自動獲得した未知語の読み・文脈情報による仮名漢字変換

笹田 鉄郎[†] ・森 信介[†,††] ・河原 達也[†,††]

未知語の問題は、仮名漢字変換における重要な課題の 1 つである。本論文では、内容の類似したテキストと音声から未知語の読み・文脈情報をコーパスとして自動獲得し、仮名漢字変換の精度向上に利用する手法を提案する。まず、確率的な単語分割によって未知語の候補となる単語をテキストから抽出する。次に、各未知語候補の読みを複数推定して列挙する。その後、テキストに類似した内容の音声を認識させることによって正しい読みを選択する。最後に、音声認識結果を学習コーパスとみなして仮名漢字変換のモデルを構築する。自動収集されたニュース記事とニュース音声を用いた実験では、獲得した未知語の読み・文脈情報を仮名漢字変換のための学習コーパスとして用いることで、精度が向上することを確認した。

キーワード ：未知語 音声認識 仮名漢字変換

Kana-Kanji Conversion by Using Unknown Word-Pronunciation Pairs with Contexts

Tetsuro SASADA[†], Shinsuke MORI[†,††] and Tatsuya KAWAHARA[†,††]

One of the significant problems of kana-kanji conversion (KKC) systems is unknown words. In this paper, for the purpose of improvement in KKC accuracy, we propose a method for extracting unknown words, their pronunciations and their contexts from similar sets of Japanese text data and speech data. Unknown word candidates are extracted from text data with a stochastic segmentation model, and their possible pronunciation entries are hypothesized. These entries are verified by conducting automatic speech recognition (ASR) on audio data on similar topics. As a result of ASR, we obtain a corpus for training a stochastic model for KKC. In the experiment, we use automatically-collected news articles and broadcast TV news covering similar topics. We made experimental evaluations with our KKC back-end enhanced with these corpora on other web news articles and observed an improvement in the accuracy.

Key Words: unknown word, automatic speech recognition, kana-kanji conversion

[†]京都大学 情報学研究科, Graduate School of Informatics, Kyoto University
[††]京都大学 学術情報メディアセンター, Academic Center for Computing and Media Studies, Kyoto University

1　はじめに

計算機の急速な普及に伴い、様々な自然言語処理システムが一般に用いられるようになっている。中でも、日本語の仮名漢字変換は最も多く利用されるシステムの 1 つである。仮名漢字変換の使いやすさは変換精度に大きく依存するため、常に高精度で変換を行うことが求められる。近年では、変換精度の向上とシステム保守の効率化を両立させるために、確率的言語モデルに基づく変換方式である統計的仮名漢字変換 (森，土屋，山地，長尾 1999) が広まりつつある。

変換精度を向上させる上で問題となるのは、多くの言語処理システムと同様、未知語の取り扱いである。統計的仮名漢字変換では、文脈情報を反映するための単語 n-gram モデル、入力である読みと出力である単語表記の対応を取るための仮名漢字モデルの 2 つのモデルによって出力文候補の生成確率を計算し、候補を確率の降順に提示するが、未知語（単語 n-gram モデルの語彙に含まれない単語）を含む候補の生成はできない。この問題に対処して変換精度を向上させる一般的な方法は、仮名漢字変換の利用対象分野における未知語の読み・文脈情報を用いたモデルの改善である。

仮名漢字変換の利用対象となる分野は多岐に渡っており、未知語の読み・文脈情報を含む対象分野の学習コーパスがあらかじめ利用可能であるという状況は少ない。このため、情報の付与されていない対象分野のテキストに必要な情報を付与して学習コーパスを新たに作成するということが行われる。しかしながら、未知語の中には、読みや単語境界をテキストの表層情報から推定することが困難な単語が少なからず存在する。このような場合には、対象分野の学習コーパスを作成するためにその分野についての知識を有する作業者が必要となるなど、コストの面で問題が多い。

上記の問題を解決するために、本論文では、テキストと内容の類似した音声を認識することで未知語の読み・文脈情報を単語とその読みの組として自動獲得し、統計的仮名漢字変換の精度を向上させる手法を提案する。以下に手法の概略を述べる。まず、情報の付与されていない対象分野のテキストから、未知語の出現を考慮した単語分割コーパスである疑似確率的単語分割コーパスを作成し、未知語候補の抽出を行う。次に、疑似確率的単語分割コーパスから音声認識のための言語モデルを構築するとともに、未知語候補の読みを複数推定・列挙し、発音辞書を作成する。その後、言語モデルと発音辞書を用いて対象分野の音声を認識し、音声認識結果から単語と読みの組の列を獲得する。最後に、獲得した単語と読みの組の列を統計的仮名漢字変換の学習コーパスに追加して言語モデルと仮名漢字モデルを更新する。

実験では、統計的仮名漢字変換のモデル構築に用いる一般分野のコーパスに、獲得した未知語の読み・文脈情報を追加し、モデルを再構築することで変換精度が向上することを確認した。本論文で提案する枠組みは、対象分野のテキストと音声の自動収集が可能であるという前提のもとで、未知語に対して頑健なモデルを構築することができるため、統計的仮名漢字変換の効率的かつ継続的な精度向上に有効である。

2　単語 n-gram モデルとその応用

確率的言語モデルとは、任意の記号列[1]に対して、その記号列がある自然言語から生成された確率を計算する枠組みを与えるためのモデルである (北 1999)。本節では、最も一般的な確率的言語モデルの 1 つである単語 n-gram モデルとその応用について述べる。

2.1　単語 n-gram モデル

本項では、確率的言語モデルとして広く用いられる単語 n-gram モデルならびにモデルパラメータの推定について述べる。

単語 n-gram モデルは、文を単語列 $\boldsymbol{w} = w_1 w_2 \cdots w_h$ とみなし、単語の生起を $(n-1)$ 重マルコフ過程で近似したモデルである。すなわち、単語 n-gram モデルにおいて、ある時点での単語 w_i の生起は直前の $(n-1)$ 単語に依存する。ここで、単語列 $\boldsymbol{w}$ の生成確率 $M_{w,n}(\boldsymbol{w})$ は以下の式で与えられる。

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

この式で、w_i $(i \leq 0)$ と w_{h+1} はそれぞれ文頭と文末を表す特別な記号である。

言語モデル構築の際には、学習コーパス内で観測されたデータの生じる確率を最大にするように最尤推定法でモデルパラメータを決定することが一般的である。最尤推定で単語 n-gram モデルのパラメータ推定を行う場合は、あらかじめ単語分割されているコーパス内に出現する単語 n-gram の頻度を計数し、以下の式によって単語 n-gram の確率を求める。

$$P(w_i) = \frac{f(w_i)}{f(\cdot)} \quad (\text{if} \quad n = 1) \tag{1}$$

$$P(w_i | \boldsymbol{w}_{i-n+1}^{i-1}) = \frac{f(\boldsymbol{w}_{i-n+1}^{i})}{f(\boldsymbol{w}_{i-n+1}^{i-1})} \quad (\text{if} \quad n > 1) \tag{2}$$

