仮名漢字変換ログを用いた講義音声認識のための言語モデル適応

山口 洋平　　森 信介　　河原 達也
京都大学 情報学研究科

1　はじめに

近年、オンラインで利用可能な学術講義アーカイブの量が飛躍的に増加している [2, 13]。これらのコンテンツの利便性を高めるために、索引を付与したり、わかりやすい字幕を表示することの需要が高まっている。そのため、音声認識を用いたインデキシング [1]、ノートテイク支援 [3]、自動要約 [14] などの研究はこれまで以上に注目を集めている。

講義音声認識における問題点の一つは、大学講義を対象とした日本語の大規模な音声とその書き起こしのデータベースが現状で存在しないことである。この問題に対して、講義で使用される教科書やスライド、また書き起こしを利用して言語モデルを補間し適応する手法 [8] が提案されている。これらの資料は講義における重要なキーワードを多く含み、音声にもそれらのキーワードが多く現れると考えられる。よって、これらの資料は言語モデルの適応において有効であると期待できる。

しかし、これらの資料は大学講義の専門性のために、多くの専門用語を含む。既存の解析器は未知語の周辺で解析を誤りやすい。講義ではそのような専門用語が頻出するため、解析誤りの影響は無視できない。これが音声認識システムの言語モデルの予測性能の低下を招き、さらには認識精度の低下につながる。

本稿では、このような解析誤りに対処するために、講義資料を作成する過程で得られる仮名漢字変換ログを活用して言語モデルを適応する方法を提案する。仮名漢字変換ログとは、仮名漢字変換エンジンで文章を作成するときの変換の入出力の履歴であり、単語境界や入力記号列の情報を含んでいる。仮名漢字変換ログは変換履歴が取れる仮名漢字変換システムを用いることで、仮名漢字変換というユーザが自然に行う言語処理を通じて、半自動的に得られるアノテーションデータと見なすことができる。これの活用により、言語モデル作成時の単語分割や発音推定の誤りを削減し、言語モデルの性能向上と認識精度の改善を実現した。

2　仮名漢字変換システム

仮名漢字変換ログの収集には、確率的モデルによる仮名漢字変換 [12] を用いた。本節では、この仮名漢字変換システムについて説明する。

2.1　確率的モデルによる仮名漢字変換

確率的モデルによる仮名漢字変換は、キーボードから直接入力可能な入力記号 $\mathcal{Y}$ の正閉包 $\boldsymbol{y} \in \mathcal{Y}^+$ を入力として、変換候補文字列 $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots)$ を確率 $P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$ の降順に提示する。ここで、$P(\boldsymbol{y}|\boldsymbol{x})$ は、確率的仮名漢字モデルであり、日本語文 $\boldsymbol{x}$ を所与とした入力記号列の生成確率を表す。また、$P(\boldsymbol{x})$ は確率的言語モデルである。

2.1.1　確率的言語モデル

確率的言語モデル $P(\boldsymbol{x})$ には、単語 2-gram モデルを用いる。このモデルは、文を単語列 $\boldsymbol{w}_1^h = w_1 w_2 \cdots w_h$ とみなし、これらを文頭から順に以下の式を用いて予測する。

$$M_{w,2}(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(w_i|w_{i-1}) \tag{1}$$

この式の中の w_i $(i \leq 0)$ は、文頭に対応する特別な記号であり、w_{h+1} は、文末に対応する特別な記号である。未知語を表わす特別な記号 UW を用意する。未知語の予測の際は、まず、単語 2-gram モデルにより UW を予測し、さらにその表記 (文字列)$\boldsymbol{x}_1^{h'}$ を以下の文字 2-gram モデルにより予測する。

$$M_{x,2}(\boldsymbol{x}_1^{h'}) = \prod_{i=1}^{h'+1} P(x_i|x_{i-1}) \tag{2}$$

この式の中の x_i $(i \leq 0)$ と $x_{h'+1}$ は、それぞれ、語頭と語末に対応する特別な記号である。

2.1.2　確率的仮名漢字モデル

確率的仮名漢字モデル $P(\boldsymbol{y}|\boldsymbol{x})$ は、日本語文を単語列 $\boldsymbol{w}$ とみなし、以下の式で表される。

$$M_{PM}(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{h} P(\boldsymbol{y}_i|w_i) \tag{3}$$

ここで、部分入力記号列 $\boldsymbol{y}_i$ は単語 w_i に対応する入力記号列である。

2.2　テキストコーパスの部分文字列の利用

仮名漢字変換ログの収集に用いた仮名漢字変換システムは、テキストコーパスの部分文字列を変換候補とするように拡張されている [9]。以下では、この拡張について説明する。

