Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

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

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において，高精度な認識を実現する有効な手段として，認識対象にマッチしたテキストを収集し，認識対象に適応した言語モデルを作成する方法がある．この言語モデル適応のために，WWW(World Wide Web) から自動的に認識対象にマッチしたテキストの収集を行う．WWW からテキストを得るには，検索のためのクエリを構成する必要がある．本研究では，認識対象の未知語を獲得するような検索クエリの自動構成方法について検討を行った．

Composition Search Query for Language Model Adaptation using WWW

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

To improve the accuracy of an LVCSR system, it is effective to gather text data related to the topic of the input speech and adapts the language model using the text data. To create an adapted language model, we collect topic-related text automatically from the WWW(World Wide Web). Search query is necessary for retrieving topic-related text from the WWW. In this paper, we investigate automatic composition of a search query to acquire out-of-vocabulary words of the input speech.

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

1. はじめに

近年音声認識技術の実用化が進んでいる．その中でも大語彙連続音声認識 (ディクテーション) 技術は，音声を利用した様々なアプリケーションの基盤になる技術であり，音声入力ワープロ，放送や会議録の書き起こしなど，今まで基本的に全て人手で行ってきたことを自動化することが期待できる．

音声認識のための言語モデルとして，n-gram が広く用いられている．しかし，n-gram の語彙サイズは有限であり，またその連鎖確率は学習データに強く依存する．そのため，ある特定の話題を持った入力に対しては，認識に必要な単語が言語モデルの語彙に含まれていなかったり (未知語)，あるいは連鎖確率が低いために認識ができなくなる等の問題がある．この問題を解決する手段として，認識対象に関連したテキストを集めることで，認識対象に適応した言語モデルを作成する方法が有効であると考えられている[1)]．言語モデルの適応を行う場合，認識対象に関連したテキストをどのように集めるかが問題となる．

そのテキスト源として，World Wide Web(以下 WWW) に注目する．現在，WWW 上には 1 兆以上のページが存在すると言われている．そして，それらのページにアクセスする方法として，Google，Yahoo などの強力な検索エンジンが存在する．ここから検索可能なページは 2005 年の調査によると 11 億 5000 万ページと報告されている[2)]．このような背景から，認識対象に関連したテキストを集めるには，WWW を利用したテキスト収集が最適であると考えられる．

本研究では，WWW から自動的にテキスト収集を行うための，有効なキーワード選択，および検索クエリの自動構成について検討を行う．テキスト収集を行う際には，認識対象に含まれる未知語を収集すること，および認識対象に含まれる重要な単語をできるだけ多く含むデータを収集することを目標とする．

2. WWW を利用した言語モデル適応

大語彙連続音声システムにおいて，WWW を利用した言語モデル適応は，教師なし適応の方法をとるのが通常であり，過去すでに検討されている[3)]．本研究でも基本的にこれと同様の適応方法をとる．適応処理の流れは以下のように行う (図 1)．

step1 話題非依存のベースラインコーパスから作成したベースライン言語モデルを使用して，入力音声を仮認識する．

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

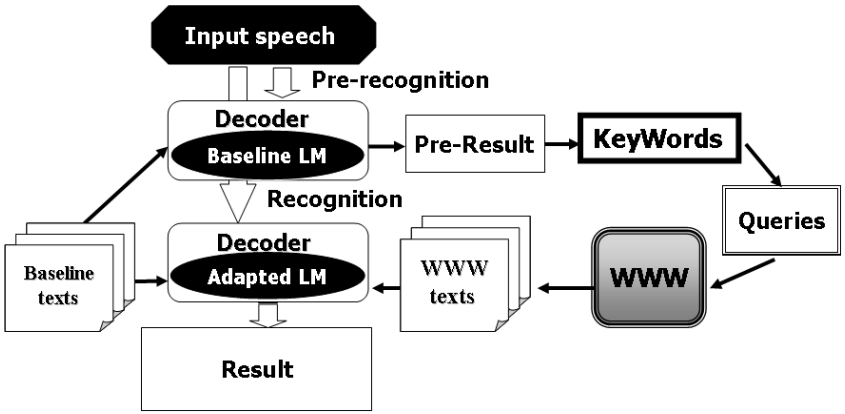

図 1 Web を利用した言語モデル適応
Fig. 1 Lamguage Model Adaptation using WWW

表 1 ベースライン言語モデル
Table 1 Baseline Langage Model

言語モデル	単語 2-gram，逆向き単語 3-gram
ベースライン言語モデルの学習コーパス	CSJ2004 年 2536 講演 毎日新聞 2000 年 100000 記事
ベースライン言語モデルの語彙の選択	top 60000 words

