Vol.2010-SLP-82 No.18
2010/7/24

教師なし言語モデル適応のための Web Document を用いた単語のトピック表現

増村 亮[†1] 咸 聖俊[†1] 伊藤彰則[†1]

我々は，Web 上の言語データを利用した言語モデル教師なし適応の高精度化を目指している．教師なし適応の場合，音声認識結果から話題に関連した検索クエリを作成することで Web 上から言語データをダウンロードする方法が一般的である．しかし，間接的な検索クエリを使用して未知語を含む言語データをダウンロードすることは非常に困難であった．そこで我々は，ある単語が出現する際の文脈を利用できるように，単語をクエリとして Web からダウンロードできる言語データを事前に単語と対応付ける方法を提案する．我々は形態素解析器が持つ全ての名詞に対して，事前に単語のトピックを表現した．この枠組みを利用して教師なしで適応実験を行い，本手法の有効性を確認した．

Topic Expression of Words using Web Documents for Unsupervised Language Model Adaptation

Ryo Masumura ,[†1] Seongjun Hahm [†1]
and Akinori Ito[†1]

We are developing a method of Web-based unsupervised language model adaptation. In the previous Web-based LM adaptation, search queries are composed from the automatic transcription of the input speech. However, it is difficult to gather documents that contain OOV words because the search queries do not contain any OOV words. For selecting relevant keywords from the transcription, we propose a method that associate each noun in the vocabulary with Web documents downloaded by that word. The downloaded documents are used to estimate the topic of the transcription. From the unsupervised LM adaptation method, we confirmed the effectiveness of the proposed method.

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，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大規模な言語データから学習する．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，入力のトピック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-Vocabulary word) 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あり，特に認識対象のトピックに関連した言語データを準備することで言語モデルを適応させる方法が有効である[1)]．しかし，適応のための言語データをどのように収集するかが大きな問題となる．

この言語データの収集源として近年注目されているのが，World Wide Web(以下 Web) である[2)]．Web 上には現在 1 兆以上のドキュメントが存在すると言われ，現状で最大のコーパス空間である．さらに，Web の特徴として Web 検索エンジンの存在が挙げられる．一般的に Web から必要な言語データを収集する場合，Google や Yahoo といった強力な検索エンジンが利用される．我々は，必要な言語データのトピックを表すキーワードを検索クエリとして準備するだけで言語データを収集できる．以前我々は，音声認識結果に出現した単語からトピックに関連した検索クエリを構成し，教師なしで Web から言語データを収集する方法を検討してきた[3)]．しかし，音声のトピックを表すキーワードが未知語である場合，そのキーワード以外の単語を使ってトピックに関連する文書群を取得することは困難である．この問題を解決するためには，特定のキーワードが認識結果に出現していなかったとしても，認識された文書全体から関連するドキュメントを選ぶことができる方法が必要となる．

そこで我々は，あるキーワードが認識結果に出現していなくても，認識結果とそのキーワードとの関連度が測ることができる方法を提案する．これを実現するため，単語をクエリとして Web からダウンロードできる言語データをその単語が出現する際の典型的な文脈と見なし，単語に対するトピックを事前に表現する．この工程を形態素解析器が持つ全ての名詞に対して事前に行っておく．これにより大規模なデータ群を構築でき，我々は単語と単語の関係性，そして単語とドキュメントの関係性を測ることが可能となる．したがって，教師なし言語モデル適応の枠組みでは，認識結果に出現していない単語についても認識結果と単

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

Vol.2010-SLP-82 No.18
2010/7/24

語の関係性を測ることができ，関連性の高いキーワードを選ぶことで容易に適応が可能となる．

本稿では，最初にダウンロードデータを用いて単語のトピックを表現する方法について述べ，事前ダウンロードにより実際に構築した大規模なデータ群の詳細を記す．次に教師なしの枠組みで，認識対象に関連するキーワードを選択する方法について記す．そして，実験に言語モデル適応を行うことで，本方法の有効性について検証する．

2. 単語のトピック表現

2.1 Web からのデータ収集の問題

Web 上から必要な言語データを収集する際には，一般的に Web 検索エンジンが利用される．Web 検索エンジンは，ドキュメントの内容だけでなく，リンク数やユーザの検索ログなどを総合的に考慮して，事前にドキュメントをキーワードに対応付けている．よってユーザはキーワードを準備するだけで，容易に必要な言語データの収集が可能である．しかしながら，前述のように，トピックに関連するキーワードが得られない場合，その他のキーワードだけから必要なドキュメントを集めることは難しい．