2.2.1　サブワードモデル

ある文字列と入力記号列との関係を記述するために、文字と入力記号列の組を単位とするサブワードモデルを用いる。このモデルでは、まず、ある表記 $w = x_1 x_2 \cdots x_m$ に対応する入力記号列を各文字 x_i の入力記号列 $\boldsymbol{y_i}$ の連接とし、次に、その出現確率 $P(\boldsymbol{y}|w)$ を各文字に対応する入力記号列が一様に出現すると仮定して、以下のように計算する。

$$P(\boldsymbol{y}|w) = P(\boldsymbol{y}|x_1 x_2 \cdots x_m) = \prod_{i=1}^{m} \frac{1}{|\boldsymbol{\mathcal{Y}}_{x_i}|} \tag{4}$$

表 1: 仮名漢字変換ログの例

入力記号列	確定結果	備考
ごいかし	語彙/ごい/RC, 化/か/IN, し/し/IN	適切な単位と入力記号列
けいたいそ	形態素/けいたいそ/RC	単位の不一致 (正: 形態, 素)
ひんしを	ひんし/ひんし/UW, を/を/IN	誤った確定 (正: 品詞, を)

IN: 一般分野の言語モデルの語彙, RC: テキストコーパスの部分文字列, UW: 未知語

ここで、$\mathcal{Y}_x$ は文字 x に対応する可能な入力記号列の集合であり、単漢字辞書を参照することで得られる。

2.2.2 文脈の記述

サブワードモデルが列挙する単語候補を適切に選択するために、その文脈を適切に記述する必要がある。このためには、仮名漢字変換を適用する分野のコーパスから言語モデルを推定することが望ましい。これを実現するために、単語分割情報がないテキストコーパスから文献 [11] の方法を用いて推定した単語 2-gram モデルを用いる。この方法では、テキストコーパスの各文字間に単語境界確率を付与し、確率的単語分割コーパスとし、単語 2-gram 確率を期待頻度から計算する。

単語境界確率は、単語分割済みコーパスから推定した最大エントロピー法に基づくモデルを用いた [11]。

テキストコーパスの部分文字列も候補にする仮名漢字変換においては、単語分割済みコーパスから推定した言語モデル P_g(式 (1)(2)) とテキストコーパスから推定した言語モデル P_r を以下のように補間して用いる。

$$P(w_i|w_{i-1}) = \lambda_g P_g(w_i|w_{i-1}) + \lambda_r P_r(w_i|w_{i-1}) \quad (5)$$

λ_g と λ_r は補間係数であり、削除補間によって求める。

2.2.3 無限語彙の仮名漢字変換

テキストコーパスの部分文字列も候補にする仮名漢字変換は、式 (3) と式 (4) で表記の候補をその生成確率とともに列挙し、式 (5) で与えられる言語モデルの確率を掛けることで得られる文全体での生成確率の降順に変換候補を提示する。

3 仮名漢字変換ログを用いた言語モデル適応

この節では、仮名漢字変換ログによる音声認識の言語モデル適応の手順について述べる。

3.1 コーパスとしての仮名漢字変換ログ

仮名漢字変換ログとは、仮名漢字変換エンジンで文章を作成するときの変換の入出力の履歴である。仮名漢字変換ログは、オープンソースな仮名漢字変換システムのソースコードに改良を加えることで容易に収集可能である。

前節で説明したテキストコーパスの部分文字列も変換候補とする仮名漢字変換システムによって得られる仮名漢字変換ログの例を表 1 に示す。仮名漢字変換ログの主な情報は、ユーザが確定した表記と入力記号列の組の列である。入力の単位は、多くの場合完全な文ではなく、文断片である。また、誤って確定した結果や、2 文字の人名などを他の単語を用いて 1 文字ずつ入力する過程などを含む。したがって、仮名漢字変換ログは、ノイズありの単語境界と入力記号列が付与された文断片からなるコーパスとして見なすことができる。

3.2 発音変換

次に、仮名漢字変換ログの各単語に付与されている入力記号列を標準的な発音 (baseform) へと変換する処理（発音変換）を行う。これは入力記号列と発音の間にある変動を解消するためである。

発音変換の手順は以下のとおりである。

1. 入力記号列を音素列に変換:
   例: オンセイ → o N s e i
2. 入力記号列の音素列を発音の音素列に変換:
   例: o N s e i ⇒ o N s e:
3. 音素列を発音に変換:
   例: o N s e: → オンセー

