社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＨＩ－109　(20)
2004－SLP－ 52　(20)
2004／7／17

関連語彙獲得に基づく認識辞書のオフライン教師なし適応

廣嶋 伸章[†]　大附 克年[‡]　林 良彦[§]
[†] 日本電信電話株式会社 NTTサイバーソリューション研究所
[‡] 日本電信電話株式会社 NTTサイバースペース研究所
[§] 大阪大学大学院 言語文化研究科

音声認識では認識辞書に含まれない単語は認識できないという、いわゆる未登録語の問題があるが、認識結果の内容に関連する語彙を獲得して認識辞書に追加することにより入力音声に適応した未登録語の少ない辞書を作成することができ、その辞書を用いて再度認識を行うことにより未登録語の影響を抑えて認識精度を改善できると考えられる。そこで本稿では、音声認識結果の内容に関連する語彙をコーパスから獲得し、獲得した語彙を辞書に追加することによって辞書を入力音声に適応させる手法を提案する。提案手法は、テキストコーパス中の語彙に対して語彙の分野を表す語彙分野ベクトルを算出しておき、入力の認識結果に対して発声内容の分野を推定し、その分野に近い語彙分野ベクトルを持つ語彙を入力に対する関連語彙として獲得し辞書に追加するという処理をオフラインで行う教師なし適応手法である。毎日新聞コーパスから各語彙の語彙分野ベクトルを求め、TV ニュース音声を用いて提案手法の評価を行った。

Off-line Unsupervised Vocabulary Adaptation based on Relevant Word Acquisition

Nobuaki Hiroshima[†]　Katsutoshi Ohtsuki[‡]　Yoshihiko Hayashi[§]
[†] NTT Cyber Solutions Laboratories, NTT Corporation
[‡] NTT Cyber Space Laboratories, NTT Corporation
[§] Graduate School of Language and Culture, Osaka University

One of the most common problems in speech recognition system is the out-of-vocabulary (OOV) problem. Although we cannot avoid that input data includes OOV words for a vocabulary, we can reduce the number of them by adapting the vocabulary to inputs. Extracting relevant words to the content of an input speech based on a speech recognition result obtained using a reference vocabulary and adding them to the vocabulary enable to build an expanded vocabulary that includes less OOV words. The second recognition process using the new vocabulary is supposed to be performed better than the first process.

In this paper, we propose vocabulary adaptation that acquires relevant words to an input from corpus and builds an expanded vocabulary by adding them to the reference vocabulary.

1. はじめに

音声認識における問題点の一つに、認識辞書の語彙に含まれない単語が認識対象の音声中に出現するために認識に失敗するという、いわゆる未登録語の問題がある。認識辞書の語彙数を増やすことにより未登録語の数は減少するが、利用できるメモリや学習データの量に限りがあるだけでなく、むやみに語彙数を増やすと認識精度の低下を招くため、この手法により未登録語の問題を解決することは難しい。

蓄積音声の書き起こしやインデクシングといったリアルタイム処理を必要としない用途の場合には、未登録語の問題を解決するもう一つの手法として、まず基準の語彙を用いて音声認識を行い、認識結果から発声内容に関連する語彙を獲得するということが考えられる。獲得した語彙を基準の語彙に追加することにより、入力音声に適応

した未登録語の少ない認識辞書を作成することができ、語彙を追加した辞書を用いて再度認識を行うことで認識精度を改善できる。また、厳選された少量の語彙を追加するため、メモリ量などの問題も発生しない。

そこで本稿では、コーパス内に含まれる語彙の中から音声認識結果の発声内容に関連する語彙を獲得し辞書に追加することにより、辞書を入力音声に適応させる手法について報告する。認識結果に対して発声内容の関連語彙をコーパスから獲得し辞書に追加する手法について述べ、TV ニュース音声における未登録語削減の評価結果について述べる。

2. 認識辞書の教師なし適応

認識辞書の適応については、これまでにも、コーパスから関連語彙を獲得し、獲得した語彙をもとに辞書を更新するという手法がいくつか提案されてきた。どの手法も正解の認識結果を必要とせず、テキストコーパスから語彙を獲得する教師なし適応手法である。

Kemp らは、コーパス中での出現頻度の高い順に語彙を追加しながら、基準の語彙の中から頻度の低いものを削除することにより、語彙サイズを一定に保ったまま辞書を更新している[1]。関連語彙の獲得では、Okapi という尺度を用いて文書データベースから認識結果に対する関連文書を検索し、関連文書に含まれるすべての語彙を関連語彙として獲得している。