「紅茶」という単語を例として説明する ．まず,「紅茶」という単語を検索クエリとして,「ティー」という単語を含む言語データを取得するのはある程度容易であろう．これは,「ティー」が「紅茶」と高頻度で共起するからである．しかし,「紅茶」という単語を検索クエリとして，Web 上にわずか 1 万 URL 程度しかアクセス可能な言語データが存在しない「イレブンジズ[⋆1]」という単語を取得するのは非常に困難である．これは,「イレブンジズ」が「紅茶」と深く関係する単語だとしても,「紅茶」を含む言語データは Web 上に 1 億 URL 以上存在するので,「イレブンジズ」と共起したデータは 1 億と比較して非常に少量であるからだ．したがって,「イレブンジズ」を含むデータにアクセスするための現実的な解は,「イレブンジズ」という単語を直接検索クエリとすることである．

そのために,「イレブンジズ」という単語が出現する際の文脈を利用したい.「イレブンジズ」が出現する文脈は「紅茶」に関連していることが期待できるので，もしその文脈をデータとして事前に持っておくことができれば,「イレブンジズ」が「紅茶」と関連していることを推定でき,「イレブンジズ」を含むデータにアクセス可能となる．

⋆1 午前 11 時ごろに楽しむ，お茶の時間のことをいう．イギリスの一般的なティータイムのことで，アフタヌーンティーよりもカジュアル感がある．(http://d.hatena.ne.jp/keyword/より)

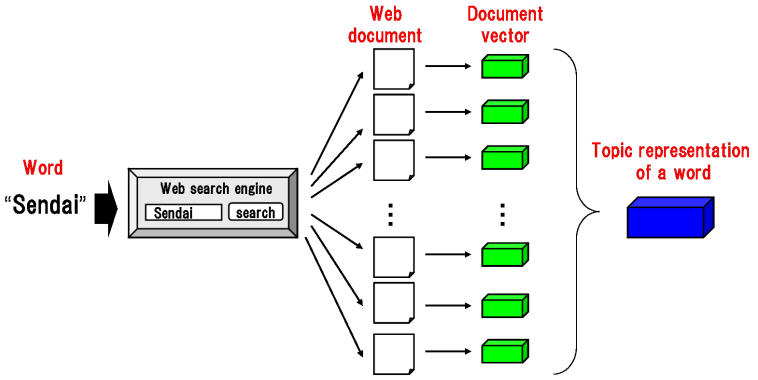

図 1 Web を利用した単語のトピック表現
Fig. 1 Topic representation of a word using Web

2.2 単語に対するトピックの表現方法

我々は，ある単語が出現する際の文脈を言語データへのアクセス時に利用できるように，単語をクエリとして Web からダウンロードできる言語データを事前に収集しておく方法を提案する．ある単語を検索クエリとした際に Web 検索エンジンから取得される Web ドキュメントは，その単語と意味的に関連が深いと考えられる．よって取得した Web ドキュメントに含まれる単語群は，その単語と共起する典型的な単語であると言える．そこで，取得される Web ドキュメントの文書ベクトルを，その単語のトピックを表現するために利用する (図 1)．

ドキュメントに対する単語の重要度には $tfidf$ を使用する．あるドキュメント D における単語 w の重要度 $T_D(w)$ を以下のように表現する．

$$T_D(w) = tf_D(w) \cdot \log \frac{N}{df(w)} \tag{1}$$

$tf_D(w)$ はドキュメント D における単語 w の出現頻度，$df(w)$ は単語 w の一般的な文書群での大局的な出現頻度であり，ここでは，単語 w でヒット可能な言語データの総数を $df(w)$ とした．N は Web 検索でヒット可能な全ての言語データの総数である．

任意のドキュメントは，ベクトル空間モデルを考えることで文書ベクトルとして表現できる．ある単語 W を検索クエリとした際に Web 検索エンジンから取得される i 番目の Web ドキュメント D_W^i を $tfidf$ で重み付けした文書ベクトル $\boldsymbol{f}_i(W)$ を以下のように定義する．

$$\boldsymbol{f}_i(W) = [T_{D^i_W}(w_1), ..., T_{D^i_W}(w_k)]^\top \quad (2)$$

そして，ある単語 W を検索クエリとした際に Web 検索エンジンから取得される上位 T ページを使用して，ある単語 W のトピックを表す特徴ベクトル $\boldsymbol{F}(W)$ を以下のように定義する．

$$\boldsymbol{F}(W) = \sum_{i=1}^{T} \frac{\boldsymbol{f}_i(W)}{|\boldsymbol{f}_i(W)|} \quad (3)$$

