音声とテキストからの語彙自動獲得*

◎倉田 岳人, 森 信介, 伊東 伸泰, 西村 雅史 (日本 IBM 株式会社 東京基礎研究所)

1 はじめに

一般分野を対象とした認識単語辞書は整備されてきたが，あらゆる分野のすべての単語を網羅できているわけではない．そのため，大語彙連続音声認識 (LVCSR) を新しい分野に導入する場合には，その分野特有の単語を認識単語辞書に追加しなければならない．コールセンタの書き起こしや議会の議事録作成のようなアプリケーションでは頻繁に LVCSR システムの適応・更新を行う必要があるが，認識単語辞書のサイズには制約があり，また高い認識率を得るためには，必要な単語のみを追加することが望ましい．

本稿では，対象分野の音声，および関連するテキストデータから，自動的に分野特有の単語を読みつきで獲得する方法を提案する．実験の結果，適切な単語と対応する正確な読みを獲得することができ，それらを選択的に利用することで，対象分野の LVCSR の精度を向上させることができた．

2 提案手法

提案手法では，最初に，対象分野の単語分割されていないテキストデータから大量の単語候補を抽出し，それらに表記から考えられる読みを割り当て，認識単語辞書を作る．次に，テキストデータから確率的単語分割を利用して，対象分野の言語モデル (LM) を作成し，対象分野用の LVCSR システムを構築する [1]．この LVCSR システムを利用して，対象分野の音声を認識し，その際に利用された単語候補とその読みを，分野特有の単語として読みつきで獲得する．最後に，獲得した単語とテキストデータから LM を再推定し，対応する読みも利用して LVCSR システムを再構築する．

Fig. 1 に提案手法の処理の流れを示した．以下ではこの図に従って，提案手法について説明を行う．括弧内に太字で示す数字は，図中の矢印に添えられている数字と一致している．なお，一般分野の単語分割済みコーパスと読み付きの単語辞書 (一般分野単語辞書) はすでに用意されているものとし，これらから一般分野 LM の推定を行う．また，対象分野の音声を分割し，一部を分野特有の単語と読みを獲得するための学習用音声，残りを評価用音声として利用した．

(**1**) 対象分野のテキストデータから単語候補を抽出した．認識単語辞書に含まれない単語は LVCSR では認識されないことを考慮すると，高い再現率を得ることができる方法を採用しなければならない．ここでは，部分文字列の頻度に基づく方法を採用した [2, 3]．精度よりも再現率を重視するため，ここでは大量の単語候補が得られることになる．

(**2**) 単語候補に対して音声合成システムで利用される未知語読みモデル [4] を利用して読みを付与した．なお，未知語読みモデルの精度が 100%ではないことを考慮し，モデルが出力する上位 10 個の読みを付与した．(**1**) で得た単語候補と各々に対する読みの集合を一次単語辞書と呼ぶ．

(**3**) テキストデータを確率的に単語分割した．ここでは，まず自動単語分割器の精度 α を算出し，その自動単語分割器でテキストデータを単語分割した．そして，単語境界と判定された部分の単語境界確率を α に，判定されなかった部分の単語境界確率を $1-\alpha$ とすることにより，確率的単語分割を行った．

(**4**) 確率的単語分割済みのテキストデータ，(**1**) で

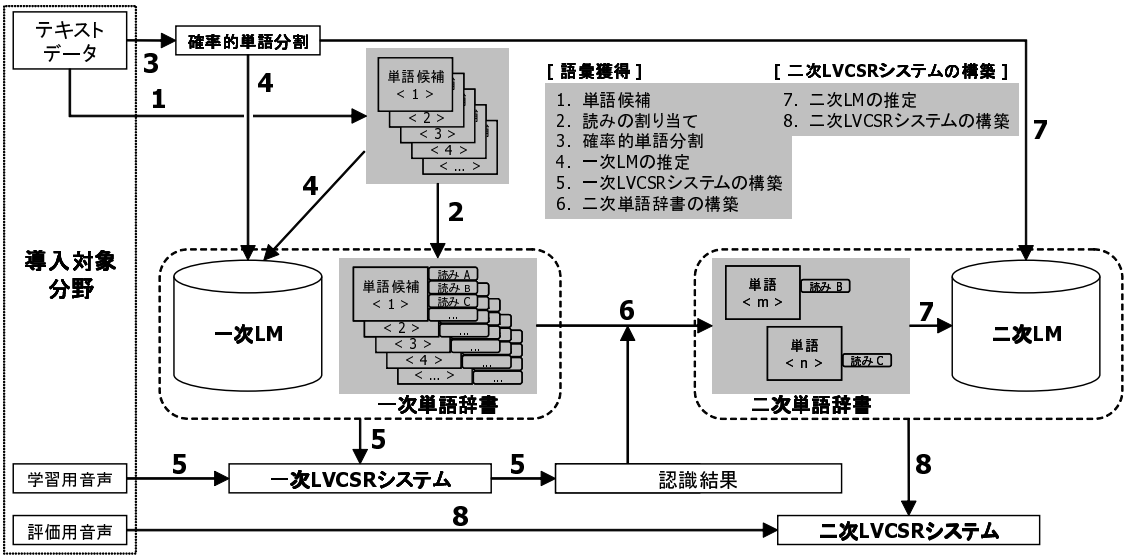

Fig. 1 Overview of Proposed Method

* Unsupervised Lexicon Acquisition from Speech and Text. by Gakuto KURATA, Shinsuke MORI, Nobuyasu ITOH, Masafumi NISHIMURA (IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.)