Yu らは、インターネット上の Web ページに含まれる語彙を追加することにより辞書を更新している[2]。インターネットサーチエンジンである Infoseek を利用して入力の内容に関連する文書を検索し、関連文書に含まれる語彙と入力の内容との相互情報量を用いて関連語彙を獲得している。

しかし、これらの手法はどれも、語彙を獲得する際に”関連文書の検索”と”関連文書からの語彙獲得”という２つの処理を実行しなければならないため、計算量が膨大になってしまうだけでなく、コーパス中の文書ごとに語彙の頻度や重みを保持しなければならないため、大量のメモリを必要とする。また、これらの手法では関連文書に含まれるすべての語彙を獲得するため、内容に関連のない語彙も大量に辞書に追加されてしまう。認識に必要のない語彙が辞書に追加されると、認識時における語彙の選択の幅が広がり、誤った語彙を選択して認識誤りを起こす可能性が高くなる。さらに、入力に認識誤りが含まれることが想定されていないため、認識誤りがあると正しく語彙を獲得することができない。

表 1：概念ベースの例

概念語	概念ベクトル			
	1	2	…	d
りんご	0.01	0.05	…	0.03
みかん	0.01	0.06	…	0.02
美術	0.09	0.01	…	0.08
絵画	0.08	0.02	…	0.07
…	…	…	…	…

3. 概念ベースを用いた関連語彙獲得に基づく認識辞書の教師なし適応

本稿では、次のようにして語彙の獲得を行うことで、従来手法の関連語彙獲得における問題点を解消する。まず、概念ベース[3]を用いて、コーパス中の各語彙に対する分野を表す語彙分野ベクトルを事前に求めておく。次に、入力となる認識結果の各話題に対する分野を表す話題分野ベクトルを求め、語彙分野ベクトルとの関連度の高いものを関連語彙として獲得する[4]。

語彙を獲得する際には、入力に対し、関連文書を検索することなく、語彙分野ベクトルをもとに直接語彙を獲得するため、高速な処理を行うことができる。また、文書ごとに各語彙の頻度や重みを保持する必要はなく、語彙ごとにベクトルを保持するだけでよいので、従来手法に比べて少ないメモリで済む。関連度の高い語彙だけを獲得するので、これらの語彙を追加することによる認識での悪影響も少ない。さらに、概念ベースを用いて認識結果に含まれる各単語をクラスタリングすると認識誤りの単語は大きいクラスタに含まれにくいため、最も大きいクラスタのみを用いて分野を推定することで、認識誤りの単語による影響をおさえて正しく語彙を獲得できる[5]。

以下では、提案手法で用いる概念ベースについて述べ、提案手法における関連語彙獲得とそれに基づく辞書更新の詳細について述べる。

3.1 概念ベース

概念ベースは、概念語とそれに対応する概念ベクトルとを収めたデータベースである。概念ベクトルを生成するには、まず学習用コーパスを用いて各単語（自立語）間の一文中における共起頻度から単語の共起行列を生成する。共起行列の各行に対応する単語を概念語と呼び、各列に対応する単語を文脈生成単語と呼ぶ。共起行列の各行が各概念語に対する共起パターンのベクトルとなる。ベクトルの次元数の圧縮とデータスパースネスの

解消のために特異値分解（SVD）により行列を変換したのち、長さ 1 に正規化したものが概念ベクトルとなる。概念ベクトルは単語の共起傾向をベクトル表現したものであり、概念ベクトルが近い単語同士は関連が高いと考えられる。概念ベースの例を表 1 に示す。「りんご」と「みかん」、「美術」と「絵画」のように関連の高い単語の概念ベクトルは近いものとなる。

3.2 関連語彙獲得

提案手法では、入力となる記事の認識結果に対して、発声内容に含まれる話題の分野を推定し、その分野に近い語彙を関連語彙として獲得する。ここで話題とは、入力の記事を構成する単位のことであると定義する。文や段落などが話題に相当し、記事自体を話題とすることもできる。文単位のように記事が複数の話題からなる場合には、各話題から関連語彙を獲得し、全ての話題に関する関連語彙の中から記事に関する関連語彙を選別する。