この特徴ベクトル $\boldsymbol{F}(W)$ は，ある単語の特徴を単語群で表現しており，単語 W が出現する際の，ドキュメントの典型的な文書ベクトルを表現していると言える．我々はこの特徴を用いることで，単語と単語の関係性，そして任意のドキュメントと単語の関係性を測ることが可能となる．

2.3 事前ダウンロードによるデータ群の構築

大語彙空間に対して本方法を適用するために，形態素解析器が持つ全ての名詞に対して事前にデータをダウンロードして単語のトピックを表現する．音声認識や自然言語処理の分野では Web を利用した大規模な試みは増加していて，Web 資源から語彙情報の自動収集を行うことで，語彙データを集約する試みも行われている[4),5)]．我々はまず，大語彙空間として ipadic[⋆2] と unidic[⋆3]の混合により構築した形態素解析器の辞書に含まれる名詞 287715 単語を対象として，この各単語に対して検索エンジンを利用して Web ドキュメントを取得した．本研究では，検索エンジンとして Yahoo! Japan 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[6)] 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．

我々は事前に 287715 単語それぞれに対して，Yahoo! Japan の検索エンジンでヒットする言語データの数を調べた．ヒットする言語データ数を上位順にソートした結果を図 2 に示す． 287715 単語のうち最も多く言語データがヒットした単語は「www」であり，約 480 億 URL の言語データが存在した．しかしながら，Web 上にさえあまり出現しないような単語も多く存在することが分かる．1 万 URL 以下しか言語データが存在しない単語は約 5 万，そして 1000URL 以下しか言語データが存在しない単語は約 2 万あった．このような単語を含む言語データを間接的な検索クエリを用いて取得するのは困難であり，やはり単語を直接検索クエリに用いなければその単語を含む言語データを獲得することは難しいと言える．

⋆2 http://chasen.aist-nara.ac.jp/stable/ipadic/
⋆3 http://www.tokuteicorpus.jp/dist/

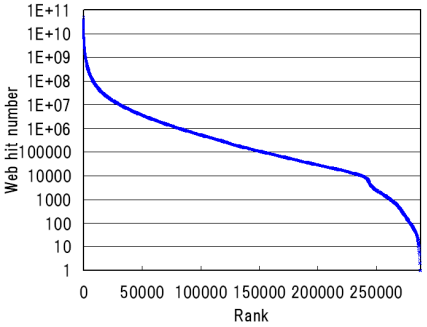

図 2 検索ヒット数
Fig. 2 Web hit number

表 1 構築した言語データ群の詳細
Table 1 Detail of constructed text data constellation

対象キーワード数	名詞 287715 単語
キーワードごとのダウンロード URL 数	50URL
ダウンロードを行った期間	2010 年 2 月から同年 4 月の約 3 カ月間
言語データ群の総形態素数	約 1406 億
言語データ群の総形態素種類数	395809 (読み付与可能な単語のみ)

我々は各単語からそれぞれ 50URL のドキュメントを Web 検索エンジンを用いて事前にダウンロードした．各 Web ドキュメントは HTML 形式でダウンロードされるので，タグを除去してテキスト部分のみを抽出し，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った[7)]．そして各 Web ドキュメントに対して，chasen を用いて形態素解析を行い，ドキュメントに含まれる名詞から文書ベクトルを構成し，(3) 式の T の値を 50 として各単語の特徴ベクトルを作成した．この特徴ベクトルは名詞の数だけ次元が存在するので，287715 単語をそれぞれ 287715 次元のベクトルで表している．

以上ような工程で構築した言語データ群の詳細を表 1 に示す． このデータ群を用いることで，形態素解析器が持つあらゆる名詞に対して単語と単語の関係性，そして単語とドキュメントの関係性を測ることが可能となる．さらに，目的のキーワードを含む言語データを容

易に取り出すこともできる．このような枠組みは，Web という大規模なコーパス空間から目的の言語データを選ぶための近似的な解なのではないかと我々は考える．

3. 教師なし言語モデル適応での利用

3.1 適応方法

構築した大規模データ群を教師なし言語モデル適応で利用したい．例えば「サッカー」というトピックに対して教師ありで言語モデル適応を行う場合,「サッカー」という単語に関連した単語をセレクションする問題と考える．これは構築したデータ群を用いて単語と単語の関係性を測ることで可能である．