手順 2. における変換は重み付き有限状態トランスデューサ WFST を用いて行う。この変換を雑音のある通信路モデルでモデル化する。与えられた入力記号列の音素列 I に対して、事後確率 $P(O|I)$ を最大化する発音の音素列 O を求める。ベイズ則を用いて $P(O|I)$ を言語モデル確率 $P(O)$ と翻訳モデル確率 $P(I|O)$ に分解すると、定式化は次の通りである。

$$\hat{O} = \underset{O}{\mathbf{argmax}} P(I|O)P(O) \quad (6)$$

また、カタカナ列以外の入力記号列は音素列に変換できないため、そのような単語に関しては、音声認識システムの言語モデルの学習コーパス作成に用いる解析器で発音推定を行う。

さらに、仮名漢字変換ログに「は/ハ」、「へ/ヘ」、「を/ヲ」が出現した場合、それぞれ「は/ワ」、「へ/エ」、「を/オ」へと置換を行う。

3.3 線形補間

言語モデルの適応は、音声認識の際に用いるベースライン言語モデル $P_b(\boldsymbol{w})$ に上述の発音変換を適用した仮名漢字変換ログから構築した言語モデル $P_l(\boldsymbol{w})$ を線形補間することにより行う。適応後の言語モデル $P_a(\boldsymbol{w})$ の式は次式に従う。

$$P_a(\boldsymbol{w}) = (1-\lambda)P_b(\boldsymbol{w}) + \lambda P_l(\boldsymbol{w}) \quad (7)$$

λ は補間係数であり、削除補間によって求める。

また、言語モデルの単位として、従来の音声認識で用いられている単語ではなく、単語と発音の組を採用する。

4 評価実験

提案手法の評価のために、音声認識の実験を行った。

4.1 テストデータ

京都大学で 2007 年度後期の工学部地球工学科 3 回生向けの「資源工学のための材料学」の中から 3 回連続する同一話者による講義音声をテスト音声とした。

表 2: 解析器の学習に用いる各言語資源の諸元

コーパス	用途	単語境界	発音 or 入力記号列	文数	単語数	文字数
BCCWJ(コアデータ)	仮名漢字変換	人手	人手	36,828	997,659	1,433,904
BCCWJ(コアデータ)	音声認識	人手	人手	36,315	909,888	1,434,141
CSJ(コアデータ)	音声認識	人手	人手	21,390	1,000,282	1,658,917

表 3: 仮名漢字変換と音声認識の言語モデルの学習に用いる各言語資源の諸元

コーパス	用途	単語境界	発音 or 入力記号列	文数	単語数	文字数
BCCWJ(コアデータ)	仮名漢字変換	人手	人手	36,828	997,659	1,433,904
ウェブテキスト	仮名漢字変換	自動	自動	340,692	5,569,616	9,136,086
CSJ	音声認識	自動	自動	358,166	6,826,501	11,394,482
スライド	音声認識	自動	自動	127	1,234	2,232
仮名漢字変換ログ	音声認識	人手	人手	379	1,223	2,122

各講義は 90 分前後である。音声はピンマイクで収録した。書き起こしの単語数は 3 講義平均で 12,548 である。

4.2 解析器

後述する仮名漢字変換システムと音声認識システムの学習コーパスの作成に用いる解析器について説明する。本実験では、京都テキスト解析ツールキット KyTea 0.3.1 [7] を解析器として用いた。KyTea は文脈を考慮した読み推定、未知語への読み推定を行う。

4.3 項で説明する仮名漢字変換システムの学習コーパス作成に用いる KyTea の学習コーパスとして、単語境界と読みが人手で付与されている現代日本語書き言葉均衡コーパス BCCWJ [5] のコアデータ、未知語の読み推定のために単漢字辞書を用いた。

また、4.4 項で説明する音声認識システムの学習コーパス作成に用いる KyTea の学習コーパスとして、単語境界と発音が人手で付与されている BCCWJ のコアデータ、日本語話し言葉コーパス CSJ [6] のコアデータ、未知語の発音推定のために単漢字辞書を用いた。

KyTea の学習に用いた各言語資源の諸元を表 2 に示した。

4.3 仮名漢字変換システム

仮名漢字変換システムの言語モデルの学習コーパスとして、BCCWJ のコアデータとウェブテキストを用いた。このウェブテキストは、本講義の電子シラバスから人手で作成したキーワードをクエリとして、Yahoo! 検索 API を使って各クエリに対して、最大 20 ページを収集した。このウェブテキストに対して、KyTea を用いて、倍率 2 で疑似確率的単語分割し、さらに倍率 2 で疑似確率的読み推定を行った [10]。