式 (1) において、$f(w_i)$ はコーパス内の単語 w_i の出現頻度（1-gram 頻度）を表し、$f(\cdot)$ はコーパス内における全ての単語の出現頻度（0-gram 頻度）を表す。式 (2) において、$f(\boldsymbol{w}_{i-n+1}^{i})$ はコーパス内における連続する n 単語の組の出現頻度（n-gram 頻度）を表す。

ここで、未知語を含む単語列の生成確率を単語 n-gram モデルで計算する場合を考える。未知語を含む単語列の生成確率が 0 となることを防ぐため、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エントリと同様に 0 より大きい確率を与えておく。未知語を予測するには、

[1] ここで述べる「記号」は処理単位としての記号であり、文字や単語、品詞など様々なものを考えることができる。

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)$\boldsymbol{x}$ を以下の文字 n-gram モデルにより予測する。

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

ここで 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。

本項で述べた n-gram モデルの応用として、文献 (永田 1999a) では日本語や中国語のように分かち書きされない言語に対する形態素解析器を提案している。また、文献 (長野，森，西村 2006) では、文献 (永田 1999a) で提案された手法の拡張として、式 (1)(2) における w_i を単語、読み、アクセント、品詞の 4 つ組に置き換えた n-gram モデルによってテキストの読みとアクセントの推定を行うシステムを提案している。

2.2　統計的仮名漢字変換

本項では、(森他 1999) で提案されている確率的モデルを用いた統計的仮名漢字変換について述べる。

日本語の仮名漢字変換システムは、計算機のキーボードからの入力記号列[2]$\boldsymbol{z}$ を仮名漢字混じり文である文字列 $\boldsymbol{x}$ に変換する。ここでは、出力を文字列 $\boldsymbol{x}$ とする代わりに単語列 $\boldsymbol{w}$ とし、入力記号列 $\boldsymbol{z}$ に対応する候補 $\boldsymbol{w}$ を以下に示す事後確率 $P(\boldsymbol{w}|\boldsymbol{z})$ が大きいものから順に列挙する。

$$P(\boldsymbol{w}|\boldsymbol{z}) = \frac{P(\boldsymbol{z}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{z})} \tag{3}$$

最尤の変換結果 $\hat{\boldsymbol{w}}$ は、$P(\boldsymbol{w}|\boldsymbol{z})$ をベイズの定理により以下のように変形することで求めることができる。

$$\hat{\boldsymbol{w}} = \mathop{\mathbf{argmax}}_{\boldsymbol{w}} P(\boldsymbol{z}|\boldsymbol{w})P(\boldsymbol{w}) \tag{4}$$

式 (4) において、後半の $P(\boldsymbol{w})$ は言語モデルであり、2.1 節で述べた単語 n-gram モデルを用いることができる。前半の $P(\boldsymbol{z}|\boldsymbol{w})$ は確率的仮名漢字モデルと呼ばれ、単語列 $\boldsymbol{w}$ が与えられた際の入力記号列の生成確率を表す。ここで述べている変換モデルでは出力を文字列 $\boldsymbol{x}$ ではなく単語列 $\boldsymbol{w}$ とみなしているため、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 $P(\boldsymbol{z}|\boldsymbol{w})$ は以下の式で表される。

$$P(\boldsymbol{z}|\boldsymbol{w}) = \prod_{i=1}^{h} P(z_i|w_i) \tag{5}$$

ここで、部分入力記号列 z_i は単語 w_i に対応する入力記号列であり、全体の入力記号列は $\boldsymbol{z}$ =

[2]入力記号列の記号とは、キーボードから入力可能なラテン文字、記号、仮名文字を表す。

$z_1 z_2 \cdots z_h$ となる。

仮名漢字モデルのパラメータ推定には、単語ごとに入力記号列が付与されたコーパスを用い、式 (5) における確率 $P(z_i|w_i)$ の値は、以下の式によって計算される。

$$P(z_i|w_i) = \frac{f(z_i,\ w_i)}{f(w_i)} \tag{6}$$

ここで $f(z_i,\ w_i)$ は単語と読みの組の出現頻度であり、$f(w_i)$ は単語出現頻度である。

2.3 確率的単語分割コーパス

n-gram モデルの性能はパラメータ学習のためのコーパスに大きく依存する。しかし、決定的な単語分割を行うコーパスを単語 n-gram モデルのパラメータ推定に用いる場合、分割誤りによって未知語の出現頻度が 0 となっている可能性がある。このようなコーパスから構築される単語 n-gram モデルは未知語に対する頑健性に欠けるため、本項では、確率的単語分割コーパス並びにその近似である疑似確率的単語分割コーパス (森，小田 2009) の枠組みを用いてこの問題に対処する方法を述べる。

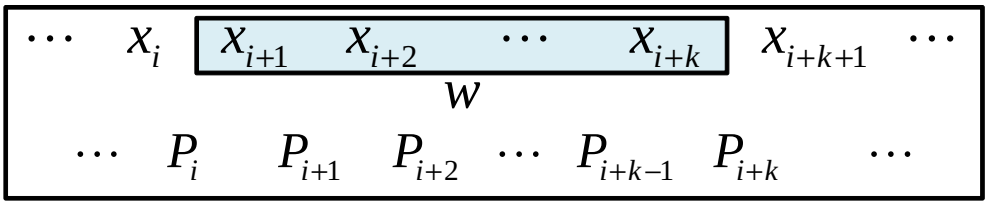
確率的単語分割コーパスを用いた n-gram 確率の推定　日本語の単語分割は、入力文における各文字間に単語境界があるかどうかを決定する問題とみなせる。入力となるコーパスを長さ n_r の文字列 $\boldsymbol{x}_1^{n_r} = x_1 x_2 \cdots x_{n_r}$ としたとき、確率的単語分割コーパスは隣接する 2 文字 x_i と x_{i+1} の間に単語境界確率 P_i を与えたものとして定義される。ここでは、確率的単語分割コーパスを作成するために最大エントロピーモデルを用いて単語境界確率の推定を行う (森，小田 2009)。単語境界をある文字列境界が単語境界であるか否かを決めるための素性として、単語境界の周辺 x_{i-2}^{i+2} の範囲の文字 n-gram（$n = 1, 2, 3$）と文字種の情報を用いる。