表 2 テストセットと仮認識
Table 2 Testset and Pre-recognition

ID(CSJ)	タイトル	単語認識精度 (%)	未知語数 (種類数)	未知語率 (%)
S04M1764	文字の歴史	41.38	51(38)	3.41
S04M1807	近代絵画	49.33	23(18)	1.39
S04M1678	木炭の効果	40.71	62(18)	1.62
S04M1730	経理事務	25.11	39(18)	0.98
S04M1725	飼い犬の話	49.95	36(20)	0.87
S04M1808	鉄鋼業	53.86	24(10)	0.37

step2 仮認識結果から，検索クエリを構成するためのキーワードを選択する．

step3 キーワードから検索クエリを構成して，WWW からテキストを取得する．

step4 ベースラインテキストと WWW テキストから新たに適応言語モデルを作成し，再認識を行う．

以上の方法で言語モデルタスク適応を行う．今回は **step2**，**step3** について検討を行った．

2.1 ベースライン言語モデル

汎用的なディクテーション用の言語モデルは，日本語の大規模なテキストから構築する．これにより，一般的な内容の音声に対応可能な言語モデルが作成できる．本研究では，汎用的かつ話し言葉に対応するために，様々なタスクに関する話し言葉の書き起こし文章を持つ CSJ[4] と，多くの話題が収められている毎日新聞の記事からベースライン言語モデルの学習を行う．

言語モデル作成の際の形態素解析には，形態素解析辞書 ipadic を用いた形態素解析システム chasen[5] を使用している．ベースライン言語モデルの詳細を表 1 に示す．

2.2 検索エンジン

本研究では，検索エンジンとして Yahoo Japan[6] 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[7] 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 個の検索結果を得ることができる．

2.3 キーワードと検索クエリ

検索エンジンに与える検索式を検索クエリとよぶ．検索クエリは，最も単純な場合はキーワードとなる文字列のみであるが，複数のキーワードに論理条件を組み合わせて指定することも可能であり，キーワードをできる限り組み合わせた「論理積」の検索を行う AND 検索クエリ，全てのキーワードの「論理和」の検索を行う OR 検索クエリを構成することが可能である．

2.4 実験に用いるテストセット

本研究では CSJ の模擬講演 S04「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」から 6 講演をテストセットに用いる．6 講演はいずれもサンプリング周波数 16kHz，16bit 量子化，モノラルで保存された音声である．以降の実験では，このテストセットを用いる．ここで，テストセットに対して CSJ のテストセット ID とタイトル，およびベースライン言語モデルを用いた仮音声認識実験の結果を表 2 に示す．

3. キーワード選択

認識対象に関連したテキストをダウンロードするための検索クエリの構成には，まずそのための質の高いキーワードが必要となる．質の高いキーワードとは，あるキーワードを単独

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

表 3 キーワード候補
Table 3 Key Words Candidacy

chasen の品詞番号	chasen の品詞の種類
2	名詞-一般
3	名詞-固有名詞
4	名詞-固有名詞-一般
5	名詞-固有名詞-人名
6	名詞-固有名詞-人名-一般
7	名詞-固有名詞-人名-性
8	名詞-固有名詞-人名-名
9	名詞-固有名詞-組織
10	名詞-固有名詞-地域
11	名詞-固有名詞-地域-一般
12	名詞-固有名詞-地域-国
16	名詞-副詞可能
18	名詞-形容動詞語幹

単語の検索クエリとした時に，未知語を含み，言語モデル適応に有効なテキストをダウンロードしてくるような単語であると考える．さらに最終的に検索クエリを構成することを考えてキーワードを集める必要がある．ここでは，そのようなキーワードの選択方法について検討を行う．