同様に教師なし適応では，音声認識結果に関連した単語をセレクションする問題として扱うことができる．これは，以前我々が検討していたような認識結果内から重要なキーワードを選択する枠組み[3] と近い考え方であり，本方法では認識対象に出現したかどうかに関わらず，可能性のある全ての単語から重要なキーワードのセレクションを行う．このような Web 上の語彙資源を利用した言語モデルの作成は，流行語や新出語などに対しても単語の読みも含めて言語データを利用でき，非常に有効である[8),9)]．教師なし適応の流れは以下の形をとる．

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て，入力音声を認識する．

step2 構築したデータ群を用いて，音声認識結果と関連したキーワードを選択する．

step3 選択されたキーワードに対応する言語データを用いて，適応のための Web コーパスを作成する．

step4 ベースラインコーパスと作成した Web コーパスから新たに適応言語モデルを学習し，入力音声を再認識する．

3.2 キーワードの選択

先程の **step3** における音声認識結果と関連したキーワードの選択方法を記す．

まず認識結果 h を $tfidf$ で重み付けすることで，以下のような文書ベクトルとして表現する．

$$\boldsymbol{S}(h) = [T_h(w_1), ..., T_h(w_k)]^\top \tag{4}$$

この文書ベクトルは認識対象のトピックを表現した文書ベクトルである．これを用いて，ある単語 W のトピックを表現する特徴ベクトル $\boldsymbol{F}(W)$ と音声認識結果 h の関係性を以下の式から算出する．

$$f(W,h) = \frac{\boldsymbol{F}(W)^\top \boldsymbol{S}(h)}{|\boldsymbol{F}(W)|} \tag{5}$$

これを形態素解析器の辞書が持つ全ての名詞 287715 単語に対して求めることで，認識対象への関連性の順位が計算できる．この上位単語を認識対象に出現し得る単語とみなし，キーワードとして選択する．

表 2 実験条件の詳細
Table 2 Detail of experimental condition

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
ベースラインコーパス	CSJ よりテストセット以外の 2536 講演
総形態素数	7652534
総形態素種類数	57970
ユニグラムエントリ数	41695 (カットオフ:1)

4. 実　　験

4.1 実験条件

我々は，本稿で提案した方法の有効性を調査するために，実際の音声データに対して言語モデル教師なし適応を行う．実験のためのテストセットとして，CSJ(Corpus of Spontaneous Japanese) の「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 40 講演を用いる．我々はまず，ベースラインコーパスからベースライン言語モデルを作成した．ベースライン言語モデルの詳細を表 2 に示す． 音声認識デコーダとして Julius 4.1.2, 音響モデルとして CSJ 付属状態共有 triphone モデルを用いて，40 講演に対して初期の音声認識を行った．その結果，40 講演の平均単語正解精度は 62.45%，補正パープレキシティは 101.5，未知語率は 1.85%であった．

4.2 実験結果

まず 40 講演それぞれに対して，認識結果と単語の関係性を調べて順位を計算した．例として,「脳卒中の話」を対象とした講演音声の自動書き起こし (単語正解精度 64.87%) に対する関連度上位 30 単語を，認識できた単語，未知語，入力音声には出現しない単語に分けて表 3 に示す． この結果から，未知語であってもキーワードとして選ぶことができることが確認できる．また入力音声には出現しない単語も，このトピックで出現してもおかしくない単語であることが分かる．

Vol.2010-SLP-82 No.18
2010/7/24

表 3 選択したキーワードの例
Table 3 Example of selected keywords

認識できた単語	脳卒中，脳梗塞 (脳こうそく，脳硬塞)，血圧，高血圧，脳，出血 (しゅっけつ)
未知語	卒中，脳溢血 (脳いっ血)，脳出血，脳塞栓，脳血栓
入力音声には出現しない単語	梗塞，脳軟化症，溢血，前触れ，前ぶれ，内出血，本態，こうけつ 内しゅっ血，一過，脈圧，卒然，立ちくらみ，硬塞，降圧

表 4 実験結果
Table 4 Experimental result

Number	Acc (%)	App	OOV (%)	Recall (%)
baseline	62.45	101.5	1.85	3.2
10	66.54	88.84	0.53	5.7
20	66.03	85.35	0.43	9.0
50	66.63	82.21	0.34	12.8
100	67.04	80.06	0.32	16.5
200	67.06	80.42	0.28	22.6
500	66.63	83.96	0.28	26.3
1000	66.21	88.31	0.3	32.0

