未知語を含む文脈情報の自動獲得による統計的仮名漢字変換システムの分野適応

笹田 鉄郎　　森 信介　　河原 達也

京都大学 情報学研究科 知能情報学専攻

1　はじめに

近年、自然言語処理や音声言語処理の分野においては、言語コーパスの統計情報を用いて言語現象の確率的モデルを構築し、言語処理のシステムに用いることが広く行われている。しかし、コーパス中に存在しない言語情報の取り扱いに関しては未だに課題が残っている。特に音声認識と統計的仮名漢字変換のように、処理過程において単語の読み[1]が必須となるシステムを用いるためには、単語分割と読み付与の両方を行った適応コーパスの作成を行うことが必要となる。しかし、これを既存の形態素解析器などを用いて自動的に行うと、未知語の周辺で解析誤りが多く発生するため、読みの付与を含めた完全な自動化は難しく、最終的には修正コストがかかるという問題がある。

本論文では、内容の類似したテキストと音声を用いて未知語の文脈情報を考慮した適応コーパスを作成する手法について述べ、それを用いた統計的仮名漢字変換システムの自動分野適応が可能であることを示す。

2　n-gram 言語モデルとその応用

本節では、基本的なモデルの一つである n-gram と、統計的仮名漢字変換 [1] の枠組みについて述べる。

2.1　単語 n-gram モデル

単語 n-gram モデルは最も一般的な確率的言語モデルの一つである。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予測する。

$$M_{w,n}(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) \qquad (1)$$

式 (1) において w_i $(i \leq 0)$ は、文頭を表す特別な記号であり、w_{h+1} は、文末を表す特別な記号である。単語 n-gram モデルにおいてあらゆる語彙を定義することは不可能であるため、未知語を表す特別な記号 UW を用意して、モデル構築の際に他の語彙と同様に 0 より大きい確率を与えておく。未知語を予測する際は、まず単語 n-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 n-gram モデルにより予測する。

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

式 (2) において x_i $(i \leq 0)$ と $x_{h'+1}$ は、それぞれ語頭と語末を表す特別な記号である。

2.2　統計的仮名漢字変換

仮名漢字変換は、日本語の入力システムにおいてキーボードから直接入力可能な記号の列 $\boldsymbol{y}$ を入力とし、変換結果を仮名漢字混じり文 $\boldsymbol{x}$ として出力する。森ら [1] の提案した確率的モデルによる統計的仮名漢字変換は、$\boldsymbol{y}$ を入力として、変換候補文字列 $(\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots)$ を確率 $P(\boldsymbol{x}|\boldsymbol{y})$ の降順に提示する。この入出力関係は音声認識と基本的に同様のものであり、$P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$ の計算による順序付けを行うことで変換候補の提示が可能となる。ここで、後半の $P(\boldsymbol{x})$ は確率的言語モデルであり、2.1 節の単語 n-gram モデルを用いることができる。また前半の $P(\boldsymbol{y}|\boldsymbol{x})$ は確率的仮名漢字モデルと呼ばれ、日本語文 $\boldsymbol{x}$ が与えられた際の入力記号列 $\boldsymbol{y}$ の確率を表す。これは、日本語文を単語列 $\boldsymbol{w}$ とみなし、単語と入力記号列との対応関係がそれぞれ独立であると仮定することで以下の式で表される。

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

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り、$\boldsymbol{y} = \boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_h$ を満たす。