3.1 キーワード候補

本研究では，最初に仮認識文からキーワードとなり得る単語のみを抽出し，その単語の中から最終的なキーワードを選択する．本研究では，キーワードとなり得る単語は名詞とする．そのため chasen が付与した品詞のうち，表 3 に示す品詞を持つ単語をキーワード候補として扱う．

3.2 従来のキーワード選択法

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，$tfidf$ が一般的に用いられる．先行研究[3)8)] では，$tfidf$ を計算する際，単語を Yahoo API で検索した時のヒット数を df として用いた．$tfidf$ の計算式は次の (1) 式の通りである．

$$tfidf(w) = tf_w \cdot idf_w = tf_w \cdot \log \frac{N}{df_w} \quad (1)$$

N：全 WWW ページ数，tf_w：対象文書内の単語 w の頻度
df_w：単語 w をクエリとした時の検索ヒット数

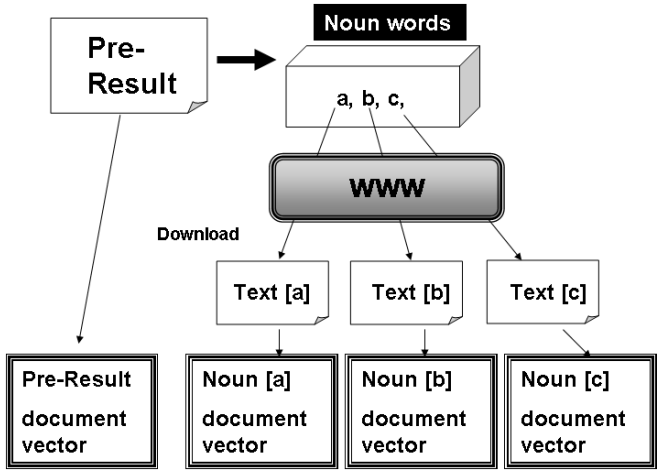

図 2 文書ベクトルの利用
Fig. 2 Using of Document Vector

先行研究では，この $tfidf$ を利用して以下のようにキーワード抽出を行っている．

従来法

(1) 仮認識文から，キーワード候補を抽出し $tfidf$ を求める．
(2) $tfidf$ 上位単語をキーワードとする．

3.3 ある単語がダウンロードするテキストの文書ベクトルの利用

従来法では，単語の頻度と検索ヒット数だけを見ており，同じ文章に出現した他の単語との関係を利用していないので，認識誤り単語であってもキーワードとして選ばれる可能性がある．この問題を解決するために，新たなキーワード選択法を検討した．

自然言語処理の技法として，文章を文書ベクトル (Document Vector) と呼ばれる高次元ベクトルで表現する方法がよく用いられる．文書ベクトルの要素には，単語の文章中の出現頻度や，$tfidf$ を用いる．本研究では，要素となる単語は先程述べたキーワード候補になり得る名詞で構成する．

さらに，文書ベクトルを高次元のベクトル空間上に配置することにより，文書ベクトル間の類似性をコサイン類似度によって表現できる．文書ベクトル間のコサイン類似度は次の

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

(2) 式の通りである．

$$CosSimilarity = \frac{\vec{v_A} \cdot \vec{v_B}}{|\vec{v_A}||\vec{v_B}|} \quad (2)$$

$\vec{v_A}$ ：文書ベクトル A

$\vec{v_B}$ ：文書ベクトル B

本研究では，仮認識文の文書ベクトル，そして，あるキーワードを単独クエリとしてダウンロードしたテキストで生成した文書ベクトルを利用したキーワード選択法を提案する (図 2)．

提案法

(1) 仮認識文から，キーワード候補の中でも $tfidf$ 上位の単語のみを抽出．今回は上位 50 単語を抽出した．

(2) 抽出単語それぞれを単独クエリとしてテキストをダウンロードし，文書ベクトルを生成．今回は 1 クエリあたり 20 ページをダウンロードした．

(3) キーワードの文書ベクトルと，仮認識文の文書ベクトル間の類似度を求め，類似度の上位単語のみをさらに抽出．今回は上位 15 単語を抽出した．

(4) 抽出したキーワードの文書ベクトル間の類似度を指標としてクラスタリングする．

(5) 保持 $tfidf$ の高いクラスタ内の単語をキーワードに選択．

3.3.1 仮認識文の文書ベクトルを用いることの信頼性

$tfidf$ は同じ文章に出現した他の単語との関係を利用していないことが問題であった．したがって，仮認識文内の他の単語からの影響も考慮に入れる．ここで，仮認識文の文書ベクトル $\vec{v_{pre}}$ とあるキーワード a の文書ベクトル $\vec{v_a}$ のコサイン類似度は以下の (3) 式のように求める．