話題の分野を推定するためには、話題の分野を表す話題分野ベクトルを求める必要がある。また、語彙が発声内容の分野と近いかどうかを判定するためには、各語彙に対し、あらかじめ語彙の分野を表す語彙分野ベクトルを算出しておく必要がある。語彙分野ベクトルは、コーパス中の各話題における話題分野ベクトルをもとに算出される。

以下では、話題分野ベクトルおよび語彙分野ベクトルを算出するためのアルゴリズムと、それらを用いて話題から関連語彙を獲得するためのアルゴリズムについて述べ、話題に関する関連語彙の中から記事に関する関連語彙を選別する方法について述べる。

3.2.1 話題分野ベクトルの算出

記事中の話題には様々な概念を持つ単語が含まれているが、その話題の分野に関する概念を持つ単語は、話題内で数多く出現すると考えられる。そこで、話題中に出現する概念語を概念ベクトルに基づいてクラスタリングし、生成された複数のクラスタの中で最も大きいクラスタが分野を表しているクラスタであるとみなして、そのクラスタに含まれている概念語の概念ベクトルの重心を話題分野ベクトルとする。

$$\vec{v}_{topic}(t) = \frac{1}{N(C_{\max})} \sum_{w \in C_{\max}} \vec{v}_c(w)$$

$$C_{\max} = \arg\max_{C \in \Omega} N(C) \qquad \cdots(1)$$

$\vec{v}_{topic}(t)$：話題 t の話題分野ベクトル

$\vec{v}_c(w)$：概念語 w の概念ベクトル

$N(C)$：クラスタ C の概念語数

Ω：クラスタリングにより生成されたクラスタの集合

このように、最も大きいクラスタだけを話題分野ベクトルの算出に用いることで、話題の分野を表していない概念語の影響を抑えることができる。また、認識誤りとなった単語は正しく認識された単語と異なる概念を持つ傾向にあるため、それらの単語もクラスタリングにより取り除くことができる。

3.2.2 語彙分野ベクトルの算出

表 1 の「りんご」と「みかん」の例のように、同じ分野の概念語であれば類似した概念ベクトルを持つので、この概念ベクトルを語彙の分野を表すベクトルと考えても差し支えない。しかし、概念ベクトルは文中の概念語と他の単語との共起頻度をもとに作成されるので、概念語自体の出現頻度が低い場合はあまり有効な概念ベクトルとならず、そのため比較的頻度の高い単語のみを概念語として概念ベースを作成するのが一般的である。一方、本稿で獲得したい語彙は高頻度語ではなく、認識辞書に出現しないような低頻度語であることが多いので、概念ベースに含まれる概念語から語彙を獲得してもあまり有益であるとはいえない。そこで、学習コーパス中の全語彙に対し、その語彙が出現する各話題における話題分野ベクトルの重心を語彙分野ベクトルとする。

$$\vec{v}_{term}(w) = \frac{1}{Z} \sum_{t \in T} \delta(w,t) \vec{v}_{topic}(t)$$

$$Z = \sum_{t \in T} \delta(w,t)$$

$$\delta(w,t) = \begin{cases} 1 & if\ w \in t \\ 0 & otherwise \end{cases}$$

$\vec{v}_{term}(w)$：語彙 w の語彙分野ベクトル

T：コーパスに含まれる話題の集合

このようにして語彙分野ベクトルを求めることで、ある語彙がコーパス中の話題に一度しか出現しなかった場合でも、その話題の話題分野ベクトルを用いて語彙の分野を正しく表すことができる。

3.2.3 関連度の算出

3.2.2 節のアルゴリズムによってあらかじめ求めておいたコーパス中の各語彙に対する語彙分野ベクトルと、3.2.1 節のアルゴリズムによって求めた入力の認識結果の話題に対する話題分野ベクトルを用いて関連語彙を獲得するためには、各語彙がどの程度話題に関連しているかを表す関連度を算出する必要がある。各語彙の関連度は、その語彙の語彙分野ベクトルと入力の話題分野ベクトルとのコサイン距離を計算することによって求める。

$$rel(w,t) = \frac{\vec{v}_{term}(w) \cdot \vec{v}_{topic}(t)}{\left|\vec{v}_{term}(w)\right|\left|\vec{v}_{topic}(t)\right|}$$