ここで、確率的単語分割コーパス内での単語の扱いについて述べる。決定的に単語分割されたコーパスにおいて、単語 0-gram 頻度はコーパス中の全単語数、単語 1-gram 頻度はそれぞれの単語の出現頻度である。確率的単語分割コーパスにおいては、単語 0-gram 頻度 $f(\cdot)$ はコーパス中に現れる全ての部分文字列の期待頻度として、以下の式で定義される。

$$f(\cdot) = 1 + \sum_{i=1}^{n_r - 1} P_i$$

また、確率的単語分割コーパス中のある 1 箇所に現れる単語 w の期待頻度 $f(w)$ は、文字列 $x_{i+1} x_{i+2} \cdots x_{i+k}$ が単語 w である確率を以下に示す式から計算することで得られる。

$$f(w) = P_i \left[\prod_{j=1}^{k-1} (1 - P_{i+j}) \right] P_{i+k}$$



$$f(w) = P_i \times (1 - P_{i+1}) \times \cdots (1 - P_{i+k-1}) \times P_{i+k} \qquad (w = x_{i+1}^{i+k})$$

図 1　確率的単語分割コーパスにおける期待頻度

これは x_{i+1} の左側 (i 番目の文字列境界) が単語境界、$x_{i+1}x_{i+2}\cdots x_{i+k}$ の間にある文字列境界が単語境界ではない、x_{i+k} の右側が単語境界である、というときに文字列 $x_{i+1}x_{i+2}\cdots x_{i+k}$ が単語 w である確率を示している。確率的単語分割コーパス中における単語 w とその期待頻度の扱いを図 1 に示す。$f(w)$ は 1 箇所の w に対する期待頻度なので、単語 1-gram 期待頻度はコーパス中の全ての出現にわたる期待頻度の合計となる。単語 n-gram 期待頻度（$n \geq 2$）についても、単語境界である確率 P_i と単語境界ではない確率 $(1 - P_i)$ から同様に期待頻度の計算を行う。単語 n-gram 確率は、式 (1)(2) における n-gram 頻度を n-gram 期待頻度として推定する。

以上に述べた確率的単語分割コーパスから構築される単語 n-gram モデルは、テキスト中に出現する全ての部分文字列が語彙となるため、未知語に対して頑健なモデルとなる。

疑似確率的単語分割コーパス　上述の確率的単語分割コーパスを用いて n-gram 確率の推定を行う場合、単語の出現頻度を計算するために多くの計算時間が必要となる。

本節では、この問題に対処するために提案されている疑似確率的単語分割コーパス (森，小田 2009) の枠組みについて述べる。これにより、決定的に単語分割されたコーパスを用いて確率的単語分割コーパスに近い n-gram 確率を推定することができ、かつ未知語に対する頑健性を保持することができる。

疑似確率的単語分割コーパスは、確率的単語分割コーパスに対して以下の処理を最初の文字から最後の文字まで $(1 \leq i \leq n_r)$ 行うことで得られる。

(1)　文字 x_i を出力する。

(2)　乱数 $r_i (0 \leq r_i < 1)$ を発生させ P_i と比較する。$r_i < P_i$ の場合には単語境界記号 (空白) を出力し、そうでない場合には何も出力しない。

これにより、確率的単語分割コーパスの特徴をある程度反映し、かつ決定的に単語分割されたコーパスを得ることができる。この処理を 1 回行って得られるコーパスにおいて、文字列としての出現頻度が低い単語 n-gram の頻度は、確率的単語分割コーパスから期待頻度を計算した場合と大きく異なる可能性がある。近似による誤差を減らすためには、上記の手続きを M 回行って得られる単

語分割コーパス全てを単語 n-gram 頻度の計数の対象とすればよい。このコーパスを疑似確率的単語分割コーパスと呼び、M をその倍率と呼ぶ。

3　未知語とその読み・文脈情報の自動獲得

本節では、仮名漢字変換の対象となる分野のテキストと音声を用いて未知語の読み・文脈情報を自動獲得し、統計的仮名漢字変換で用いられる言語モデルならびに仮名漢字モデルの性能を改善させる手法について述べる。

3.1　提案手法の概略

本項では提案手法の概略について述べる。図 2 に提案手法全体の概要を示す。本研究では、人手によって読みと単語境界が付与されている一般分野のコーパス C_b があらかじめ用意されているものとする。また、以下では一般分野のコーパスから読みを取り除いたコーパスを一般分野の単語分割コーパスと記述し、その中に存在する単語を既知語、それ以外の単語を未知語と定義する。

提案手法では、以下に示す 4 段階の処理により、未知語の読み・文脈情報を未知語を含む単語と読みの組の列として音声認識結果から獲得[3]し、統計的仮名漢字変換のモデルを更新する。

(1) 情報の付与されていない対象分野のテキストから疑似確率的単語分割コーパスを作成し、未知語の候補となる単語（以下、未知語候補と記述する）の抽出を行う（3.2 項を参照）。
(2) 疑似確率的単語分割コーパスを用いて音声認識のための言語モデルを構築する。また、未知語候補の読みを複数推定し、音声認識のための発音辞書を作成する（3.3 項参照）。
(3) 準備した言語モデル、発音辞書、音響モデルを用いて対象分野の音声を認識し、音声認識結果から単語と読みの組の列を獲得する（3.4 項を参照）。
(4) 獲得した単語と読みの組の列を統計的仮名漢字変換の学習コーパスに追加して言語モデルと仮名漢字モデルを更新する（3.5 項を参照）。

以下では、これらの処理について詳細を述べる。

3.2　疑似確率的単語分割コーパスを用いた未知語候補の抽出

まず、獲得対象となる未知語候補を単語境界の付与されていない対象分野のテキストから抽出する。本項では、2.3 項で述べた疑似確率的単語分割コーパスを用いた未知語候補の抽出について述べる。

疑似確率的単語分割コーパスは決定的に単語分割されたコーパスの集合であるが、全く同様の文であっても単語境界に揺れが存在するため、未知語の分割誤りを抑制可能である。しかしながら、テキスト中に出現する全ての部分文字列が単語になり得るという疑似確率的単語分割コーパスの性