$$\frac{\vec{v_{pre}} \cdot \vec{v_a}}{|\vec{v_{pre}}||\vec{v_a}|} = \frac{\sum^{k} tfidf_{pre} \cdot tf_a}{\sqrt{\sum^{k} tfidf_{pre}^2} \cdot \sqrt{\sum^{k} tf_a^2}} \quad (3)$$

仮認識文の文書ベクトルの要素には $tfidf$，あるキーワードの文書ベクトルの要素には tf を用いた．これにより，仮認識文に似た内容のテキストをダウンロードしてくる単語を考慮することができる．しかし本来は，仮認識文ではなく，正解文に近い内容のテキストをダウンロードすることが理想である．ここで，正解文の文書ベクトルを用いた場合と，平均単語

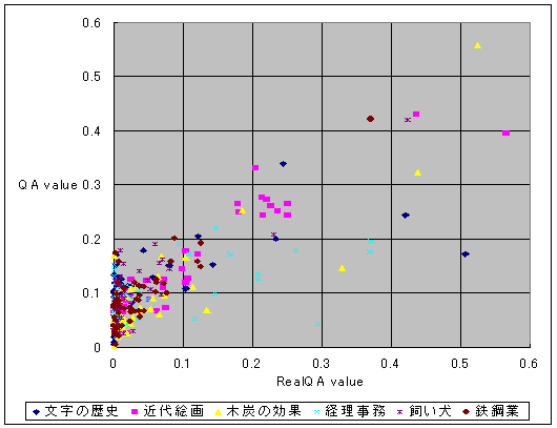

図 3 文書ベクトルの違いによる類似度の相関

Fig. 3 Correlation of Smilarity from Difference of Document Vector

正解精度 43.18 %の仮認識文の文書ベクトルを用いた場合の，キーワードの文書ベクトルとの類似度の相関を図 3 に示す．X 軸が正解文との類似度，Y 軸が仮認識文との類似度である．

この結果，互いに平均 0.78 の正の相関があり，仮認識文の文書ベクトルを用いる場合でも正解文の文書ベクトルを用いる場合に近い結果が得られることが分かる．したがって，正解文に近い内容のテキストをダウンロードする単語であることを十分考慮できると言える．

3.3.2 キーワードの文書ベクトル間の類似度の考慮

仮認識文の文書ベクトルと比較することで，類似度上位のキーワードの文書ベクトルは，仮認識文の文書ベクトルと近くなっていると言える．

しかし，仮認識文の文書ベクトルと近いからといって，キーワードの文書ベクトル間の類似度も高いとは言えない．それぞれのキーワードの文書ベクトルは，仮認識文の文書ベクトルの異なった領域との類似性を考慮されていることが考えられるからである．したがって，同じような領域との類似性が考慮されているキーワードをグループ化することを考える．

ここで，あるキーワード a の文書ベクトル $\vec{v_a}$ とあるキーワード b の文書ベクトル $\vec{v_b}$ のコサイン類似度は次の (4) 式によって求める．

