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

廣嶋伸章　大附克年　別所克人　林良彦
日本電信電話株式会社　NTT サイバースペース研究所
{hiroshima.nobuaki, ohtsuki.katsutoshi, bessho.katsuji, hayashi.yoshihiko}@lab.ntt.co.jp

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登録することにより入力音声に対する未登録語を削減することができ、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、音声認識結果を入力文書として、その内容に関連する語彙をコーパスから獲得する手法を提案する。提案手法では、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、入力文書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文書に対する関連語彙として獲得する。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。

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、入力音声に対する未登録語を削減することができ、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。

本稿では、音声認識における未登録語の削減を目的として、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する手法を提案する。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する手法と、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報告する。

2. コーパスからの関連語彙獲得

コーパスからの関連語彙獲得については、これまでいくつかの手法が提案されてきたが、どの手法もまず入力に対して間接的に関連文書を検索し、関連文書の中から関連語彙を獲得するというものであった。

Kemp らは、未登録語削減を目的として、データベースから認識結果に対する関連文書を検索し、関連文書から関連語彙を獲得している[1]。関連文書の検索には、情報検索でよく用いられる Okapi を用いている。関連文書に含まれる語彙をすべて獲得し、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、標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認識辞書を更新している。

Yu らも、未登録語削減を目的として、Web ページから話題に関連する文書を検索し、関連文書から関連語彙を獲得している[2]。関連文書の検索には、インターネットサーチエンジンである Infoseek を利用している。語彙の獲得には、語彙と話題との相互情報量を用いている。

しかし、これらの手法はどれも、語彙を獲得するために”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という問題がある。また、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。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。

3. 概念ベースを用いた関連語彙獲得

本稿では、次のようにして語彙の獲得を行うことで、従来手法の問題点を解消する。まず、コーパス中の各語彙に対し、概念ベース[3]を用いて分野を表す語彙分野ベクトルを事前に求めておく。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。さらに、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があることを利用し、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ことで、認識誤りに対して頑健であることが期待される。

以下では、提案手法で用いる概念ベースについて述べ、提案手法の詳細なアルゴリズムについて述べる。

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 アルゴリズム

提案手法では、入力となる認識結果に対して、発声内容の分野を推定し、その分野に近い語彙を関連語彙として獲得する。しかし、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。以下では、語彙分野ベクトルを算出するためのアルゴリズムについて述べ、その語彙分野ベクトルを用いて関連語彙を獲得するためのアルゴリズムについて述べる。

3.2.1 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、そのため概念ベースは比較的頻度の高い概念語のみから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、学習コーパス中の各語彙の語彙分野ベクトルは、その語彙と同一文中で共起する概念語の概念ベクトルを用いて補完することによって求める。

語彙分野ベクトルを算出するアルゴリズムは以下のとおりである。

(1) コーパス中の各文において、その文に出現する概念語の概念ベクトルの重心を求め、文ごとの分野ベクトルとする。

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
…	…	…	…	…

(2) 各語彙に対し、その語彙が含まれるすべての文における文ごとの分野ベクトルの重心を求め、語彙分野ベクトルとする。

このようにして語彙分野ベクトルを求めることで、コーパス中に一度しか出現しなかった語彙についても、その語彙と同一文中に概念語が含まれていれば、語彙の分野を正しく表すことができる。

3.2.2 関連語彙獲得

認識結果に対し、語彙分野ベクトルを用いて語彙を獲得する。そのアルゴリズムは以下のとおりである。

(1) 認識結果からすべての概念語を抽出する。

(2) 概念語を概念ベクトルに基づいてクラスタリングし、概念語を多く含む順に上位 M 個のクラスタについて概念ベクトルの重心を求め、文書分野ベクトルとする。

(3) 語彙ごとに、語彙分野ベクトルと文書分野ベクトルとの距離の最小値である関連度を求め、その値が大きい順に N 個の語彙を関連語彙として獲得する。

認識結果中の概念語は発声内容の分野を表していると考えられるが、すべての概念語が分野を表しているわけではないので、クラスタリングを行って概念語を多く含むクラスタだけを文書分野ベクトルの算出に用いることで、分野を表していない概念語を取り除くことができる。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。文書分野ベクトルを M 個求めているのは、例えば、イラクの治安問題という内容の場合に、「戦争」と「政治」についての話題が取り上げられるというように、発声内容の分野が複数にまたがる場合を考慮しているためである。