[3]音声認識には大語彙音声認識システム Julius (Lee, Kawahara, and Shikano 2001) を用いる。

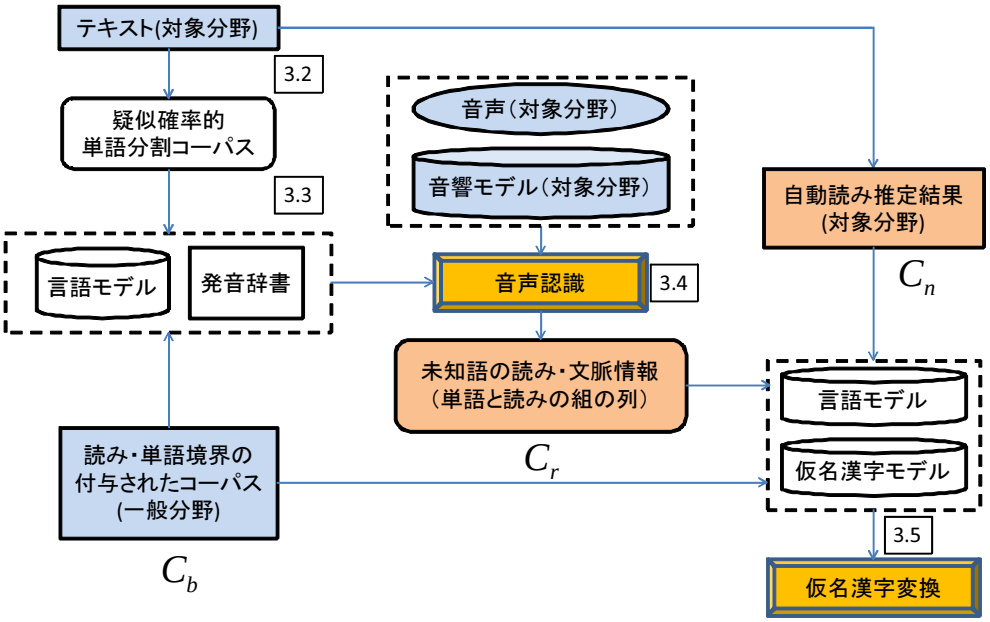


図 2　提案手法の概要図

質上、低頻度の文字列は単語として適切ではないものが多い。このため、出現頻度閾値を設定して適切な未知語候補を抽出する。

以下では、未知語候補「守屋」を抽出する場合を例にとり、その手続きを示す。

(1) 一般分野の単語分割コーパスから単語境界確率を推定するためのモデル（2.3 項を参照）を構築し、対象分野のテキストに単語境界確率を付与する。

···	昨	日	、	守	屋	前	次	官	が	···
··· 0.8		0.1	0.9	0.9	0.4	0.7	0.8	0.3	0.8	0.8 ···

(2) 単語境界確率と乱数の比較を行い、倍率 M の疑似確率的単語分割コーパスを作成する。

（試行 1）	··· 昨日　、　守屋　前　次官　が ···
（試行 2）	··· 昨日　、　守屋　前次官　が ···
（試行 3）	··· 昨日　、　守　屋前　次官　が ···
⋮	⋮
（試行 M）	··· 昨日　、　守屋　前　次　官が ···

(3) 作成した疑似確率的単語分割コーパス内に出現する単語のうち、頻度 F_{th} 以上の未知語（一般分野のコーパスに出現しない単語）を未知語候補として抽出する。

次項では、未知語候補の音声認識を行うための言語モデルと発音辞書について述べる。

3.3 未知語候補を含む言語モデルと発音辞書の作成

音声認識システムを用いて未知語候補を正しい読みとともに認識するためには、未知語候補が語彙に含まれる言語モデルと発音辞書が必要である。本項では、未知語候補を考慮した言語モデルならびに発音辞書の作成方法について述べる。

まず、音声認識のための言語モデルを構築する。大語彙連続音声認識システムを用いる場合には、対象分野のコーパスと一般分野のコーパスを用いて対象分野に適合した言語モデルの構築を行うことが一般的である (Matsunaga, Yamada, and Shikano 1992) (伊藤，好田 2000)。本研究では、3.2 項で作成した疑似確率的単語分割コーパスを一般分野の単語分割コーパスに追加し、言語モデルを構築する。

次に、未知語候補の読みを複数推定し、既知語から作成された発音辞書に追加する。読みの推定は、2.1 項の n-gram モデルにおける単語 w を文字とその読みの組に置き換えた n-gram モデルによって行う。以下では、未知語候補「守屋」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。

守:　マモ, シュ, モリ
屋:　ヤ, オク

(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。

マモヤ, マモオク, シュヤ, シュオク, モリヤ, モリオク

(3) 文字と読みの組を単位とする n-gram モデルにより、単語表記からの読みの生成確率を計算する。

$$\begin{aligned} P(\text{マモヤ} \mid \text{守屋}) &= 0.53 \\ P(\text{モリヤ} \mid \text{守屋}) &= 0.22 \\ P(\text{シ ュ オク} \mid \text{守屋}) &= 0.05 \\ &\vdots \end{aligned}$$

(4) 読みが付与されている一般分野のコーパスから発音辞書を作成し、(3) で推定した未知語候補と読みの組の中から、確率の上位 L 個を追加する。この際、L 個の未知語候補と読みの組の生成確率を反映させるため、単語の読みごとの確率を発音辞書に記述する[4]。

[4] 上位 L 個の確率の合計が 1 となるように正規化を行う。

既知語			未知語候補		
⋮			⋮		
国会	1.00	コッカイ	守屋	0.53	マモヤ
前	0.50	ゼン	守屋	0.22	モリヤ
前	0.50	マエ	守屋	0.05	シュオク
⋮			⋮		

上記の例における「守屋」の正しい読みは「モリヤ」であるが、(3) で述べた n-gram モデルによって与えられる確率 P(モリヤ | 守屋) は最大とならないため、確率の比較による正しい読みの選択は難しい。次項では、本項で作成した言語モデルと発音辞書を用いた音声認識によって未知語候補の正しい読みを選択する方法について述べる。

3.4 未知語の読み・文脈情報の獲得

前項の処理で発音辞書中に列挙される未知語候補の読みの中に正しい読みが含まれている場合には、音声認識によって未知語候補を含む単語と読みの組の列が得られる。しかし、前項の処理で推定した読みの多くは誤った読みであるため、音声認識の際に似た発音の単語を取り違え、誤った読みの未知語候補を出力する可能性がある。この問題に対処するため、ここでは言語モデルならびに音響モデルの尤度を反映した事後確率から計算される信頼度 (Wessel, Schlüter, and Ney 2001) [5] を用いて、認識結果における単語の文脈上の妥当性を判定する。ある単語の信頼度 CM は 0 から 1 の間の値で与えられ、大きい値であるほど信頼性が高いとみなされる。

以下では、音声認識を用いて未知語の読み・文脈情報を単語とその読みの列として獲得する手順を示す。

(1) 対象分野のテキストと同様の話題を扱った音声と、その音声に適合した音声認識用の音響モデルを用意する。

(2) (1) の音響モデルと、3.3 項の処理によって得られた言語モデルならびに発音辞書を用いて (1) の音声に対し音声認識を行い、単語、読み、単語信頼度の 3 つ組の列を出力する。

```
···  守屋/モリヤ/0.7  前/ゼン/0.8  事務/ジム/0.8  次官/ジカン/0.9  ···
··· 全体/ゼンタイ/0.4 の/ノ/0.7 守屋/シュヤ/0.05 が/ガ/0.8 狭/セマ/0.9 い/イ/0.9 ···
```