$rel(w,t)$：話題 t に対する語彙 w の関連度

このようにして関連度を算出し、関連度の大きい順に上位 N 個の語彙を関連語彙として獲得する。

3.2.4 記事からの関連語彙獲得

これまでは、話題は文や段落というような記事の構成単位であると定義し、ある話題に関連する語彙を獲得する方法について述べてきた。しかし、実際の音声は話題ごとに存在することは少なく、記事ごとに存在するか、あるいは複数の記事について述べられたニュース番組ごとに存在することが多い。ニュース番組は、トピックセグメンテーション[6]などにより記事に分割することができる。そのため、記事に相当する音声の認識結果に対して関連語彙を獲得することが望まれる。

記事自体を 1 つの話題であるとした場合は、話題に関する関連語彙をそのまま記事に関する関連語彙であるとすればよい。文単位のように記事が複数の話題からなるものである場合には、それぞれの話題に対して関連語彙を獲得することができる。このとき、記事に対する関連語彙を以下のようにして獲得する。

(1) 全話題から得られた関連語彙をマージ
(2) 関連語彙を関連度の大きい順にソート
(3) 関連度の大きい順に上位Ｎ個の関連語彙を獲得

(1)において複数の話題から同一の語彙が獲得された場合には、その中で関連度が最大となるものを残し、それ以外を削除する。

3.3 認識辞書の更新

Kemp らの方法では、獲得した語彙を追加しながら、基準の語彙の中で頻度の低いものを削除することで、語彙サイズを一定に保ったまま認識辞書の更新を行っているが、頻度の低い語彙が認識に不要であるという保証はまったくない。また、語彙分野ベクトルを用いれば、本手法により基準の語彙に対する関連度を算出し、関連度の小さい語彙を辞書から削除するということも考えられるが、この方法だとどの分野にも出現するような一般的な語彙の関連度が小さくなり、誤って削除されてしまうおそれがある。そこで、本研究においては基準の語彙の削除は行わず、獲得した語彙の追加のみを行って辞書を更新する。

追加した語彙に対する言語モデル確率としては、学習時の未登録語クラスの確率を適用し、クラス内 unigram 確率は追加した語彙数の逆数とする。

4. 評価実験

提案手法の有効性を検証するため、放送ニュース音声を用いて評価を行った。以下では実験条件について述べ、実験結果を報告する。

4.1 実験条件

学習および評価に利用したデータについて述べる。

4.1.1 学習データ

概念ベクトルの作成には新聞記事テキスト 1 年分（毎日新聞 2002 年）の見出しと本文を用いた。概念語として高頻度語約 47,000 語を用い、文脈生成単語として上位 50 語を除く高頻度語 1,000 語を用いた。概念語との共起頻度ベクトルを SVD により 100 次元に圧縮し概念ベクトルとした。上述の新聞記事テキストに出現するすべての語彙（約 16 万語）について、3.2.2 で述べた手法で語彙分野ベクトルを作成した。

4.1.2 評価データ

2002 年 12 月に放送された TV ニュース番組 30 番組を評価に用いた。評価データ全体でのトピック数は 265、発話数は 2,898、総単語数は 69,068 であった。音声認識エンジンには NTT で開発された VoiceRex[7]を使用し、ニュース番組の書き起こしなどのテキスト約 45 万文（約 1500 万語）を用いて語彙サイズ約 25,000 語（最低頻度 10 の

表 2：話題の単位に関する評価結果

話題の単位	25k		50k	
	#oov	%red.	#oov	%red.
記事	1399	4.9	683	4.1
文	**1233**	**16.1**	**583**	**18.1**

表 3：クラスタリングの有無に関する評価結果

話題の単位	クラスタリング	25k		50k	
		#oov	%red.	#oov	%red.
記事	あり	1435	2.4	697	2.1
	なし	1399	4.9	683	4.1
文	あり	1283	12.8	616	13.5
	なし	**1233**	**16.1**	**583**	**18.1**

高頻度語）と約 50,000 語（最低頻度 2 の高頻度語）の trigram を学習して認識辞書とした。学習データに対する被覆率は 25,000 語と 50,000 語の語彙でそれぞれ 99.18%、99.87%であり、評価データに対する単語誤り率は 25,000 語と 50,000 語の語彙でそれぞれ 27.5%、27.3%であった。