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

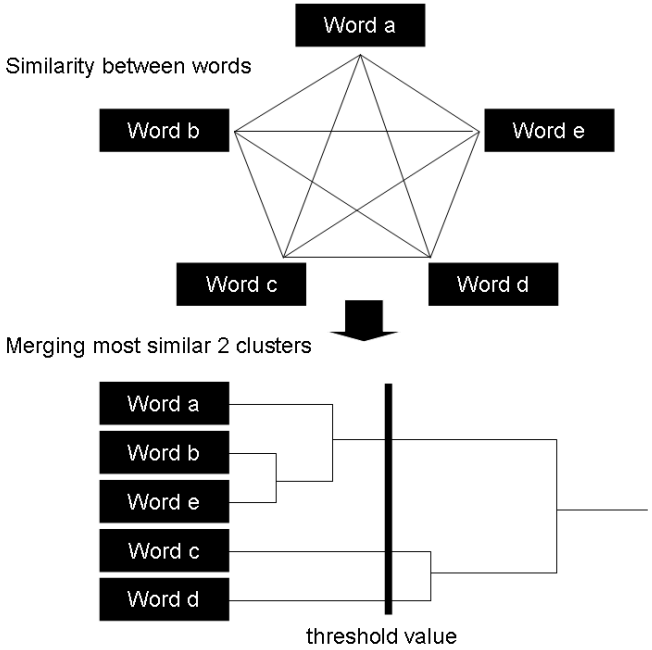

図 4 類似度によるクラスタリング
Fig. 4 Clustering using Similarity

$$\frac{\vec{v_a}\cdot\vec{v_b}}{|\vec{v_a}||\vec{v_b}|}=\frac{\sum^{k} tf_a\cdot tf_b}{\sqrt{\sum^{k} tf_a^2}\cdot\sqrt{\sum^{k} tf_b^2}} \tag{4}$$

このキーワードの文書ベクトル間の類似度を指標として，M 個の単語をいくつかのグループにクラスタリングする方法を考える．これにより，ダウンロードするテキストが似ている単語同士をグループ化できる．

クラスタリングは，凝集的階層化クラスタリングで行う (図 4)．以下手順を示す．

s1 M 個の単語に対して，2 単語間の類似度を全て求める．
s2 M 個の単語をそれぞれ個別のクラスタとする．
s3 最も類似度の高い 2 クラスタを結合した時に，新たなクラスタ内の全ての単語の共起検索ヒット数が閾値以上となるならば結合する．閾値以下ならば終了．(なお，今回は検索ヒット数が 1000 を閾値としている．)
s4 新たなクラスタと既存クラスタの類似度は最遠近傍値をとり，**s3** に戻る．

ここで検索ヒット数 1000 を閾値としているのは，YahooAPI の 1 検索クエリあたりの最大取得ページ数が 1000 であるからである．グループ化されたキーワードは，仮認識文の文書ベクトルの同じような領域との類似性が考慮された単語同士であると言える．

各クラスタごとの重要度の指標には，各クラスタが保持している各単語の $tfidf$ の総和を用いる．クラスタの重要度が高ければ，仮認識文の文書ベクトルの領域でも，重要な領域と類似しているキーワードと考えられる．

3.4 キーワード集合を単位とした従来法と提案法の比較

ここで従来法と提案法の比較を行う．キーワード選択の質を考えるためにキーワード集合 (キーワードとして扱う単語の集まり) を単位として比較する．キーワード集合には次の 4 種類を用いる．

- **従来法** $tfidf$ 上位 15 単語
- **提案法 (途中)** キーワードの文書ベクトルと，仮認識文の文書ベクトル間の類似度上位 15 単語
- **提案法 (最終 1)** クラスタリング後，保持 $tfidf$1 位クラスタの単語
- **提案法 (最終 2)** クラスタリング後，保持 $tfidf$2 位クラスタの単語

3.4.1 認識誤り単語率による評価

梶浦らによると構成する検索クエリの中に認識誤りの単語があると，単語認識精度の低下を招くと報告されている[8]．したがって，キーワード集合に対して，集合中の認識誤り単語の割合を調べた．その結果を図 5 に示す．

キーワードの文書ベクトルと仮認識文の文書ベクトル間の類似度を考慮することで，認識誤りの単語が大きく減っているのが分かる．さらにキーワードの文書ベクトル間の類似度でクラスタリングすることで，仮認識文の文書ベクトルの重要な領域と類似していると思われるクラスタ内には，認識誤りの単語がないことが分かる．