(3) 音声認識結果のうち、単語信頼度が CM_{th} 以上の単語を抽出し、連続する単語とその読みの組の列を作成する。なお、単語信頼度が CM_{th} より小さい単語は抽出せず、それまでに抽出

[5] ある単語を含む全ての単語列候補（音声認識結果）の相対的な尤度の比率を、その単語の信頼度として表す。なお、本研究で用いる音声認識システム Julius に実装されている単語信頼度は、信頼度計算の対象となる単語を含む最尤パスの確率で全体の確率の和を近似することによって計算される (李, 河原, 鹿野 2003)。

された単語とその読みの列を独立した文とみなす。

···　守屋/モリヤ　前/ゼン　事務/ジム　次官/ジカン ···
···　全体/ゼンタイ　の/ノ
が/ガ　狭/セマ　い/イ　···

3.5　統計的仮名漢字変換のためのモデル構築

仮名漢字変換のモデル性能を改善するには、対象分野の学習コーパスを大量に用意することが重要である。人手によって十分な量のコーパスを作成することはコストの面で実用的ではないため、まずテキストの読み推定を行うことによって対象分野のテキストに単語境界と読みを自動的に付与する。ここでは、2.1 項の式 (1)(2) において単語 w を単語と読みの組に置き換え、読み推定のための n-gram モデルを一般分野のコーパス C_b から構築する。この結果得られるコーパスを C_n とする。一般的には、情報の付与されていない対象分野のテキストのみを大量に入手可能である、という状況が多いため、上述の読み推定システムや形態素解析器の利用によって大規模なコーパス C_n を作成し、C_b と C_n からモデルを構築することによって変換精度を向上させることが可能である。しかしながら C_n は一般分野のコーパス C_b から構築されるモデルを用いたシステムによって単語境界や読みを付与されるため、C_b の内部に出現しない未知語の情報をモデルに反映させることは難しい。この問題を解決するため、提案手法では 3.4 項の処理によって獲得される、未知語を含む単語と読みの列をコーパス C_r とみなし、C_r によって未知語の読み・文脈情報をモデルに反映させ、未知語の変換精度の向上を図る。

4　評価

本節では、3 節で述べた提案手法の評価実験について述べる。まず、3.2 項〜3.4 項で述べた手法に従って、未知語の読み・文脈情報を単語とその読みの組の列として獲得した。その後、3.5 項で示した学習コーパスから統計的仮名漢字変換の言語モデルならびに仮名漢字モデルを構築して精度評価を行い、提案手法の有効性を検証した。

4.1　実験で利用するテキストと音声

本項では、実験を行う際にあらかじめ準備するデータ、ならびに実験の過程で利用するデータについて述べる。

テキスト　本実験において利用するテキストコーパスを以下に示す。

一般分野のコーパス C_b には現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary

表 1　コーパスの一覧

	コーパス	文数	単語数	文字数
C_b	一般分野	14,645	485,604	693,156
C_n	対象分野 (自動読み推定結果)	30,552	$1,856,237^*$	2,528,722
C_r	音声認識結果	-	$109,313^*$	150,646
C_t	テストセット	888	50,018	73,020

仮名漢字変換の精度評価では、C_b,C_n,C_r をモデル学習に、C_t をテストに用いた。

* 自動読み推定ならびに音声認識システムの出力から単語数を計数した。

表 2　テストセット C_t の未知 n-gram 率 (%)

n-gram の単位	1-gram	2-gram
単語	5.54 (= 2,772 / 50,018)	29.40 (= 14,705 / 50,018)
単語と読みの組	5.85 (= 2,925 / 50,018)	30.61 (= 15,313 / 50,018)

Written Japanese; BCCWJ) (小椋，小磯，冨士池，原 2008) を用いた。BCCWJ はあらかじめ単語分割がされており、各単語に読みが付与されている[6]。ここで、BCCWJ の内部に出現する全ての単語が既知語となる。

対象分野のテキストとして、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を自動収集したものを用いた。このウェブニューステキストには情報が付与されていないため、このテキストに対して 3.5 項で示した手法を適用することで、単語分割と読みの付与を自動的に行い、コーパス C_n を作成した。また、ウェブニュースのテキストは 3.2 項で述べた疑似確率的単語分割コーパスの作成に用いた。

後述する 4.2 項の実験により、音声認識結果 C_r として単語と読みの組の列が獲得される。C_r は、C_n と同様に仮名漢字変換のためのモデル構築に用いた。

テストセット C_t として、2008 年 1 月 9 日、2008 年 1 月 10 日の 2 日間のウェブニュースを単語分割し、読みを付与したものを用いた。

以上に述べたテキストコーパスの文数、単語数、文字数を表 1 に示す。なお、表 1 において、対象分野のテキストに対する自動読み推定結果、ならびに音声認識結果の単語数は、各システムの出力結果から単語数を計数したものである。また、音声認識結果の出力から文境界を同定することは

[6] 本研究では、人手による修正が入ったコアデータのみを使用し、さらに活用語を語幹と語尾に分割する等の変更を加えている。

表 3　対象分野のテキストから抽出した未知語候補の再現率 (%)

日数 \ M	1	2	5	10	1(決定的)
7	6.75	11.69	22.37	30.74	7.83
14	10.61	16.67	28.68	40.51	9.38
35	18.80	28.43	45.96	58.48	19.12
68	28.93	39.57	57.00	**66.56**	22.85

M は疑似確率的単語分割コーパスの倍率を示す。

困難であるため、単語数と文字数のみを示す。

表 2 に、テストセットにおける未知の 1-gram 率（未知語率）、未知の 2-gram 率を、単語を単位とする場合と単語と読みの組を単位とする場合のそれぞれについて示す。

音声　読みを選択するために用いる音声として、収集したウェブニュース記事と同時期に当たる 2007 年 12 月 5 日から 2008 年 1 月 8 日の間に放送された 30 分のニュース番組の合計 17 時間の音声を用いた。ここで、対象分野のテキストと音声の類似度として、音声の一部の書き起こし（2008 年 1 月 7 日、8 日の 2 日分）に対するパープレキシティを示す。後述する対象分野の疑似確率的単語分割コーパスから単語 3-gram モデルを構築し、書き起こしに対するパープレキシティを求めたところ、58.5 となった。これは、本実験で用いる疑似確率的単語分割コーパスから構築される音声認識用言語モデルは認識対象となる音声に対して十分な単語予測性能を持っている（対象分野の音声と対象分野のテキストが十分に似ている）ことを示している。

4.2　未知語とその読み・文脈情報の自動獲得

本項では、対象分野のテキストと対象分野の音声を用いた未知語とその読み・文脈情報の自動獲得について述べる。また、処理の途中段階で獲得した未知語とテストセット中の未知語を比較し、各処理における未知語の検出精度を示す。