4.2 関連語彙獲得に関する評価

語彙サイズが 25,000 語（25k）と 50,000 語（50k）のそれぞれについて、提案手法により 100 語の関連語彙を獲得して認識辞書の語彙に追加することで、未登録語がどの程度削減されるかの評価を行った。記事ごとに獲得した語彙を追加して未登録語数を求め、その合計を求めた。語彙を追加しない場合の未登録語数は、25,000 語のとき 1471 語、50,000 語のとき 712 語であった。

4.2.1 話題の単位に関する評価

まず、話題の単位に関する評価を行った。3. 2. 4 節で述べたように、入力記事に対する話題の単位には記事、段落、文など様々なものが考えられる。このうち、記事や文を単位とした場合には比較的容易に実験が行えるが、音声から段落の切れ目を検出することは容易ではない。そこで、記事および文を話題の単位とした。話題分野ベクトルの作成時にはクラスタリングを行わず、話題に含まれるすべての概念語を用いた。評価結果を表 2 に示す（表において、#oov は未登録語数、%red. は未登録語削減率を表す）。

表 2 より、どちらの場合も未登録語が削減されていることがわかる。また、記事を単位とした場合にはあまり未登録語が削減されないのに対し、文を単位とした場合は大幅に未登録語が削減されていることがわかる。話題の単位を細かくし、複数の話題から得られた関連語彙をマージして上位のものを獲得することによって、得られる関連語彙の質が向上するということが予想される。

表 4：語彙の条件を固定した場合の評価結果

記事の条件		25k		50k	
話題の単位	クラスタリング	#oov	%red.	#oov	%red.
記事	あり	1410	4.1	682	4.2
	なし	1369	6.9	676	5.1
文	あり	1273	13.5	634	11.0
	なし	**1233**	**16.1**	**583**	**18.1**

表 5：記事の条件を固定した場合の評価結果

語彙の条件		25k		50k	
話題の単位	クラスタリング	#oov	%red.	#oov	%red.
記事	あり	1294	12.0	633	11.1
	なし	**1183**	**19.6**	**578**	**18.8**
文	あり	1255	14.7	603	15.3
	なし	1233	16.1	583	18.1

4.2.2 クラスタリングの有無に関する評価

次に、それぞれの話題に対し、クラスタリングを適用した場合と適用しない場合の比較評価を行った。クラスタリングは重心法を適用し、クラスタ数が初期クラスタ数の 50%未満になるまでクラスタリングを行った。結果を表 3 に示す。

表 3 より、クラスタリングを行うと未登録語削減の性能が悪くなってしまうということがわかる。原因としては、次の２つが考えられる。

・ 正しいクラスタの選択に失敗している
・ クラスタが正しく分野を表していない

これらについては、4.2.4 節で詳しく検討する。

4.2.3 記事と語彙の組み合わせに関する評価

ここまでの実験では、入力記事と語彙について同一の条件を適用してきた。例えば、記事について話題の単位を文とし、クラスタリングを行った場合には、語彙についても話題の単位を文とし、クラスタリングを行うというものであった。しかし、入力記事と語彙で異なる条件を適用したほうがよいことも考えられる。そこで、4.3.2 節の実験で最もよい結果となった、文単位でクラスタリングなしという条件を基準とし、入力記事か語彙のどちらか一方をこの条件に固定して、もう一方を異なる条件とした場合の比較評価を行った。

まず、語彙の条件を文単位・クラスタリングなしという条件に固定し、記事の条件を変化させた場合の結果を表 4 に示す。

表 4 を見ると、その傾向が表 3 とまったく同じであることがわかる。表 3 と表 4 で異なるのは語

彙の条件であり、これを変化させてもほとんど影響がないということは、未登録語削減の性能が語彙の条件の影響をあまり受けず、記事の条件の影響を強く受けるということを意味している。

次に、記事の条件を文単位・クラスタリングなしという条件に固定し、語彙の条件を変化させた場合の結果を表 5 に示す。

表 5 を見ると、その傾向は表 3 と異なり、どのような語彙の条件でも未登録語が大幅に削減されている。このことは、未登録語削減の性能が記事の条件の影響を強く受けることをあらためて示しており、記事の条件として文単位・クラスタリングなしという条件が優れているということがいえる。

表 5 からもう一つわかることは、語彙の条件として記事単位・クラスタリングなしとした場合が最もよい結果になっていることである。記事の条件と語彙の条件は必ずしも同一の条件がよいというわけではないといえる。しかし、文単位・クラスタリングなしという条件と比較すると、50,000 語の場合に削減された未登録語数の差はわずかに 5 であり、有意な差であるとはいえない。