次にキーワードとして選択する単語数を変化させて言語モデル適応を行った．Web コーパスの作成には，選択されたキーワードに対応する全ての Web ドキュメントを用いた．また適応言語モデルは，ベースラインコーパスと Web コーパスを 1:1 で足し合わせて学習した．なお言語モデルの語彙には，選択したキーワードを優先的に選択し，残りは足し合わせたコーパスの単語出現頻度から最大 55000 単語を選択した．その他の実験条件は表 2 と同様である．キーワードの数を変化させた際の単語正解精度，補正パープレキシティ，未知語率，そして，認識結果には出現しなかった未知語を直接キーワードとして再現できた割合を表 4 に示す．

この結果から，適応により単語正解精度，補正パープレキシティ，および未知語率が大きく改善していることが分かる．しかしキーワードを多く選択してしまうと関連性が低い単語も多くなってしまい，性能が上がらないことが分かる．また，多くの未知語を実際にキーワードとして再現できることが確認できる．適応前と比較して，上位 100 単語をキーワードに用いた際に単語正解精度を約 4.6%改善，補正パープレキシティを約 21.4%改善した．また上位 200 単語をキーワードに用いた際に未知語率を約 1.6%改善した．

さらに，大語彙空間からキーワードを選択する場合のメリットを調べるために，次の 3 条件の場合で比較を行う．

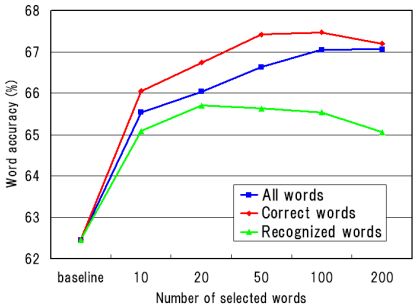

図 3 単語正解精度の結果
Fig. 3 Result of word accuracy

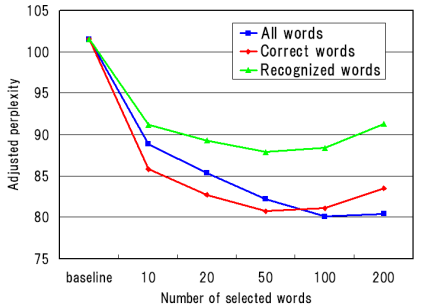

図 4 補正パープレキシティの結果
Fig. 4 Result of adjusted perplexity

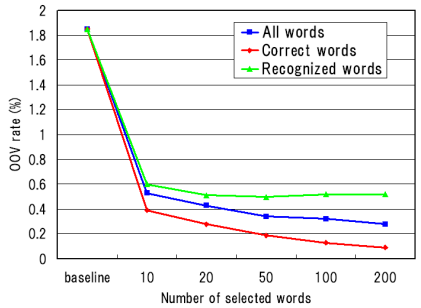

図 5 未知語率の結果
Fig. 5 Result of OOV rate

- **All words:** 大語彙空間の全ての名詞からキーワードを選択する場合 (表 4 と同様).
- **Correct words:** 正解文に存在する名詞のみからキーワードを選択する場合.
- **Recognized words:** 認識文に存在する名詞のみからキーワードを選択する場合.

各条件で音声認識結果と関連性の高い単語を選択した．テストセットの 40 講演はいずれも名詞が 200 種類程度含まれているので，最大 200 単語を対象にキーワードの数を変化させて言語モデル適応を行った場合の単語正解精度の結果を図 3 に，補正パープレキシティの結果を図 4 に，未知語率の結果を図 5 に示す．

この結果から，従来法である認識結果の単語のみをキーワードに選択する場合と比べて，

表 5 取得した未知語の考察
Table 5 Consideration of retrieved OOV words

Web 検索エンジンで ヒット可能な言語データの数	テストセットの 未知語の総数	従来法で取得した 未知語の数	提案法で取得した 未知語の数
100 億 URL 以上	1	0	0
10 億 URL 以上 100 億 URL 以下	1	0	0
1 億 URL 以上 10 億 URL 以下	17	11	13
1000 万 URL 以上 1 億 URL 以下	181	140	152
100 万 URL 以上 1000 万 URL 以下	328	217	268
10 万 URL 以上 100 万 URL 以下	121	47	84
1 万 URL 以上 10 万 URL 以下	52	10	35
1000URL 以上 1 万 URL 以下	7	0	4
1000URL 以下	8	0	4