未知語候補の抽出　まず、3.2 項で述べた手法に従って対象分野のテキストから疑似確率的単語分割コーパスを作成し、未知語候補の抽出を行った。ここで、疑似確率的単語分割コーパスの倍率は $M = 10$ とした。また、未知語候補を決定する際の閾値は、$F_{th} = 50$ とした。

また、対象分野のテキストの規模と最終的に獲得可能な未知語の数との関係として、表 3 に、未知語候補のテストセット C_t 中の未知語に対する再現率を示す。表 3 では、利用するウェブニュースの日数と疑似確率的単語分割コーパスの倍率 M を変えることでテキストの規模を調節し、それぞれについて再現率を示した。また、確率的単語分割コーパスを作成せず、決定的に単語分割を行った場合の再現率についても、併せて表 3 に示した。C_t 内の未知語の集合を UW_t、疑似確率的単語分

表 4　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
合計	20,712	26,880

割コーパス内の未知語候補の集合を UW_c とし、コーパス C における単語 w の出現頻度を $f(C,w)$ とすると、再現率は

$$\frac{\sum_{w \in UW_t \cap UW_c} f(C_t, w)}{\sum_{w \in UW_t} f(C_t, w)}$$

で表される。ここで、$\sum_{w \in UW_t} f(C_t, w) = 2,772$ である（表 2 参照）。

表 3 から、未知語の抽出を行う場合には、決定的な単語分割を行ったコーパスではなく、疑似確率的単語分割コーパスを利用することが有効であることがわかる。

未知語候補を含む言語モデルと発音辞書の作成　3.3 項で述べた手法を用いて音声認識用の言語モデルと発音辞書を作成した。本実験で用いる音声認識システム Julius は言語モデルとして順向き 2-gram モデル、逆向き 3-gram モデルを必要とする。ここでは、一般分野の単語分割コーパス (BCCWJ) と対象分野の疑似確率的単語分割コーパス（ウェブニュース）から単語表記を単位とする順向き 2-gram モデルならびに逆向き 3-gram モデルを構築した。

次に、抽出した未知語候補の読みを、文字と読みの組を単位とする 2-gram モデルによって推定し、生成確率の上位 L 個の単語と読みの組を既知語から作成される発音辞書に追加した。本実験では $L = 5$ とした。

作成した発音辞書の詳細を表 4 に示す[7]。言語モデルにおける語彙の総数は表 4 における既知語と未知語候補の単語数を合計した数である 20,712、発音辞書のエントリ（単語と読みの組）の総数は 26,880 となった。

ここで、L の値の妥当性を検証するため、L を変えた場合に得られる未知語候補と読みの組の、テストセット C_t 内の未知語と読みの組に対する再現率を表 5 に示す。コーパス C における単語と読みの組 u の出現頻度を $f(C,u)$、未知語候補と推定された読みの組の集合を UU_e、テストセット

[7]片仮名のように文字ごとの読み候補が少ない場合、もしくは単語長が短い場合など、5 個まで読みの列挙を行うことができない未知語候補が存在する。このため、未知語候補のエントリ数は単語数の 5 倍未満になることがあり得る。

表 5　発音辞書に列挙された未知語候補と読みの組の再現率 (%)

L	1	2	5	10	20	50
再現率	49.9	51.2	52.6	53.1	53.3	53.3

C_t 内の未知語と読みの組の集合を UU_t とすると、再現率は

$$\frac{\sum_{u \in UU_t \cap UU_e} f(C_t, u)}{\sum_{u \in UU_t} f(C_t, u)}$$

で表される。ここで、$\sum_{u \in UU_t} f(C_t, u) = 2,925$ である（表 2 参照）。表 5 より、$L \geq 5$ では再現率に大きな変化が見られないことから、L の値を単純に大きくしても最終的に獲得可能な未知語と読みの組の量は変わらないことが予想される。また、L を大きくするに従って、誤った読みを持つエントリがより多く発音辞書に登録され、認識誤りが増加する。本実験では以上の 2 点を考慮し、$L = 5$ とした。

未知語の読み・文脈情報の獲得　作成した言語モデルと発音辞書を利用し、音声認識によって読みを選択し、音声認識結果から未知語を含む単語と読みの組の列を獲得した。音声認識システムには、Julius 3.5.3 を用いた。なお、Julius の動作に必要となる音響モデルは、連続音声認識コンソーシアム 2003 年度版ソフトウェア[8] に同梱されている、新聞記事読み上げ音声コーパス（JNAS）から学習された 3,000 状態、64 混合の PTM triphone モデル (李，河原，武田，鹿野 2000) を用いた。

音声認識結果のうち、単語信頼度が CM_{th} を超えている単語のみを抽出し、単語と読みの組の列を単語境界と読みの付与されたコーパス (C_r) の形で獲得した。この際、単語信頼度の閾値は $CM_{th} = 0.1$ とした。また、獲得頻度の少ない未知語候補には音声認識誤りと考えられるものが多かったため、上記の閾値による制限に加えて 2 回以上認識した未知語候補のみを獲得した[9]。

C_r の単語数ならびに文字数は表 1 で示した通りである。また、表 6 に音声認識結果 C_r の未知語率を示す。ここでは、テストセット C_t の未知語率（表 2 参照）と同様に、単語ならびに単語と読みの組を単位とした場合の未知 n-gram 率を示す。なお、最終的に獲得された未知語候補と読みの組（異なり数）は 872 となった。

最後に、対象分野の音声の規模と獲得した未知語と読みの組の数との関係を調べるため、使用するニュースの日数を変更した場合の C_t に対する再現率を表 7 に示す。C_r 内の未知語と読みの組の

[8] http://www.lang.astem.or.jp/CSRC/

[9] アナウンサーの発話のように、音声が明瞭である部分の認識精度は 80%程度、記者発表のように、背景に雑音が多く含まれる部分の認識精度は 30%程度であった。

表 6　音声認識結果 C_r の未知 n-gram 率 (%)

n-gram の単位	1-gram	2-gram
単語	3.19	24.95
	(= 3,486 / 109,313)	(= 26,182 / 109,313)
単語と読みの組	3.42	28.17
	(= 3,739 / 109,313)	(= 30,796 / 109,313)

表 7　音声認識結果内の未知語候補と読みの組の再現率 (%)

用いたニュースの日数	1	7	14	35
再現率	9.15	20.0	25.46	31.6

集合を UU_r とすると、再現率は

$$\frac{\sum_{u \in UU_t \cap UU_r} f(C_t, u)}{\sum_{u \in UU_t} f(C_t, u)}$$

で表される。

獲得した未知語と読み・文脈情報の再現率と適合率　本実験の目的は、後述する仮名漢字変換の精度評価において、音声認識結果 C_r から獲得した未知語とその読み・文脈情報を利用してテストセット C_t を対象とした仮名漢字変換の変換精度を向上させることにある。C_r を用いて仮名漢字変換のモデルを構築する場合、C_t と C_r に共通して出現する未知語と読みの組、または単語を単位とする未知の 2-gram が多いほど仮名漢字変換の精度が向上する[10]。以下では、それぞれの再現率ならびに適合率を示す。