Table 1　講義音声とテキストデータのサイズ

テキストデータの文字数	講義音声 [分]	
	学習用音声	評価用音声
73,437	12.3	6.2

Table 2　二次単語辞書の正解率 [%]

2 回以上	1 回のみ	平均
97.2	68.9	79.5

Table 3　獲得された単語の例

頻度	単語	読み
27	受容体	ju yo o ta i
13	リン酸化	ri n sa n ka
2	単量体	ta n ryo o ta i
2	残基	za n ki

Table 4　一次，二次 LVCSR システムの比較

LVCSR System	単語辞書		未知語率 [%]		CER [%]	
	単語数	読みの総数	学習用	評価用	学習用	評価用
一般	—	—	5.88	6.15	27.0	26.1
一次	3,999	26,169	2.31	1.71	12.3	**9.9**
二次	326	326	2.31	1.90	11.7	**9.7**

得られた単語候補，および一般分野単語辞書から対象分野の LM を推定した．この LM を一次 LM と呼ぶ．

(**5**) 一次単語辞書，一般分野単語辞書，一次 LM と一般分野 LM の補間 LM を利用して一次 LVCSR システムを構築し，学習用音声に対して LVCSR を行った．

(**6**) 一次単語辞書の中で LVCSR 時に利用された単語候補とその読みの組み合わせを選択した．これらを二次単語辞書と呼ぶ．

(**7**) テキストデータと二次単語辞書，および一般分野単語辞書から LM を再推定した．ここで得られる LM を二次 LM と呼ぶ．

(**8**) 二次単語辞書，一般分野単語辞書，二次 LM と一般分野 LM の補間 LM を利用して二次 LVCSR システムを構築し，評価用音声に対して LVCSR を行った．

3　評価実験とその結果

3.1　評価実験の概要

放送大学の「生物」の講義音声を利用して実験を行った．講義の内容は専門的であり，一般的な認識単語辞書には含まれない単語が出現すると考えられる．また講義に関連するテキストデータとして，講義の教科書を利用した．Table 1 に利用したテキストデータのサイズ，分割した講義音声のサイズを示した．

3.2　実験結果

まず，一次 LVCSR システムの認識結果から得られる二次単語辞書について検証した．利用された単語候補が分野特有の単語としてふさわしいかどうかを単語内の係り受け関係も考慮して判断し [5]，さらに同時に利用された読みが正しかった場合に，正しく分野特有の単語が獲得された，と判定した．認識時に複数回使われた単語と一度しか使われなかった単語の正解率，およびその平均を Table 2 に示した．認識時に 2 回以上利用された単語については非常に高い正解率が得られた．これらの単語は，学習用音声にその正しい読みが，高い LM 確率を得る文脈で複数回出現していた，ということであり，実際に正しい単語である場合が多かったと推測される．逆に，一度しか現れなかった単語については，単なる挿入・置換誤りの可能性もあり，正解率が低くなっていると推測される．また，Table 3 に，得られた単語の例とその読み，および認識時に利用された頻度を示した．

次に学習用音声，評価用音声に対する一次，二次 LVCSR システムの文字誤り率 (CER: Character Error Rate)，および未知語率を Table 4 に示した．一次，二次単語辞書のサイズについても同時に示した．比較のため，一般分野 LM と一般分野単語辞書のみを利用した場合についての結果も示した．学習用音声に対する結果を比較すると，二次 LVCSR システムは一次 LVCSR システムよりも高い精度を得ることができた．二次 LVCSR システムは学習用音声で学習を行っているが，提案手法により自動的に精度を向上させることができた．一次 LM においては単語として意味をなさない単語候補に対しても確率値が割り当てられていたが，二次 LM では学習用音声によって検証された単語にのみ正しく確率値が割り当てられようになることが効果的であったと考えられる．次に，評価用音声に対する結果を比較すると，追加されている二次単語辞書のサイズは一次単語辞書のサイズの 10%程度となり，さらに二次 LVCSR システムは一次 LVCSR システムよりも高い性能を得ることができた．これは，学習用音声によって検証された二次単語辞書が正確な分野特有の単語を含んでいることを示唆している．

4　まとめ

本稿では，日本語における自動語彙獲得方法を提案した．提案手法では，テキストデータ中に出現する部分文字列の中から，対象分野の音声を利用して，分野特有の単語とその読みを選択した．実験により，正確な分野特有の語彙が獲得されることを確認した．また，それが対象分野の LVCSR の精度向上に貢献することも検証した．

謝辞　放送大学の番組制作に携わっておられる方々に深謝します．

参考文献

[1] S. Mori *et al.*, "Word *N*-gram Probability Estimation From A Japanese Raw Corpus," in *Proc. Interspeech*, 2004.

[2] G. Kurata *et al.*, "Unsupervised Adaptation of a Stochastic Language Model Using a Japanese Raw Corpus," in *Proc. ICASSP*, 2006.

[3] H. Feng *et al.*, "Accessor Variety Criteria for Chinese Word Extraction," *Computational Linguistics*, vol. 30, no. 1, pp. 75–93, 2004.

[4] T. Nagano *et al.*, "A Stochastic Approach to Phoneme and Accent Estimation," in *Proc. Interspeech*, 2005.

[5] M. Asahara *et al.*, "Japanese Unknown Word Identification by Character-based Chunking," in *Proc. COLING*, 2004, pp. 459–465.