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。以下では学習および評価に利用したデータについて述べ、実験結果を報告する。

4.1 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、3.2.1 で述べた手法で語彙分野ベクトルを作成した。

4.2 評価データ

2002 年 12 月に放送された TV ニュース番組を 30 番組用意し、トピックの音声ごとに認識結果と正解データを作成して評価に用いた。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[4]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。

4.3 未登録語削減に関する評価

まず、語彙サイズが 25,000 語（25k）と 50,000 語（50k）のそれぞれについて、語彙を追加しない場合と、提案手法により 100 語、1000 語の関連語彙を獲得して認識辞書の語彙に追加した場合で、未登録語がどの程度削減されるかの評価を行った。トピック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。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 20%未満になるまでクラスタリングを行った。文書分野ベクトルの数は 1 とした。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、どちらの語彙サイズに対しても、獲得した語彙を追加することにより未登録語が削減されていることがわかる。100 語を追加するだけで 17%程度の未登録語が削減され、1000 語を追加すると 30%前後の未登録語が削減されている。よって、提案手法は未登録語の削減に有効であることが示された。

表 3 は、通信技術について述べられている音声の認識結果を入力として 10 個の語彙を獲得した結果である。獲得された語彙を見ると、通信技術に関連する語彙が適切に獲得できていることがわかる。第 3 位の「ＡＤ」は、形態素解析により「ＡＤＳＬ」が「ＡＤ／ＳＬ」という２単語に解析されていたことが原因である。使用した音声には「ＦＴＴＨ」という単語が含まれており、標準の語彙に含まれていなかったため認識誤りとなっていたが、関連語彙として獲得することができた。

表 2：未登録語削減に関する評価結果

語彙の追加	25k		50k	
	#oov	%red.	#oov	%red.
なし	1471	-	712	-
100 語	1207	17.9%	586	17.7%
1000 語	1002	31.9%	506	28.9%

表 3：獲得された語彙の例

順位	獲得語彙	関連度
1	プロトコル	0.912
2	ＢＢ	0.876
3	ＡＤ	0.873
4	メガビット	0.873
5	ＦＴＴＨ	0.872
6	ＯＣＮ	0.871
7	テレマティクス	0.869
8	モバイル	0.866
9	ＰＤＡ	0.864
10	メガビット	0.862

4.4 従来手法との比較評価

次に、提案手法と従来手法との比較評価を行った。従来手法として、Kemp らの手法[1]を用いた。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 語の語彙を獲得した結果を表 4 に示す。

表 4 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ないことがわかる。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。Kemp らが語彙を追加しながら削除を行ったのは、大量の語彙の追加によりメモリ量の問題が発生したためであると予想される。追加と削除を行うにはかなりの処理時間を必要とするため、高速に認識辞書を構築することができない。これに対し、提案手法では少ない語彙の追加で十分な効果が得られる。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ことで、メモリ量の問題が発生することはなく、高速に認識辞書を構築することができる。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。

4.5 クラスタリングに関する評価

提案手法では、入力の分野を推定するためにクラスタリングを行い、多くの概念語を含んでいるク

表 4：従来手法との比較

手法	25k		50k	
	#oov	%red.	#oov	%red.
提案手法	1207	17.9%	586	17.7%
従来手法	1441	2.1%	711	0.1%

ラスタから文書分野ベクトルを作成している。このクラスタリングの効果を調査するため、クラスタリングを行った場合と、クラスタリングを行わずに入力中のすべての概念語における概念ベクトルの重心を文書分野ベクトルとした場合について比較評価を行った。

