ N-gram 出現回数の混合によるタスク適応の性能解析 ”，電子情報通信学会論文誌，Vol.J83-DII，No.11，pp.2418-2427，2001.
2) 南條浩輝，河原達也，山田篤，内元清貴，“ 講演音声認識のための言語モデルの教師なし適応 ”，電子情報通信学会技術研究報告，SP2002-152，NLC2002-75，(SLP-44-32)，2002.
3) A.Sethy，P.G.Georgiou，S.Narayanan，“ BUILDING TOPIC SPECIFIC LANGUAGE MODELS FROM WEBDATA USING COMPETITIVE MODELS ”，In Proc.Interspeech，pp1293-1296，2005.
4) M.Suzuki，Y.Kajiura，A.Ito and S.Makino，“ Unsupervised language model adaptation based on automatic text collection from WWW ”，In Proc.Interspeech，pp.2202-2205，2006.
5) R.Nisimura，K.Komatsu，Y.Kuroda，K.Nagatomo ，A.Lee，H.Saruwatari and K.Shikano，“ Automatic n-gram language model creation from Web resources ”，In Proc.Eurospeech，pp.2127-2130，2001.
6) T.Misu，T.kawahara，“ A Bootstrapping Approach for Developing Language Model of New Sphoken Dialogue Systems by Selecting Web Texts ”，In Proc.Interspeech，pp.9-12，2006.
7) 根本雄介，秋田祐哉，河原達也，“ 講演音声認識のためのスライド情報を用いた言語モデル適応 ”，第 1 回音声ドキュメント処理ワークショップ講演論文集，pp89-94，2007.
8) A.Ito，Y.Kajiura，S.Makino and M.Suzuki，“ Unsupervised language model adaptation based on keyword clustering and query availability estimation ”，In Proc.Conf.on Audio，Language and Image Processing，pp.1412-1418，2008.
9) 増村亮，伊藤仁，伊藤彰則，牧野正三，“ WWW を利用した言語モデル適応のための検索クエリ構成の検討 ”，情報処理学会研究報告，Vol.2009-NL-191-12，Vol.2009-SLP-76-12，2009.
10) Yahoo! Japan，http://www.yahoo.co.jp/
11) Yahoo! developer’s network，http://developer.yahoo.com/