E-023

ウェブ上の言語知識を利用した音声認識用単語辞書の更新手法

A method to update ASR lexical information using Web resources

鈴田健太郎† 西村竜一† 河原英紀† 入野俊夫†

Kentaro Suzuta Ryuichi Nisimura Hideki Kawahara Toshio Irino

1．はじめに

一般的な大語彙連続音声認識では，音素とその特徴を定義した音響モデルと単語とその読みを定義した単語辞書，および単語間の接続確率を定義した言語モデルという３つの辞書を参照し認識結果を出力する（以下，単語辞書と言語モデルを合わせて言語モデルと呼ぶ）．新出語や流行語，死語に代表されるように，言語は時間と共に刻々と変化する．高精度な音声認識精度を保持するためには，言語の使用形態の変化と共に言語モデル中の知識も更新しなければならない．

また，認識タスクがあるトピックに特定される場合を考える．この際，言語モデルをトピックに関連するテキストデータから構築することによって音声認識率は向上することが知られている[1]．[2][3]では，ウェブテキストから抽出したテキストを用いた学習を行っていた．その過程においては，自然文章から単語を数え上げるための前処理として，形態素解析による単語分割および読み情報の付与が必要である．しかし，収集したウェブ テキスト中には，形態素解析の辞書には未登録であり，読みの付与ができない単語が頻出する．このため，統計量の算出が正しく行えない．また，音声認識に必要な単語辞書に対して読みを正しく決定できない問題が発生していた．本研究では，この問題に対し，ウェブ ネットワーク上に存在する言語知識（はてなキーワードAPI[4]）を利用して，正しい読みを獲得し，形態素解析用の形態素辞書及び音声認識用の言語モデルを更新する手法を提案する．

2．提案手法の概要

提案手法は，図1 のように三段階のステップを持つ．このうち ウェブ テキスト収集及び単語辞書更新の２つのステップにおいて，ウェブネットワークからの知識の獲得を実現している．提案手法の手順を以下に示す．本手法は，特定の認識タスクに応じた言語モデルの自動構築を実現するものである．

Step1 認識タスクに関するキーワードをクエリとしたウェブ 検索を行い，言語モデルの学習元となるテキストを収集する．

Step2 収集テキストから形態素辞書に未登録な単語を自動抽出し， ウェブ 知識を元に，その単語の読みを獲得する．その結果を用いて，形態素辞書を更新する．

† 和歌山大学大学院システム工学研究科

*1 “ROCKIN ’ ON JAPAN 2006 年 8 月号,” pp42-75, ロッキングオン社

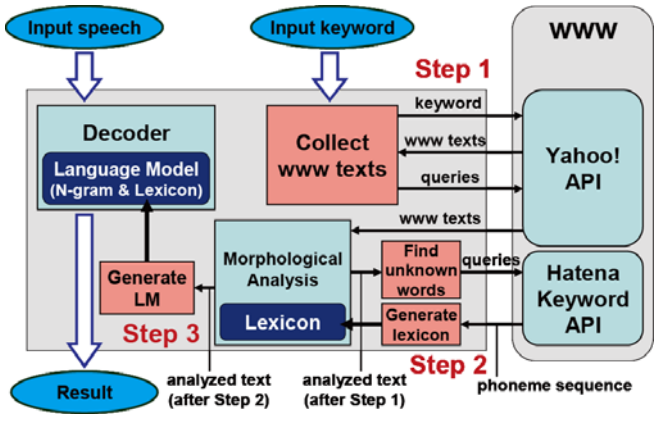

図１．提案手法の構成

Step3 収集テキストを，更新された形態素辞書で再解析する．その結果から単語の出現頻度を求めて，N-gram言語モデルを構築する．

3．ウェブテキスト収集

従来，認識タスクに関連するトピックを持ったテキストを人手で選択・収集するのは，非常に大きな手間が必要であった．本手法のStep 1 では，この作業を自動化した．安易なウェブ検索によってウェブテキストを収集した場合，トピックに関連のないウェブテキストを収集してしまうことが予想される．これは音声認識率の低下の原因となる．そこで，提案手法ではウェブテキスト収集を２段階で行なうことでこの問題に対処する．１度目のウェブテキスト収集では，認識タスクを指定するキーワードを検索クエリとしてウェブ検索をし，HTML テキストを収集する．ウェブ検索には Yahoo! Search API [5] を利用した．また，収集するページ数を30 に設定した．次に，収集したテキストに出現する名詞の中から，２度目のウェブ検索における補助的な検索クエリを選定する．補助クエリには，同一ページに出現するキーワードと類似・関連性の高い単語を選択する．単語間類似度W(t) は，以下の式(1) から求めた．DF(t) は単語t が出現する文書数を表す．

$$W\ (keyword, t_i)\ =\ \frac{2\ DF\ (keyword, t_i)}{DF\ (keyword) + DF\ (t_i)} \qquad (1)$$

また，Yahoo!関連検索ワードウェブサービス[5] が推薦する単語に対しても上記の類似度計算を行い，補助クエリの候補とした．最終的に，単語間類似度上位 50 名詞を補助的な検索クエリとした．これと，入力されたキーワードを組み合わせ，２度目のウェブ検索（AND