まず、C_r から獲得した未知語と読みの組の再現率、適合率を示す。再現率ならびに適合率はそれぞれ

$$\text{再現率} = \frac{\sum_{u \in UU_t \cap UU_r} f(C_t, u)}{\sum_{u \in UU_t} f(C_t, u)} \quad , \qquad \text{適合率} = \frac{\sum_{u \in UU_t \cap UU_r} f(C_r, u)}{\sum_{u \in UU_t} f(C_r, u)}$$

で表される。計算の結果、再現率は 31.6%、適合率は 38.2%となった。

次に、未知の 2-gram の再現率、適合率を示す。コーパス C における単語 2-gram（w_{i-1}^{i}）の出

[10]前者は 2.2 項で述べた仮名漢字モデルの性能に影響し、後者は言語モデルの性能に影響する。

現頻度を $f(C, w_{i-1}^i)$、テストセット C_t 内の未知の単語 2-gram の集合を UB_t、音声認識結果 C_r 内の未知の単語 2-gram の集合を UB_r とすると、再現率ならびに適合率は

$$\text{再現率} = \frac{\sum_{w_{i-1}^i \in UB_t \cap UB_r} f(C_t, w_{i-1}^i)}{\sum_{w_{i-1}^i \in UB_t} f(C_t, w_{i-1}^i)} \quad , \qquad \text{適合率} = \frac{\sum_{w_{i-1}^i \in UB_t \cap UB_r} f(C_r, w_{i-1}^i)}{\sum_{w_{i-1}^i \in UB_t} f(C_r, w_{i-1}^i)}$$

で表される。計算の結果、再現率は 31.9%、適合率は 25.5%となった。

4.3 統計的仮名漢字変換による精度評価

本項では、3.5 項で挙げた学習コーパスを用いて統計的仮名漢字変換の精度評価を行い、提案手法の有効性を検証する。

実験の条件　本実験では、一般分野のコーパス C_b、対象分野のテキストの自動読み推定結果 C_n、音声認識結果 C_r を用いて統計的仮名漢字変換のためのモデルを構築した。各コーパスの規模は 4.1 の表 1 に示した通りである。

本実験では、3 種類のコーパスを以下のように組み合わせて学習コーパスとし、言語モデル（単語 2-gram モデル）ならびに仮名漢字モデルを構築した。

(1) C_b: ベースライン
(2) $C_b + C_n$: テキストのみを用いた手法（既存手法）
(3) $C_b + C_n + C_r$: テキストと音声に共通して現れる未知語の読み・単語文脈を反映させる手法（提案手法）

統計的仮名漢字変換システム全体の精度を評価する基準として、文字単位の再現率と適合率を計算し、(1)–(3) について比較を行った。また、提案手法において未知語の読みと単語文脈を共に利用することの有効性を検証するため、(2) を基準として、C_r から言語モデル (LM) のみを更新した場合[11]と、仮名漢字モデル (PM) のみを更新した場合[12]についても変換精度の評価を行った。

本実験における評価指標として、文字単位の再現率と適合率を用いる。それぞれの定義を以下に示す。

$$\text{再現率} = \frac{\text{正解文字数}}{\text{テストセット中の文字数}}$$

$$\text{適合率} = \frac{\text{正解文字数}}{\text{システムの出力した文字数}}$$

[11] $C_b + C_n + C_r$ から言語モデルを構築し、$C_b + C_n$ から仮名漢字モデルを構築する。
[12] $C_b + C_n$ から言語モデルを構築し、$C_b + C_n + C_r$ から仮名漢字モデルを構築する。

表 8　統計的仮名漢字変換による評価 (%)

学習コーパス	再現率	適合率
C_b	87.96	84.60
（ベースライン）	(= 64,231/73,020)	(= 64,231/75,919)
$C_b + C_n$	96.90	96.28
（既存手法）	(= 70,760/73,020)	(= 70,760/73,496)
$C_b + C_n + C_r$	**97.26**	**96.76**
（提案手法）	(= 71,019/73,020)	(= 71,019/73,396)
$C_b + C_n$	96.93	96.31
$+C_r$（LM のみ）	(= 70,780/73,020)	(= 70,780/73,493)
$C_b + C_n$	97.07	96.55
$+C_r$（PM のみ）	(= 70,884/73,020)	(= 70,884/73,416)

実験結果と考察　(1)–(3) で示した学習コーパスから構築されるモデルによる再現率、適合率を表 8 に示す。

C_b を用いる場合（ベースライン）の変換精度と C_b, C_n を用いる場合（既存手法）の変換精度を比較した結果、再現率で 8.94%、適合率で 11.68%の精度向上が確認された。ここで C_n と C_t は同分野のコーパスであり、C_b は C_n に比較すると小規模なコーパスであるため、この精度向上は単純に学習データの量を増やしたことに起因すると考えられる。

次に、C_b, C_n を用いる場合（既存手法）の変換精度と C_b, C_n, C_r を用いる場合（提案手法）の変換精度を比較した結果、仮名漢字変換の精度は再現率で 0.36%、適合率で 0.48%の改善が見られた。既存手法において、コーパス C_n は対象分野の未知語を考慮しない手法で読みを付与されているため、未知語の正しい分割と読みの付与が行われず、C_b と C_n のみを用いて構築されるモデルでは未知語の誤変換が発生する。しかし、提案手法では 4.2 項の実験で得られた C_rを用いて 未知語の読み・文脈情報をモデルに反映させることが可能である。上記の精度増加は、4.2 項で示した未知語の読み・文脈情報の獲得の実験で獲得した未知語と読みの組、未知の 2-gram の量に対応しており、より多くの未知語を獲得するほど変換精度が向上すると考えられる。

また、C_r を追加することによる精度向上の要因を明らかにするため、C_b と C_n から構築したモデルによる変換精度を基準に、C_r を利用して言語モデルと仮名漢字モデルを独立に更新して精度を比較した。言語モデルのみを更新した場合は、再現率、適合率ともに 0.03%の改善となり、仮名漢字モデルのみを更新した場合は、再現率で 0.17%、適合率で 0.27%の改善となった。

言語モデルのみを更新する場合、未知語と読み（仮名漢字変換における入力記号列）との対応付けを行うことが不可能であるため、未知語周辺の文脈が変換精度の向上にほとんど寄与しない。こ

の際、変換精度の向上に寄与する要素は C_r に現れる既知語周辺の文脈情報のみであり、かつ C_n に比較して C_r の規模は非常に小さいために、精度がほぼ変化していないと考えられる。