式 (3) における確率 $P(\boldsymbol{y}_i|w_i)$ の値は以下の式を用いて最尤推定する。

$$P(\boldsymbol{y}_i|w_i) = \frac{f(\boldsymbol{y}_i,\ w_i)}{f(w_i)}$$

$f(e)$ は事象 e のコーパス内頻度である。上記の計算は単語ごとに入力記号列が付与されたコーパスから頻度を計数することで可能となる。本論文の焦点は、テ

[1] 厳密には、音声認識には単語の音素列、仮名漢字変換にはキーボードから入力可能な記号列が必要となる。本論文では前者を読み、後者を入力記号列と呼ぶ。

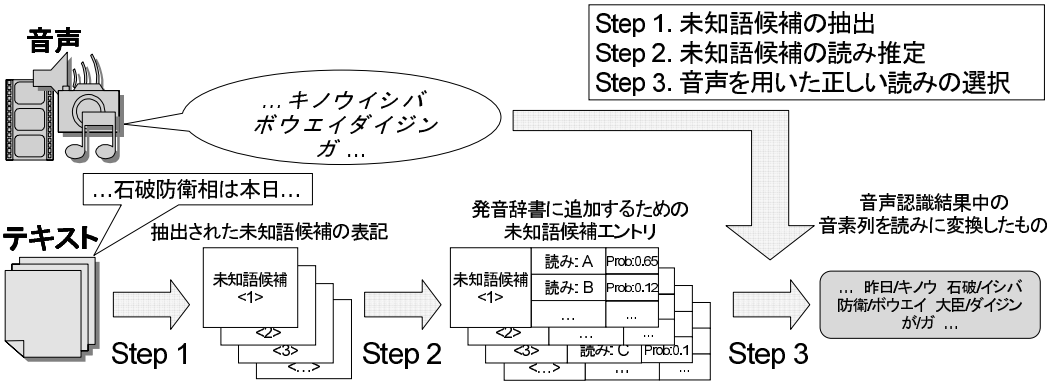

図 1: 未知語を含むコーパスの自動獲得

キストの表層からは得られない入力記号列の情報を、音声情報から得られる読みによって代替し、変換精度の向上を目指すというものである。

3 未知語を含むコーパスの自動獲得

本節では、テキストと音声から未知語とその読みを自動獲得する手法 [2] について述べる。図 1 に手法の概要を示す。

3.1 未知語候補の抽出

一般的に、単語分割を行う際はテキスト中に分野特有の未知語があるとその周辺で解析を誤ることが多く、結果として必要な未知語が単語として得られないことがある。そこで本手法では生テキストから確率的単語分割コーパス [3] を作成し、さらに疑似確率的単語分割コーパス [4] を作成することで未知語候補の表記を抽出する。

確率的単語分割コーパスはコーパス中の全ての文字列間に単語境界確率が付与されたコーパスであり、ここでは最大エントロピーモデルに基づいて単語境界確率の推定を行う [4]。確率的単語分割コーパスはテキスト内のあらゆる部分文字列を単語として取り扱うため、必要な未知語の出現頻度がゼロになるという状況は発生しない。ただし、確率的単語分割コーパスからの n-gram 確率の計算には多くの計算量が必要となる、また既存の n-gram 確率計算用ツールを用いることができない、という 2 点を考慮し、本手法では確率的単語分割コーパスを決定的に単語分割されたコーパスで近似した疑似確率的単語分割コーパスを用いる。

疑似確率的単語分割コーパスは単語境界確率と乱数 $r(0 \leq r < 1)$ を比較し、その大小で各文字間の境界が単語境界であるかどうかを順次決定するという処理を確率的単語分割コーパス全体に対して行い、それを複数回（M 回）行うことで得られる。疑似確率的単語分割コーパス中の未知語のうち、F_{th} 回以上出現したものを未知語候補として抽出する。

3.2 未知語候補の読み推定

抽出した未知語候補の読みを n-gram モデルにより複数推定する。以下では「石破」が未知語候補である場合を例にとって説明する。

1. 単語を 1 文字ごとに分割し、それぞれの文字について単漢字辞書から得られる読みを列挙する。
   ex.) 石 (イシ, セキ), 破 (ヤブ, ハ, バ)
2. 各文字の読みを組み合わせ、可能性のある単語の読みを列挙する。
   ex.) イシヤブ, イシハ, イシバ,
   セキヤブ, セキハ, セキバ
3. 文字と読みの組を単位とする n-gram モデルにより、単語と読みの同時確率を計算する。
   ex.) $P(\text{イシヤブ}, \text{石破}) = 0.53$
   $P(\text{イシバ}, \text{石破}) = 0.22$
   $\vdots$

例に示しているように、テキスト中における「石破」の正しい読みが「イシバ」である場合、n-gram モデルを用いた読み推定のみでは正しく読みを推定することができない。しかし、複数推定した読みの中に正しいものが含まれていれば、それを以下に示す方法で選択することができる。

3.3 音声を用いた正しい読みの選択

前項で列挙された読みのうち、どれが正しいかを音声認識によって判定する。一般的に音声データには明確な区切りが無く、また雑音成分を多く含む。そのため音声認識においては似た発音の単語を取り違えて