4.5.1 トピックごとの評価

まず、4.3 の実験と同様に、トピックごとの未登録語数の合計を求めた。獲得する語彙数は 100 語とした。その結果を表 5 に示す。

表 5 より、クラスタリングを行った場合よりも行わない場合のほうがよい結果であることがわかる。これは、語彙分野ベクトルの作成時にはクラスタリングを行わないのに対し、文書分野ベクトルの作成時にクラスタリングを行うため、両者にずれが生じていると考えられる。語彙分野ベクトルも文書分野ベクトルと同様に、文書中の概念語をクラスタリングして求めるなどの手法について検討が必要であると思われる。また、この結果は、概念語の全体の指す分野と、クラスタリングにより選別された概念語の集合が指す分野にあまり違いがないということを表している。たいていの場合、１つのトピックは１つの分野について述べられているので、これは妥当な結果と言える。

4.5.2 ニュース番組ごとの評価

4.5.1 では、評価の単位をトピックとしたが、通常、扱う音声がトピックごとに分かれているということは稀である。例えば、ニュース番組は、番組の開始時間と終了時間は決まっていることが多いが、トピックによって所要時間が異なるため、それぞれのトピックごとに開始時間と終了時間が決まっていることはない。トピックごとに区切るような手法としてトピックセグメンテーションがあるが、このような手法を用いずに、ニュース番組１本というような複数のトピックを含む音声から、その内容に関連する語彙を獲得できることが望ましい。そこで、ニュース番組ごとに、クラスタリングを行った場合と行わない場合で未登録語数の比較実験を行った。獲得する語彙数は 1000 語とし、文書分野ベクトルの数は可変とした。その結果を表 6 に示す。

表 6 を見ると、表 5 とは対照的に、クラスタリングを行った場合のほうが、行わない場合よりも、多くの未登録語が削減されていることがわかる。ニュース番組のように複数のトピックを含む音声の認識結果に対しては、クラスタリングを行って複数のクラスタから文書分野ベクトルを求めるという提案手法が有効であることを示している。しかしながら、表 2 と表 6 を比較すると、ニュース番組単位で 1000 語の語彙を追加した結果は、トピック単位で 100 語の語彙を追加した結果よりも劣っていることがわかる。獲得された語彙を観察してみると、過去に起こった犯罪に関連する地名が大量に獲得されていた。このように、ある分野で多くの語彙が獲得されてしまうと、残りの分野で獲得される語彙が少なくなり、語彙の少ない分野については未登録語の削減が適切に行われなくなるということが原因であると思われる。各文書分野ベクトルから獲得される語彙が均一になるようにするなどの工夫が必要であると思われる。

表 5：クラスタリングに関する評価結果（トピック単位）

クラスタリング	25k		50k	
	#oov	%red.	#oov	%red.
あり	1207	17.9%	586	17.7%
なし	1165	20.8%	565	20.6%

表 6：クラスタリングに関する評価結果（ニュース番組単位）

クラスタリング	25k		50k	
	#oov	%red.	#oov	%red.
あり	1256	14.6%	622	12.6%
なし	1337	9.1%	656	7.9%

5. まとめ

本稿では、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、入力文書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文書に対する関連語彙として獲得する手法を提案した。本手法を音声認識における未登録語の削減に適用し、基準となる辞書による認識結果を用いた実験の結果、獲得した語彙を追加することで未登録語が削減されることを示した。

今後の課題としては、まず語彙分野ベクトルの算出手法について改良が必要であると思われる。また、今回は概念ベクトルを用いてクラスタリングを行ったが、語彙分野ベクトルを用いることも考えられるので、その場合についても評価したい。

参考文献

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

[2] Yu et al., “New Developments in Automatic Meeting Transcription,” Proc. of ICSLP, 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] 野田他, “音声認識エンジン VoiceRex の開発,” 音講論, 2-1-19, pp.91-92, 1999-9.