語彙の条件はあまり性能に影響しないことが判明したため、以降の評価では語彙の条件として、表 5 でもっともよい結果となった記事単位・クラスタリングなしという条件を使用し、記事の条件を変化させて評価を行う。

4.2.4 クラスタ選択の尺度に関する評価

これまでの実験では、クラスタリングを行った場合に良好な結果が得られなかった。4.2.2 節で述べたとおり、正しいクラスタが正しく選択されていないか、そもそもクラスタ自体が正しく分野を表せていないかということが原因であると考えられる。そこで、これらが原因であるかどうかを究明するため、クラスタの条件を変化させて実験を行った。

まず、クラスタが正しく選択されているかどうかについて検証した。複数のクラスタからどのクラスタを選択するかは、3.2.1 節の式におけるクラスタ $C_{\max}$ をどのように求めるかに依存する。これまでは、3.2.1 節の(1)式に示したとおり、クラスタリングによって得られた複数のクラスタの中から、最も多くの語彙を含むクラスタを選択していた。しかし、実は語彙数は重要ではなく、クラスタ内での各語彙の概念ベクトルがどれだけ近くに集まっているか（以下、結合度と呼ぶ）がクラスタの選択に重要である可能性が考えられる。そこで、結合度を考慮した尺度として、以下のように $C_{\max}$ を求めることを考える。

表 6：クラスタ選択の尺度に関する評価結果

語彙の条件		25k		50k	
話題の単位	クラスタ選択尺度	#oov	%red.	#oov	%red.
記事	語彙数	1418	3.6	685	3.8
	結合度	1451	1.4	707	0.7
	両方	1418	3.6	688	3.4
文	語彙数	1303	11.4	631	11.4
	結合度	1407	4.4	702	1.4
	両方	**1290**	**12.3**	**621**	**12.8**

$$C_{\max} = \arg\max_{C\in\Omega} \frac{1}{N(C)} \left| \sum_{w\in C} \vec{v}_c(w) \right| \quad \cdots(2)$$

このように、クラスタ内のすべての概念語における概念ベクトルの重心の絶対値を尺度とする。(2)式の絶対値の内側にある式は概念ベクトルの和を表しているが、ベクトル同士の距離が近いほど和の絶対値は大きくなり、遠いほど小さくなるため、概念語数が同じ場合は概念ベクトル同士が近いクラスタが選択される。これを概念語数により正規化することで、概念語数によらず結合度のみを考慮するようになっている。

さらに、クラスタの語彙数と結合度の両方が重要である可能性もある。そこで、語彙数と結合度の両方を考慮した尺度として、以下のように $C_{\max}$ を求めることを考える。

$$C_{\max} = \arg\max_{C\in\Omega} \left| \sum_{w\in C} \vec{v}_c(w) \right| \quad \cdots(3)$$

(2)式では概念ベクトルの重心の絶対値を考えたのに対し、ここでは概念ベクトルの和の絶対値を尺度としている。概念語数で正規化を行っていないので、概念語数が多いほど値は大きくなり、概念語数が同程度のクラスタが複数存在する場合には、重心の場合と同じく概念ベクトル同士が近いクラスタが選択されるようになっている。

それぞれの尺度によりクラスタを選択し、記事に対する話題分野ベクトルを作成した場合の比較評価を行った。話題の単位としては、文と記事の両方を試みた。

表 6 を見ると、結合度を用いた場合には、ほかの２つを用いた場合と比べて性能が悪いということがわかる。また、語彙数のみを用いた場合と、語彙数と結合度の両方を用いた場合を比較すると、同程度の未登録語が削減されている。これより、クラスタの選択にはクラスタに含まれる語彙数が重要であるといえる。両方を用いたほうが文単位のときに多少性能がよいが、これはおそらく文単位のときに語彙数が最大となるクラスタが複数存在することが多く、このような場合に、語

表 7 : クラスタリング手法に関する評価結果

語彙の条件		25k		50k	
話題の単位	手法	#oov	%red.	#oov	%red.
記事	重心	1418	3.6	688	3.4
	最短一致	1436	2.4	695	2.3
	最長一致	1422	3.3	686	3.7
文	重心	1290	12.3	621	12.8
	最短一致	1314	10.7	637	10.5
	最長一致	**1264**	**14.1**	622	12.6