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

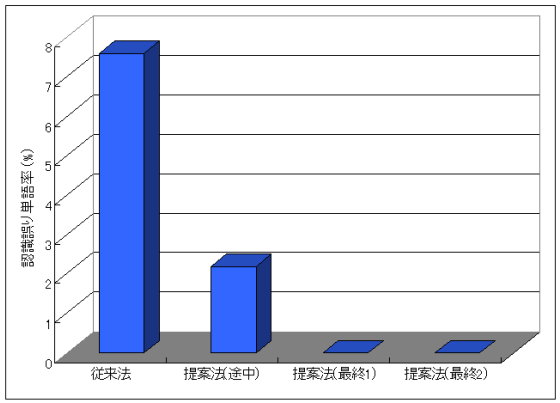

図 5 認識誤り単語の割合
Fig. 5 Rate of Mistaking Recognition Word

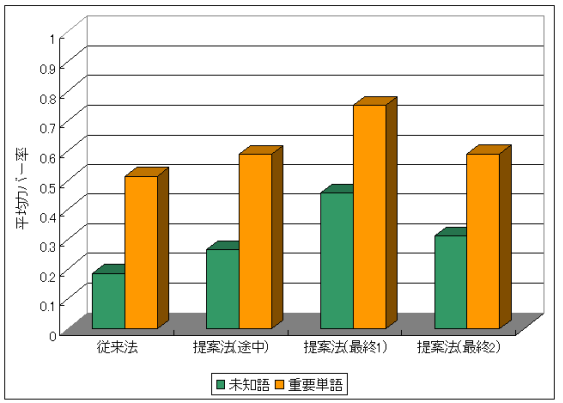

図 6 キーワード集合の平均カバー率
Fig. 6 Average Cover Rate of Key Words

3.4.2 カバー率による評価

次に，提案法及び従来法のキーワード集合の各単語を単独クエリとしてダウンロードしたとき，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[9]．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．また，正解テキストに含まれる未知語の集合を 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．

単語集合 W について，クエリ q のカバー率を以下の (5) 式のように定義する．

$$C(q,W) = \frac{|V(q) \cap W|}{|W|} \tag{5}$$

また，キーワード集合 V_k による平均カバー率を

$$C(V_k,W) = \frac{1}{|V_k|} \sum_{w \in V_k} C(w,W) \tag{6}$$

とする．このとき，$C(V_k,U)$ を平均未知語カバー率，$C(V_k,V_I)$ を平均重要単語カバー率と呼ぶことにする．

各キーワード集合に対して，未知語と重要単語の平均カバー率を調べた．その結果を図 6 に示す．なお 1 検索クエリあたり 1000 ページのダウンロードを行っている．

この結果から，キーワードの文書ベクトルと仮認識文の文書ベクトル間の類似度を考慮することで，キーワード集合の平均カバー率が上昇していることが分かる．つまりキーワード集合の中に，重要なキーワードが増え，そうでないキーワードが減ったと言える．さらにキーワードの文書ベクトル間の類似度でクラスタリングすることで，仮認識文の文書ベクトルの重要な領域と類似していると思われるクラスタのキーワード集合には，重要なキーワードのみが残っていると言える．このように，ある単語がダウンロードするテキストの文書ベクトルを利用することで，従来法よりも重要なキーワード選択を行うことができる．

4. キーワードからの有効な検索クエリ構成

提案法により有効なキーワード選択を行うことができた．さらに，認識対象に特化したページをダウンロードするために，キーワードから有効な検索クエリを構成する方法を考える．ここでは，提案法で選択するキーワードからの有効な検索クエリ構成について検討する．

4.1 提案法のキーワード選択からの有効な検索クエリ構成

提案法では，最終的に仮認識文の文書ベクトルの同様な領域との類似性が考慮されたキーワードがグループ化されている．このように似た内容のテキストをダウンロードするようなキーワードを組み合わせて検索クエリを構成する際，どのような検索クエリを構成することが有効なのかを考える．提案法の最終的なクラスタ単位での検索クエリの構成方法として，

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

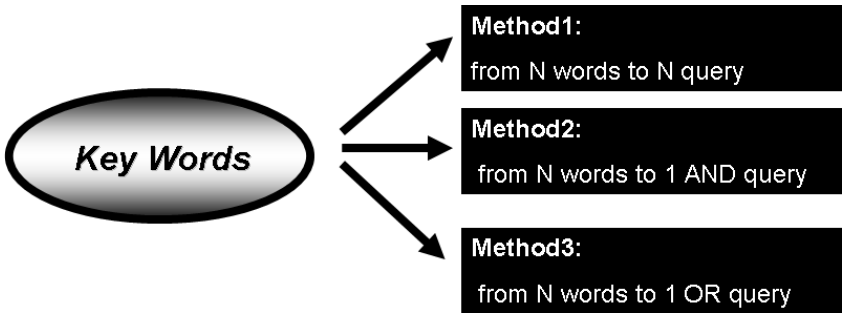

図 7 検索クエリ構成の単純な方法
Fig. 7 Simple Method of Composing Search Query

以下の単純な 3 つの方法[10] について検討を行う (図 7).