仮名漢字変換システムの学習に用いた BCCWJ のコアデータとウェブテキストの諸元を表 3 中に示した。

4.4 音声認識システム

音声認識システムのベースライン言語モデルは、CSJ の学会・模擬講演 2,720 講演の書き起こしから学習した。書き起こしは 4.2 項で説明した KyTea を用いて、単語分割と発音推定を行った。言語モデルの作成には Palmkit [1]を用いた。語彙と N-gram に関するカットオフ値はそれぞれ 5, 0 とし、語彙サイズは 20,145 となった。

音響モデルは CSJ に含まれる 257 時間の学会講演の収録音声から学習した 3,000 状態、16 混合の状態共有トライフォン HMM である。この音響モデルに、ケプストラム平均正規化、ケプストラム分散正規化、声道長正規化の正規化処理を行った。すべての実験においてこの音響モデルを共通して用いる。

発音辞書に関しては、単語と発音の組を一つの語彙エントリとして発音辞書に登録する。

音声認識デコーダには、大語彙連続音声認識システム Julius 3.5.3 [4] を用いた。音声認識の評価は単語正解精度で行った。

4.5 発音変換における重み付き有限状態トランスデューサ

3.2 項で説明した発音変換に用いる WFST の言語モデルと翻訳モデルの学習データとして、解析器の学習に用いた BCCWJ のコアデータを利用した。利用手順を説明する。まず、コーパス中の単語に付与された読みと発音の頻度を計算する。次に、それらを音素列に変換し、読みと発音の音素列のアラインメントをとる。最後に、アラインメントされた音素対を言語モデルと翻訳モデルの学習データとし、両モデルとも 1-gram モデルを用いる。BCCWJ の 3 万異なり単語のうち 78 % の単語の読みと発音が一致していた。残りのうちのほとんどが長音化による変化であった。

WFST デコーダとして京都有限状態トランスデューサデコーダ Kyfd 0.2 [2] を用いた。

4.6 仮名漢字変換ログの収集

本実験では、講義者のスライド作成過程を模擬した。すなわち、第一著者がスライドにあるテキストを日本語入力することで仮名漢字変換ログを収集した。

今回、スライドのテキストを日本語入力する際に、仮名漢字変換システムにとって未知語が存在しなかったため、辞書登録する必要はなかった。仮に、BCCWJ コアデータのみを仮名漢字変換システムの学習コーパスとした場合、辞書登録の回数は 24 回になる所であったが、全ての未知語をウェブから収集したテキストコーパスの部分文字列として獲得することができた。

語彙と N-gram に関するカットオフ値はそれぞれ 1, 0 とした。語彙サイズは 3 講義平均で 20,174 となった。予備実験の結果、補間係数は 0.2 とした。

4.7 スライドを用いた言語モデルの適応

本実験では、提案手法との比較対象を、スライドを利用して言語モデルの補間を行う手法 [8] とした。電子的な講義スライドからは簡単にテキスト情報を抽出することができるため、スライドは広く利用可能な資料である。スライドの解析に 4.2 項で説明した KyTea

[1] http://palmkit.sourceforge.net/ (2012/1/25 アクセス)

[2] http://www.phontron.com/kyfd/ (2012/1/25 アクセス)

表 4: 各手法による未知語率、パープレキシティ、認識精度

手法	未知語率 [%]	パープレキシティ	認識精度 [%]
B	5.66	189	62.51
S	2.85	137	68.94
L	2.21	139	70.11

B:ベースライン S:スライド適応 L:仮名漢字変換ログ適応

を用いた。カットオフ値は仮名漢字変換ログの場合と同様にした。スライドによる適応後の言語モデルの語彙サイズは 3 講義平均で 20,175 となった。予備実験の結果、補間係数は 0.2 とした。

以上、音声認識システムの学習に用いた言語資源の諸元を表 3 中に示す。ただし、仮名漢字変換ログにおける文数とは文断片の数を意味する。

4.8 実験結果

表 4 中に 3 講義平均の実験結果を示す。ベースライン言語モデルを B、スライドによる適応言語モデルを S、仮名漢字変換ログによる適応言語モデルを L としている。

まず、未知語率とパープレキシティである。L は S よりも未知語率に関して 0.64 ポイント上回った。表記としてはほぼ同じテキストを用いているのにも関わらず、この差が生じているのは、スライドを単語分割したときに単語分割誤りが発生しているのが原因であると考えられる。また、パープレキシティに関しては、未知語率の差に関わらずほぼ同じ値になった。これはスライドの方が仮名漢字変換ログよりも長い N-gram を観測するからであると考えられる。