表 8 : クラスタリング終了条件に関する評価結果

語彙の条件		25k		50k	
話題の単位	終了条件	#oov	%red.	#oov	%red.
記事	なし	1399	4.9	683	4.1
	25%未満	1410	4.1	682	4.2
	50%未満	1422	3.3	686	3.7
	75%未満	1417	3.7	692	2.8
文	なし	**1183**	**19.6**	**578**	**18.8**
	25%未満	1198	18.6	590	17.1
	50%未満	1264	14.1	622	12.6
	75%未満	1322	10.1	642	9.8

彙数のみの場合は語彙数以外の選択基準がないためランダムにクラスタを選択したのに対し、両方を考慮した場合は結合度の高いクラスタを選択することができたという違いがあらわれたものと考えられる。よって両方を考慮した(3)式が最も優れていると考えられる。しかし、この結果はクラスタリングなしの場合の結果に比べるとはるかに及ばないものとなっており、クラスタリングを用いた場合に性能がよくない原因は、クラスタが正しく選択されているかどうかによるものではないと考えられる。

4.2.5 クラスタリング手法に関する評価

次に、クラスタが正しく分野を表せているかどうかについて検証した。分野を表すクラスタが生成されているかどうかは、クラスタリング手法に依存すると考えられる。これまでは、クラスタリング手法として重心法を用いていたが、それ以外の手法を用いるとよい結果が得られる可能性がある。そこで、重心法のほかに最短一致法、最長一致法を用いてクラスタリングを行った場合との比較を行った。その結果を表 7 に示す。

表 7 を見ると、最短一致法の場合に性能が悪くなることがわかる。最短一致法を用いると、最初に大きくなり始めたクラスタが小さいクラスタを吸収していき、最終的に巨大なクラスタとなる傾向があるが、最初に大きくなるクラスタが分野を表すものであるとは限らないため、これにより誤ったクラスタが生成されて選択され、性能の低下を引き起こしたと考えられる。

重心法と最長一致法を比較すると、文単位で 25,000 語の場合に最長一致法のほうがわずかによいという結果が得られた。最長一致法を用いると大きいクラスタが生成されにくく、小さいクラスタが複数できあがる傾向があるため、クラスタの選択が難しくなるにも関わらず、多くの未登録語が削減されている。クラスタが小さい場合には、結合度の高いクラスタが分野をよく表しているため、正しいクラスタが選択されたのではないかと考えられる。しかし、この結果もクラスタリングなしの場合の結果に比べるとはるかに及ばないものとなっており、クラスタリングを用いた場合に性能がよくない原因は、クラスタリング手法によるものではないと考えられる。

4.2.6 クラスタリングの終了条件に関する評価

クラスタが正しく分野を表しているかどうかは、クラスタの大きさにも依存する。クラスタリングを行って一部の概念語を用いるよりも、クラスタリングを行わずにすべての概念語からなるクラスタを用いた場合のほうが性能がよいということは、分野と関係ないように見える語彙でも実は分野の推定に貢献している可能性が考えられる。これまでは、1 つの概念語からなるクラスタが概念語の数だけ存在する状態からクラスタリングを開始し、クラスタ数が開始状態の 50%になるまでクラスタリングを行っていたが、この終了条件を変更することで得られるクラスタの大きさを変化させることができる。そこで、終了条件を開始状態の 25%にした場合と 75%にした場合について比較を行った。その結果を表 8 に示す。ここで「終了条件なし」とはクラスタリングを行わない場合を指す。

表 8 を見ると、クラスタ数を少なくしていくことで未登録語が多く削減されていくようになることがわかる。分野の推定には一部の語彙ではなく、ほぼすべての語彙が必要といえるかもしれない。あるいは、クラスタ数の減少にしたがってクラスタの選択に失敗しにくくなるため、クラスタが大きくなり分野と関係ない概念語や認識誤りが多少含まれてしまっていても、クラスタの選択に失敗するよりはよい結果をもたらすと考えることもできる。クラスタの大きさと、その中に含まれる不要な概念語の割合、およびクラスタの選択に失敗する割合との相関関係については、今後詳しく検討していきたい。

表 9：認識性能に関する評価(25k)

手法	語彙数	#oov	%oov	%red.	%wer
なし	-	1471	2.10	-	27.50
従来手法	100	1440	2.06	2.1	27.71
	1000	1159	1.66	21.2	**27.38**
提案手法	100	1183	1.69	19.6	**27.17**
	1000	1017	1.45	30.9	**27.00**