表 1: 単語分割、読み付与済みのコーパス

文数	単語数	文字数
14,645	485,604	693,156

表 2: 未知語抽出と言語モデル推定に用いるテキスト

文数	文字数
30,552	2,528,722

認識することがしばしば起こる。この問題に対処するためには単語の文脈を考慮する必要があり、大語彙連続音声認識システムを用いる場合には、ドメインを限定して言語モデルの学習を行うことが一般的である。本論文では未知語候補の抽出元となったテキストを用いて適応用の言語モデルを作成し、テキストと同じ話題を扱った音声を用意する。また、実験で用いる大語彙連続音声認識システム Julius [5] の音声認識結果には、単語ごとに信頼度 [6] が付与されている。音声認識結果には一般に多くの認識誤りが蓄積されるため、この単語信頼度を用いて認識結果のフィルタリングを行うことで、認識誤りに対してより頑健な単語と読みの抽出を行う。

以下に構築したシステムを用いて未知語候補とその読みを含むコーパスを得る手順を示す。

1. テキストと同じ話題を扱った音声と、それに合った音声認識用の音響モデルを用意する。
2. 既知語リストならびに未知語候補とその読み推定結果から音声認識用の発音辞書を作成する。
3. 未知語抽出時に作成した疑似確率的単語分割コーパスと一般分野のコーパスを用いて音声認識用言語モデルを作成する。
4. 1〜3 の音響モデル、言語モデル、発音辞書を用いて 1 の音声に対し音声認識を行う。
5. 音声認識の出力の中から、音声認識の信頼度が C_{th} 以上の部分単語列を抽出する。
6. 各単語に対応する音素列を読みに変換する。

以上の処理により、テキストと音声に共通して現れる未知語の表記と読みを含めたコーパスを獲得できる。これは 2.2 節で述べた仮名漢字変換モデルのための学習コーパスとしてそのまま用いることができる。

4 評価

本節では、まずウェブニュースを対象として未知語を含めた音声の認識を行い、コーパスとして獲得する実験の条件について述べる。次に、仮名漢字変換による精度評価について述べる。

表 3: 音声認識用の発音辞書

	単語数	エントリ数
既知語	17,208	17,826
未知語候補	3,504	9,054

4.1 未知語を含む音声の認識

単語境界確率を推定するために、あらかじめ単語分割が行われたコーパスが必要となる。本実験では「現代日本語書き言葉均衡コーパス」[2]（以下、BCCWJ と略す）を用いた。ここでは自動解析結果の人手による修正が行われているものだけを用いており、これは仮名漢字変換の言語モデルならびに仮名漢字モデル学習時にも用いられる。また、既知語リストとしてコーパス内に出現する全ての単語と読みの組の集合を用いた。BCCWJ の詳細を表 1 に示す。

未知語候補の抽出元となる単語境界と読みの情報が共にないテキストとして、2007 年 11 月 2 日から 2008 年 1 月 8 日のうち、68 日間のウェブニュース記事を用意した。ここから擬似確率的単語分割コーパスを $M = 10$ として作成し、未知語候補を抽出した。未知語候補を決定する際、閾値 F_{th} は、単語の読み推定による発音辞書のサイズ増大 (3.2 節参照) を考慮して調節し、$F_{th} = 50$ とした。実験で用いたウェブニューステキストの詳細を表 2 に示す。ここで作成した単語分割コーパスは音声認識の言語モデル推定にも用いられる。既知語リストと未知語候補の読みを合わせて作成された発音辞書の詳細を表 3 に示す。

未知語候補の獲得に用いる音声認識システムとして、Julius 3.5.3 を用いた。音響モデルには新聞記事読み上げ音声コーパス（JNAS）から学習した 3,000 状態、64 混合の triphone HMM を用いた。

正しい読みを選択するために用いる音声として、2007 年 12 月 5 日から 2008 年 1 月 8 日の間に放送された、1 日 30 分のニュース番組の音声 34 日分を用いた。

音声認識結果のうち、単語信頼度が一定値を超えている部分 ($C_{th} > 0.1$) を分野適応コーパスとして獲得した。

4.2 仮名漢字変換による評価

仮名漢字変換の評価指標として、文字ごとの再現率と適合率を用いた。実験では、表 1 に示したコーパス (BCCWJ) に表 2 のウェブニュースから作成した疑似確率的単語分割コーパスを加えたものから言語モデル

[2]「現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese) 」モニター公開データ（2008 年度版）