- **Method1** N 個のキーワードそれぞれで 1 単語検索クエリを構成し，各クエリで等しいページ数ずつテキストをダウンロードし，合計 1000 ページとする．
- **Method2** N 個のキーワードで 1 個の AND 検索クエリを構成し，1000 ページダウンロードする．
- **Method3** N 個のキーワードで 1 個の OR 検索クエリを構成し，1000 ページダウンロードする．

この 3 つの方法を，提案法の最終的なクラスタの保持 $tfidf$ 上位 2 クラスタに対して行った場合の，(5) 式による未知語カバー率 $C(q,U)$，重要単語カバー率 $C(q,V_I)$ を調べた．その結果を図 8 に示す．

この結果から **Method1** と **Method3** ではあまり結果が変化しないということが分かる．つまり OR 検索クエリは，ほとんど 1 単語の検索クエリを用いる場合と変わらないと言える．また，**Method2** の AND 検索クエリを構成する場合に一番有効であることが分かる．つまり，似たようなテキストをダウンロードするキーワードから検索クエリを構成する場合，AND で組み合わせることが有用な検索クエリ構成であると言える．

4.2 各検索クエリ構成の比較

検索クエリの構成方法として，従来法のキーワード選択を利用して $tfidf$ 上位順に組み合わせる方法が一般的によく用いられる[11]．しかし検索クエリを構成する際，認識誤り単語

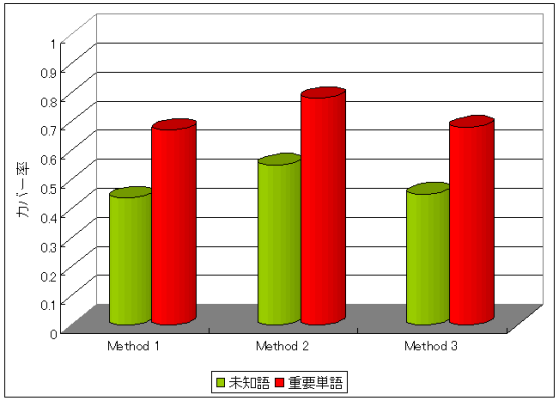

図 8 各方法のカバー率
Fig. 8 Cover Rate of Each Method

も組み合わせてしまう可能性があることが問題であった．

したがって，本研究では提案法のキーワード選択からの各クラスタ単位での AND 検索クエリを用いた検索クエリ構成を考える．認識対象の一番重要な部分を獲得できる可能性が高い，保持 $tfidf$1 位のクラスタを用いる場合が一番有効であると思われる．しかし 1 位以下のクラスタを用いる場合も，1 位のクラスタとは別領域で重要な部分を獲得できることが考えられる．よって，1 位以下のクラスタも用いて，複数の検索クエリを構成する場合についても考える．

ここで，以下のように 4 種類の検索クエリ構成する．

(a) 従来法のキーワード選択から $tfidf$ 上位順に検索ヒット数が 1000 以下になるまでの AND 検索クエリを構成
(b) 保持 $tfidf$1 位クラスタで AND 検索クエリを構成
(c) 保持 $tfidf$2 位クラスタで AND 検索クエリを構成
(d) 保持 $tfidf$1 位クラスタと保持 $tfidf$2 位クラスタでそれぞれ AND 検索クエリを作り，複数のクエリを構成