表 10：認識性能に関する評価(50k)

手法	語彙数	#oov	%oov	%red.	%wer
なし	-	712	1.02	-	27.32
従来手法	100	710	1.01	0.3	27.39
	1000	580	0.83	18.5	**27.22**
提案手法	100	578	0.83	18.8	**27.08**
	1000	500	0.72	29.8	**26.99**

4.3 認識性能に関する評価

これまでは未登録語がどの程度削減されるかについて検証してきたが、語彙数を多くすると語彙の選択の幅が広がり、誤った語彙を選択する可能性が高まるため、認識性能に悪影響を及ぼすこともある。そのため、未登録語が削減されていても認識性能が向上するとは限らない。

そこで、提案手法により入力音声に適応させた辞書を用いて認識精度がどのように変化するかを調査した。Kemp らの手法[1]を従来手法として、提案手法との比較も行った。Kemp らの手法では、認識辞書の語彙サイズを一定に保っており、獲得した語彙の中でコーパスでの出現頻度が高い順に語彙を追加しながら、標準の語彙から頻度の低いものを取り除いているが、今回は比較のため、単純に獲得した語彙を追加して実験を行った。それぞれの手法で 100 語および 1000 語の語彙を獲得した結果を表 9 および表 10 に示す（表において#oov は未登録語数、%oov は未登録語率、%red. は未登録語削減率、%wer は単語誤り率を表す）。

表 9、表 10 を見ると、従来手法では、100 語という少ない語彙を追加してもほとんど未登録語の削減ができず、語彙を追加した辞書を用いて認識を行うと単語誤り率が増加してしまうことがわかる。従来手法を用いて未登録語の削減を行うためには、大量の語彙を追加する必要があるが、その際に不要な語彙も多く追加してしまうため、単語誤り率はほとんど低下しない。これに対し、提案手法では少ない語彙の追加により大幅に未登録語が削減され、認識性能も向上していることがわかる。標準の語彙サイズは数万語であり、それらに 100 語という少ない語彙を追加するだけなので、高速に認識辞書を構築することができるだけでなく、単語誤り率の低下の割合も大きい。従来手法に比べ、提案手法のほうがはるかに優れているといえる。

5. まとめ

コーパス中の語彙に対して語彙の分野を表す語彙分野ベクトルを算出しておき、入力音声に対してその話題の分野を推定し、その分野に近い語彙分野ベクトルを持つ語彙を入力に対する関連語彙として獲得することにより、認識辞書を適応させる手法を提案した。TV ニュース番組を対象として実験を行った結果、本手法により適応させた辞書を用いることで未登録語が削減され、認識精度が向上することを示した。

今回の実験では、クラスタリングを用いた場合によい結果が得られなかったが、認識誤りのみを取り除いたクラスタを得ることができれば、クラスタリングを行わない場合と比較してよい結果が得られると考えられる。認識誤りのみを取り除くためのクラスタリング手法について検討していきたい。また、今回は基準の辞書から語彙を削除するということは行わなかったが、基準の辞書の中にも特定の分野でしか用いられず、入力の分野に関係ない語彙が存在すると思われる。そのような語彙を基準の辞書から取り除く手法についても検討していきたい。

参考文献

[1] T. Kemp et al., “Reducing the OOV Rate in Broadcast News Speech Recognition,” Proc. of ICSLP-98, pp.1839-1842, 1998.

[2] H. Yu et al., “New Developments in Automatic Meeting Transcription,” Proc. of ICSLP-2000, Vol. IV, pp.310-313, 2000.

[3] T. Kato et al., “Idea-Deriving Information Retrieval System,” Proc. of 1st NTCIR Workshop, pp.187-193, 1999.

[4] 別所克人, “クラスター内変動最小アルゴリズムに基づくトピックセグメンテーション,” 情報処理学会研究報告,NL-154,pp.177-183, 2003.

[5] 廣嶋伸章他, “音声認識における未登録語削減を目的としたコーパスからの語彙獲得,” 言語処理学会第10回年次大会,pp.79-82, 2004.

[6] K. Ohtsuki et al., “Multi-Pass ASR using Vocabulary Expansion,” Proc. of ICSLP-2004, 2004. (to be appear)

[7] 野田喜昭他, “音声認識エンジン VoiceRex の開発,” 音講論, 2-1-19, pp.91-92, 1999.