表 4: 仮名漢字変換による評価

	再現率	適合率
baseline	94.36(%)	92.33(%)
	= 68,900/73,020	= 68,900/74,623
baseline	96.07(%)	94.69(%)
+ rec.	= 70,147/73,020	= 70,147/74,077

を学習し、仮名漢字モデルを BCCWJ のみから学習した場合をベースラインとした。BCCWJ のみから両方のモデルを学習した場合をベースラインとしてもよいが、疑似確率的単語分割コーパスの追加による仮名漢字変換精度の向上はすでに報告されており [4]、音声を用いることなく実現可能であるため、本実験では上記の設定をベースラインとした。ベースラインとの比較対象として、4.1 節で獲得したコーパスをベースラインに追加したものを用いた。なお、本実験における言語モデルは単語 2-gram モデルとした。テストセットとして、2008 年 1 月 9 日、10 日のウェブニュース 888 文に読みを付与したものを用いた。実験結果を表 4 に示す。

本論文で提案した手法を用いて得られた音声認識結果を適応コーパスとして用いることで、再現率が 1.71%、適合率が 2.36%向上した。また、変換誤り文字数の減少率は 4,120 文字中 1,247 文字 (30.3%) であり、音声を用いることで未知語を含めた読みの情報が正しく獲得されていることが確認された。また、未知語のなかには 3.2 節で正しい読みの確率が最大にならなかった（テキストのみでは正しい読みの特定ができなかった）ものも含まれており、音声を用いた読みの選択が適切に行われていることがわかった。

以上の結果から、本手法で自動獲得した未知語とその読み、文脈情報を用いて仮名漢字変換の言語モデルならびに仮名漢字モデルを更新し、分野適応を行うことが可能であることが確認された。

5 関連研究

テキストから未知語とその読みを獲得する手法として、倉田ら [7] は音声認識システムを対象として、本手法と同様の手法で未知語の獲得を行い、音声認識を行っている。しかし、読みの不明な未知語への対策という観点から見ると、音声認識結果からは単語レベルの情報を取り出すのみにとどまっている。それに対して提案手法では、音声認識結果をコーパスという形で追加し、言語モデルの再学習を行うため、読みを合わせた未知語周辺の文脈情報というより多くの情報を用いてシステムの改善を行うことができる。

森ら [8] は仮名漢字変換を用いる際の入力とその変換結果から未知語の獲得と言語モデルの更新を行うことを繰り返し、それによって仮名漢字変換システムの精度が徐々に向上すると報告している。ここで行われている実験はユーザによるシステムの利用を想定したシミュレーションであり、本論文で示した音声とテキストを入力とする自動獲得とは異なっている。

6 おわりに

本論文では、仮名漢字変換の自動分野適応を目的とし、音声認識を用いて読みの付与されたコーパスを自動作成する手法を提案した。実験の結果、音声認識結果をコーパスとしてそのまま利用することで言語モデルと仮名漢字モデルを自動的に更新し、変換精度を向上させることが可能であることを確認した。

参考文献

[1] 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, 情報処理学会論文誌, Vol. 40, No. 7, pp. 2946–2953 (1999).

[2] 笹田鉄郎, 森信介, 河原達也：テキストと音声を用いた単語と読みの自動獲得, 情報処理学会研究報告, SLP-72 (2008).

[3] 森信介, 宅間大介, 倉田岳人:確率的単語分割コーパスからの単語 N-gram 確率の計算, 情報処理学会論文誌, Vol. 48, No. 2, pp. 892–899 (2007).

[4] 森信介, 倉田岳人, 小田裕樹：最大エントロピー法による単語境界確率の推定, 情報処理学会研究報告, SLP-63 (2006).

[5] Lee, A., Kawahara, T. and Shikano, K.: Julius – An Open Source Real-Time Large Vocabulary Recognition Engine, *Proc. of the EuroSpeech2001*, pp. 1691–1694 (2001).

[6] 李晃伸, 河原達也, 鹿野清宏:2 パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法, 情報処理学会研究報告, SLP-49 (2003).

[7] 倉田岳人, 森信介, 西村雅史:講義関連コーパスを利用した音声認識システムの自動適応, 電子情報通信学会論文誌, Vol. J90-D, No. 9, pp. 1780–1789 (2005).

[8] 森信介, 小田裕樹：自動未知語獲得による仮名漢字変換システムの精度向上, 言語処理学会第 13 回年次大会 (2007).