次に、認識精度である。L は S よりも 1.17 ポイント上回った。この結果は、仮名漢字変換ログが認識精度に効果があることを示している。よって、仮名漢字変換ログを用いることは有用である。

4.9 仮名漢字変換ログによる改善

仮名漢字変換ログは専門用語を含んだ単語境界と入力記号列の情報を含んでいる。これは、仮名漢字変換システムのユーザが日本語入力したい単語を仮名漢字変換の確定結果として選択ないし辞書登録といった操作を行うからである。そのため、「き裂」、「へき開」などの単語分割が難しい単語を 1 つの単語として獲得することが可能である。また、発音推定が困難な単語に対しても有用であると言える。

仮名漢字変換ログの利用により、実際に得られた単語分割と発音推定の改善例を表 5 と表 6 に示す。

こういった操作を仮名漢字変換において行うことは、アノテーションデータを作成して解析器の辞書に追加することと同等であると見なせる。しかしながら、前者の方がユーザが日常的に行なっている操作であるため、ユーザにとって負担が少ないと考えられる。

したがって、仮名漢字変換ログを用いることは有用である。

5 おわりに

本稿では、講義の音声認識を目的として、講義資料を作成する際の仮名漢字変換ログを活用して言語モデルを専門分野に適応する方法を提案した。提案手法を使った評価実験において、一定の有効性を示した。

表 5: 仮名漢字変換ログによる単語分割の改善例

スライド	仮名漢字変換ログ
き 裂 先端	き裂 先端
でき 裂	で き裂
せん 断	せん断
へき 開破 面	へき開 破面
板 厚	板厚
弾塑 性	弾塑性

表 6: 仮名漢字変換ログによる発音推定の改善例

単語	スライド	仮名漢字変換ログ
板厚	バンアツ	イタアツ
切欠	キカキ	キリカキ
塑性	デクセー	ソセー
破面	ハズラ	ハメン
変位	カイ	ヘンイ

本研究は、講義に限らず、発話内容と同様のテキストを入力する場合に有効である。例えば、予稿やスライドなどの資料を事前に作成する講演やプレゼンテーションの音声認識、さらに音声検索などにも応用可能である。

参考文献

[1] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. of the InterSpeech*, pp. 2553–2556, 2007.

[2] J. Glass, T. J. Hazen, L. Hetherington, and C. Wang. Analysis and Processing of Lecture Audio Data: Preliminary Investigations. In *Proc. of the WIASIR at HLT-NAACL*, pp. 9–12, 2004.

[3] T. Kawahara, N. Katsumaru, Y. Akita, and S. Mori. Classroom Note-taking System for Hearing Impaired Students using Automatic Speech Recognition Adapted to Lectures. In *Proc. of the InterSpeech*, pp. 626–629, 2010.

[4] A. Lee and T. Kawahara. Recent Development of Open-Source Speech Recognition Engine Julius. In *APSIPA ASP*, pp. 131–137, 2009.

[5] K. Maekawa. Balanced Corpus of Contemporary Written Japanese. In *Proc. of the WALR*, pp. 101–102, 2008.

[6] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous Speech Corpus of Japanese. In *Proc. of the LREC*, pp. 947–952, 2000.

[7] G. Neubig and S. Mori. Word-based Partial Annotation for Efficient Corpus Construction. In *Proc. of the LREC*, 2010.

[8] I. Trancoso, R. J. F. Nunes, L. Neves, C. Viana, H. Moniz, D. Caseiro, and A. I. Mata. Recognition of Classroom Lectures in European Portuguese. In *Proc. of the InterSpeech*, 2006.

[9] 森信介. 無限語彙の仮名漢字変換. 情処論, 48(11):3532–3540, 2007.

[10] 森信介, 笹田鉄郎, G. Neubig. 確率的タグ付与コーパスからの言語モデル構築. 自然言語処理, 18(2):71–87, 2011.

[11] 森信介, 小田裕樹. 擬似確率的単語分割コーパスによる言語モデルの改良. 自然言語処理, 16(5):7–21, 2009.

[12] 森信介, 土屋雅稔, 山地治, 長尾真. 確率的モデルによる仮名漢字変換. 情報処理学会研究報告. 自然言語処理研究会報告, 98(48):93–99, 1998.

[13] 土屋雅稔, 小暮悟, 西崎博光, 太田健吾, 山本一公, 中川聖一. 日本語講義音声コンテンツコーパスの作成と分析. 情処論, 50(2):448–459, 2009.

[14] 藤井康寿, 山本一公, 北岡教英, 中川聖一. 重要文抽出に基づく講義音声の自動要約. 情処論, 51(3):1094–1106, 2010.