以上のそれぞれで合計 1000 ページをダウンロードした．複数のクエリを構成した場合は，

Vol.2009-NL-191 No.10
Vol.2009-SLP-76 No.10
2009/5/21

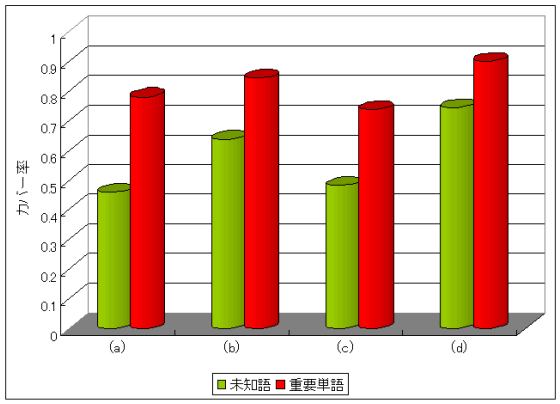

図 9 各検索クエリ構成の有用性
Fig. 9 Availability of Each Composing Search Query

等しいページ数ずつテキストを獲得する．この時の (5) 式による未知語カバー率 $C(q, U)$，重要単語カバー率 $C(q, V_I)$ を調べた．その結果を図 9 に示す．

提案法によるキーワード選択の特性を生かして AND 検索を行うことで，**(a)** の $tfidf$ 上位単語を組み合わせる場合よりも，有効な検索クエリを構成していることが分かる．また，**(d)** の保持 $tfidf$1 位クラスタと保持 $tfidf$2 位クラスタからそれぞれ AND 検索クエリを作り，複数の検索クエリを構成することで，**(b)** や **(c)** の単独クラスタのみで検索クエリを構成する場合よりもカバー率が上がっていることが分かる．つまり各 AND 検索クエリは，互いに異なる領域の重要な部分をダウンロードしていると言える．

5. ま と め

本稿では，WWW を利用した言語モデルタスク適応のための，有効なキーワード及び有効な検索クエリについて検討を行った．

キーワード選択の過程では，ある単語を検索クエリとした時の少量のテキストを利用することで，従来法よりも未知語を多く含んだテキストをダウンロードするキーワードの選択を行うことができた．キーワード集合単位では，未知語の平均カバー率が従来法よりも最大約 25 %上昇した．

さらに，提案法によるキーワード選択を行う場合は，クラスタ単位で AND 検索クエリを作ることが有用であることが確認できた．

また，保持 $tfidf$ 上位クラスタをいくつか組み合わせて，複数の検索クエリを構成することで，さらに有用となることが分かった．従来の検索クエリ構成と比較して，最大 30 %の未知語のカバー率の上昇があった．

今後は，より多くのテストセットを用いて実験することで，本方法の頑健性を調べる．また，認識対象に関連し，その未知語を含むような WWW テキストを用いた場合の，有効な言語モデル適応方法について検討を行う．

参 考 文 献

1) 伊藤 彰則，好田 正紀，“ 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
2) “ The Indexable Web is more than 11.5 billion pages ”，http://www.cs.uiowa.edu/asignori.web-size/
3) 梶浦泰智，鈴木基之，伊藤 彰則，牧野 正三，“ WWW を利用した言語モデル教師なしタスク適応の検討 ”，日本音響学会春季講演論文集，2-1-4，pp77-78，2006
4) 独立行政法人国立国語研究所，“ 日本語話し言葉コーパス ”，2004
5) 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸，“ 形態素解析システム「茶筌」 ver 2.3.3 使用説明書 ”，奈良先端科学技術大学院大学，2003
6) Yahoo! Japan，http://www.yahoo.co.jp/
7) Yahoo! developer’s network，http://developer.yahoo.com/
8) 梶浦泰智，鈴木基之，伊藤 彰則，牧野 正三，“ WWW を利用した言語モデル教師なしタスク適応における有効検索クエリ決定法 ”，電気情報通信学会技術研究報告，NLC2006-51，pp131-135，2006
9) 宇野有，伊藤仁，伊藤 彰則，牧野 正三，“ 音声ドキュメントの索引付けに向けたウェブ検索を用いたデータ収集における未知語率の検討 ”日本音響学会春季講演論文集，3-Q-30，pp275-276，2009
10) M.Suzuki，Y.Kajiura，A.Ito，S.Makino，“ Unsupervised language model adaptation based on automatic text collection from WWW ”，Proc.Interspeech，pp.2202-2205，2006
11) 翠輝久，河原達也，“ 音声対話システムのための web テキストの選択による効率的な言語モデル構築 ”，日本音響学会春季講演論文集，2-11-12，pp119-120，2006