検索）を行う．この結果に対してウェブクロールを行い，言語モデル学習用テキストデータの収集を実施した．

4. 単語辞書更新

Step 2 では，ウェブ上に存在する既存の知識集合を用いて形態素辞書を更新した．収集したテキストに含まれる形態素辞書に未登録な単語を適切に処理するために，本研究では，形態素解析を以下のように2 回適用する．1 度目は辞書に未登録な単語を抽出するための解析である．更新前の形態素辞書を用いて，以下のように解析された単語列を辞書に未登録な単語として抽出した．

1. 形態素解析の結果，「未知語」と解析された形態素
2. 英数字列
3. 形態素2-gram
4. カタカナ文字列

1.～3. の未登録と判断された単語に対しては，はてなキーワードAPI から読み情報を与え，辞書に追加した．はてなキーワードAPI は，入力された単語に対し，読み情報の要素を持つXML 形式の検索結果を出力する．APIでヒットし，新たに登録された単語の品詞は「名詞-サ変接続」とした．また，形態素生起コストには100 を暫定的に設定した．カタカナ文字列の読みは，カタカナ文字列そのものとし，1.～3. と同様の条件で辞書に追加した．

3. の形態素2-gram とは，“ロック”+“歌手”のように出現した形態素の前後を組につないだ単語である．はてなキーワードAPI は完全一致検索ではないため，2-gramが登録単語の一部になっていればヒットする．そのため，2-gram を未登録語の候補とすることで，分割して形態素解析されてしまった単語にも読みを与えることができる．

Step 3 の言語モデル構築では，2 度目の形態素解析を行う．このとき，更新された形態素辞書を用いて処理を行った．この結果，正しい単語出現頻度を求めることができるようになった．

5. 音声認識実験

提案した手法を用いて言語モデルの構築を行い，音声認識実験を行った．形態素辞書を更新した場合（提案法）と未更新の場合（従来法）の単語正解精度を比較する．音声認識デコーダはJulius ver. 3.5.3，形態素解析器はChaSen ver. 2.3.3，形態素辞書はipadic-2.6.3 を使用した．作成したのは，言語モデル（2-gram，逆向き3-gram）と単語辞書である．言語モデルの学習に用いたウェブ ページ数は500 ページ（検索クエリ50 対× 10 ページ， 1.8MB）である．認識タスクは邦楽アーティストの名前から「くるり」に設定した．テストセットとして「くるり」を紹介する雑誌記事*1から100文章を抽出して，男女各5 名の話者が読み上げた音声を用意した．

認識結果への悪影響が予想されるため，アルファベットの一文字は単語辞書から除外した．提案手法による更新処理の結果，形態素辞書に新たに追加された単語数は5,860 であった．「くるり」タスクに対して追加登録された形態素辞書のエントリ例を図2 に示す．図3 に，テストセットに対する平均単語正解精度および話題語認識率を示す．話題語とは，今回の認識タスクである「くるり」に関連する単語（アルバム名や他のアーティスト名など）である．

人名やカタカナ語(“カオティック”など) に対しては認識精度の向上を確認した．しかし，英語表記の曲名など，はてなキーワードにも未登録な単語に関しては従来法と変わりがなかった．

```
(品詞(名詞サ変接続))((見出し語
  (ＲＯＳＳＯ 100))(発音ロッソ))

(品詞(名詞サ変接続))((見出し語
  (大村達身100))(発音オオムラタッシン))
```

図2: 追加登録された形態素辞書のエントリ例

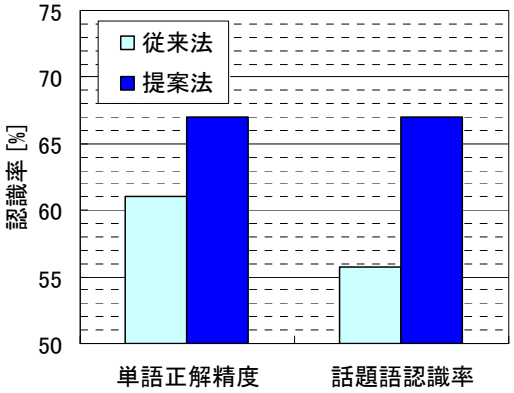

図3: テストセットに対する音声認識率

6. おわりに

単語辞書の更新にウェブ知識を利用した音声認識の改善手法を提案した．形態素辞書の更新により，単語正解精度の向上を確認した．従来法との比較において，6.0 %の改善を確認した．また，話題語認識率においても11.3 %の改善があった．今回は，特定のトピックに特化した言語モデルを作成した．今後は，複数のトピックに対応できるように，複数の言語モデルを融合する手法について検討する．

参考文献

[1] 駒谷他, 情処学論, vol. 44, no. 5, pp. 1333-1342, 2003.
[2] 翠他, 情処学論, vol. J90-D, No. 11, pp. 3024-3032, 2007.
[3] 梶浦他, 音講論(春), pp. 75-76, 2007.
[4] “はてなキーワードAPI とは,” はてな, 2005. http://d.hatena.ne.jp/keyword/
[5] “Yahoo!デベロッパーネットワーク,” ヤフー, 2008. http://developer.yahoo.co.jp/