大語彙空間からキーワードを選択する場合の方が高い適応効果を示していることが分かる．また，実際に正解文に出現した単語をキーワードに選択する場合と大語彙空間からキーワードを選択する場合では，単語認識精度や補正パープレキシティの結果が大差ないことが分かる．ただし未知語率に関しては，正解文に出現した単語を用いると未知語を必ずキーワードに選択できるため，大語彙空間からキーワードを選択する場合よりも大きく改善している．なお，単語認識精度や補正パープレキシティの値が正解文や認識文で 200 単語をキーワードに選択した場合に低下してしまうのは，認識対象と関連しているとは言えない汎用的な単語がキーワードになるためである．

ここで名詞の未知語に着目して考察する．40 講演の未知語の種類数は全部で 839 種類あったが，9 割近くである 735 種類が名詞の未知語であった．そこで我々は，Web 検索エンジンでヒット可能な言語データの総数で未知語の種類を区分けして，40 講演に含まれる名詞の未知語の数，さらに従来のような認識結果に出現した単語から上位 200 単語をキーワードを選択する場合に取得できた未知語の数，および提案した大語彙空間から上位 200 単語をキーワードを選択する場合に取得できた未知語の数を調べた．その結果を表 5 に示す．

この結果から，どの種類の未知語も提案法で多く取得できていることが分かる．さらに Web 上に言語データがあまり存在しない未知語に対しては，提案法が有効であることが顕著である．特に従来法では 1 万 URL 以下しか言語データが存在しないような単語は取得できていなかったが，提案法では Web 上でわずか 382URL しか存在しないような未知語も取得できていた．このことから，提案法が未知語の取得に優れていることが分かる．

5. まとめと今後の課題

本稿では，ある単語が出現する際の文脈を言語データへのアクセス時に利用できるように，単語をクエリとして Web からダウンロードできる言語データを事前に単語と対応付ける方法を提案した．そのために，形態素解析器が持つ全ての名詞に対して事前に言語データをダウンロードして各単語との対応付けを行い大規模なデータ群を構築した．そして，構築したデータ群を用いて教師なし言語モデル適応を行った．提案法は，従来の音声認識結果に出現した単語をキーワードとする方法よりも性能を大きく改善し，従来の枠組みでは取得できない未知語を多く取得できることが分かった．今後は，教師なし言語モデル適応の枠組みでさらなる実験，評価を行い，キーワードの選択方法についても検討を行う．さらに構築したデータ群を用いて，トピック言語モデルとしての表現を模索したい．

参 考 文 献

1) A.Ito, H.Saitoh, M.Katoh and M.Kohda, “ N-gram language model adaptation using small corpus for spoken dialog recognition ”, In Proc.Eurospeech, pp.2735-2738, 1997.
2) D.Vaufreydaz, M.Akbar and J. Rouillard, “ Internet documents: A rich source for spoken language modeling ”, In Proc.Workshop ASRU, pp.277-280, 1999.
3) 増村亮, 伊藤仁, 伊藤彰則, 牧野正三, “ Web 検索結果を利用したトピック関連語推定に基づく言語モデルの教師なし適応 ”, 日本音響学会春季講演論文集, 2-6-3, 2010.
4) 赤峯享, 加藤義清, 河原大輔, レオン末松豊インティ, 新里圭司, 乾健太郎, 黒橋禎夫, 木俵豊, “Web 情報分析のための大規模 Web ページの収集・選択・検索”, 言語処理学会第 16 回年次大会発表論文集, pp.238-241, 2010.
5) 中野 鐵兵, 佐々木浩, 藤江真也, 小林哲則, “集合知を利用した語彙情報の収集・共有・管理システム”, 情報処理学会研究報告, Vol.2008-SLP-71-12, pp.77-84, 2008.
6) Yahoo! Japan Developers Network, http://developer.yahoo.co.jp/.
7) 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.
8) 松原勇介, 緒方淳, 後藤真孝, “ポッドキャスト音声認識の性能向上手法：集合知によって更新される Web キーワードを活用した言語モデリング”, 情報処理学会研究報告, Vol.2008-SLP-71-6, pp.39-44, 2008.
9) 佐々木浩, 中野 鐵兵, 緒方淳, 後藤真孝, 小林哲則, “集合知に基づく語彙情報を用いたトピック依存言語モデリング”, 情報処理学会研究報告, Vol.2009-SLP-75-11, pp.57-62, 2009.