仮名漢字モデルのみを更新する場合については、一定の精度向上が観察された。しかしながら、ある読みを持つ未知語に対し、同じ読みを持つ既知語、もしくは結合の結果同じ読みとなる既知語の連続が存在するという状況では、未知語を含む変換候補の言語モデル確率は既知語を含む変換候補の確率に比較して小さくなる。言語モデルと仮名漢字モデルの両方を更新する場合（提案手法）との精度の差は、上述の言語モデル確率の差に起因する。

最後に、提案手法を用いることで未知語の変換誤りが改善した例を示す。

$C_b + C_n$:	前 事務 次官 の 森 や タケマサ
$C_b + C_n + C_r$:	前 事務 次官 の 守屋 武昌

3.2〜3.4 項において例として示した未知語（守屋）は、本実験において実際に獲得された未知語の例であり、音声認識結果 C_r を用いることによって未知語の誤変換が改善されることを確認した。

以上の結果より、テキストと音声から獲得される未知語の読み・文脈情報は統計的仮名漢字変換システムの精度向上に有効であることが確認された。

5 関連研究

第 1 節で述べた通り、人手によって任意の分野における未知語の情報を収集することはコストの面で現実的ではない。このため、未知語に関する情報を自動獲得する研究が多く行われている。

まず、形態素解析など、自動単語分割を行うシステムにおいて単語辞書に未知語を追加することを目的とした研究について述べる。

文献 (永田 1999b) では、ある文の自動単語分割候補における N-best の相対確率を、それぞれの候補において出現する未知語の出現頻度の期待値として与える。その後、出現した未知語の中から一定の閾値より大きい出現頻度の期待値を持つ未知語を獲得している。また、単語分割の際には、未知語を構成する字種によって 9 種類の未知語タイプを定義し、それぞれのタイプにおける単語長の分布を考慮した未知語モデルを用いることで、未知語モデルの性能向上を図っている。

形態素解析のため、品詞を考慮して未知語を獲得する研究として、文献 (Mori and Nagao 1996) では、コーパス中に出現する任意の部分文字列 α に注目し、α の前後の文字から、α が未知語として出現する可能性の高い品詞に属する確率を推定している。その後、出現頻度が一定値以上かつ 2 文字以上の文字列 α を単語として抽出しておき、形態素解析器にかけた結果に辞書未登録語が含まれている文字列 α を未知語として獲得している。

日本語は分かち書きを行わない言語であるため、自動単語分割器や形態素解析器において必須となる未知語の情報は正しい単語単位である。このため、形態素解析器のための未知語獲得を行う研

究では未知語の読みには言及しないことが多い。しかしながら、本研究では統計的仮名漢字変換の精度向上を目的としているため、未知語の表記ならびにその読みに関する情報を同時に獲得することが望ましい。

文献 (森，小田 2007) では、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手法を提案している。また、言語モデルの更新を繰り返すことで、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。ただし、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、本論文で扱う自動獲得とは性質が異なる。

音声認識の分野においては、未知語を原因とする認識誤りの影響を抑制するため、単語より小さい単位の語彙であるサブワードを擬似的な単語とし、未知語をサブワードの連続として認識する手法が提案されている (甲斐，廣瀬，中川 1999)(Bisani and Ney 2005) (Choueiter, Seneff, and Glass 2007) 。しかしながら、日本語の音声認識においてサブワードは基本的に仮名文字列から構成されるため、サブワードをそのまま未知語獲得に用いても仮名漢字変換への寄与は低いと考えられる。

文献 (倉田，森，伊東，西村 2008) では、規則を用いてテキストから未知語の候補を抽出、音声認識を用いて読みを自動的に獲得し、発音辞書に追加する手法が提案されている。この手法は、テキストと音声から未知語と読みの情報を獲得する点で本研究と共通しているが、未知語候補の抽出方法と獲得する情報の粒度が本研究と異なる。本研究では、疑似確率的単語分割コーパスを用いることにより、一貫した単語単位で言語モデルと発音辞書を作成する。また、音声認識結果から未知語の読みだけではなく文脈情報を獲得し、統計的仮名漢字変換で利用する確率的言語モデル全体の性能向上を図っている。

6　結論

本論文では、類似した話題を扱っているテキストと音声から未知語の読み・文脈情報を単語と読みの組の列として自動獲得し、統計的仮名漢字変換の精度向上に利用する手法を提案した。

自動的に収集可能なニュース記事とニュース音声を用いた実験の結果、音声認識結果から得られる単語と読みの組の列を学習コーパスとして統計的仮名漢字変換のモデルを学習することにより、システム全体の精度が向上することを確認した。

以上の結果から、テキストと音声を用いることにより、仮名漢字変換システムの効率的かつ継続的な精度向上を行うことが可能であることが示された。

参考文献

Bisani, M. and Ney, H. (2005). “Open Vocabulary Speech Recognition with Flat Hybrid Models.” In *Proceedings of the Interspeech 2005*, pp. 725–728.

Choueiter, F., Seneff, S., and Glass, J. (2007). “New Word Acquisition Using Subword Modeling.” In *Proceedings of the Interspeech 2007*, pp. 1765–1768.

Lee, A., Kawahara, T., and Shikano, K. (2001). “Julius – an open source real-time large vocabulary recognition engine.” In *Proceedings of the Eurospeech 2001*, pp. 1691–1694.

Matsunaga, S., Yamada, T., and Shikano, K. (1992). “Task adaptation in stochastic language models for continuous speech recognition.” In *Proceedings of the ICASSP 1992*, Vol. 1, pp. 165–168.

Mori, S. and Nagao, M. (1996). “Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis.” In *Proceedings of the COLING 1996*, pp. 1119–1122.

Wessel, F., Schlüter, R., and Ney, H. (2001). “Confidence measures for large vocabulary continuous speech recognition.” *IEEE Transactions on Speech and Audio Processing*, **9** (3), pp. 288–298.

倉田岳人，森信介，伊東伸泰，西村雅史 (2008). “音声とテキストを用いた認識単語辞書の自動構築.” 情報処理学会論文誌, **49** (8), pp. 2900–2909.

北研二 (1999). 確率的言語モデル, 言語と計算, 4 巻. 東京大学出版会.

李晃伸，河原達也，武田一哉，鹿野清宏 (2000). “Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識.” 電子情報通信学会論文誌, **J83-D-II** (12), pp. 2517–2525.

李晃伸，河原達也，鹿野清宏 (2003). “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, 2003-SLP-49-48, pp. 281–286.

小椋秀樹，小磯花絵，冨士池優美，原裕 (2008). 『現代日本語書き言葉均衡コーパス』形態論情報規程集.

甲斐充彦，廣瀬良文，中川聖一 (1999). “単語 N-gram 言語モデルを用いた音声認識システムにおける未知語 · 冗長語の処理.” 情報処理学会論文誌, **40** (4), pp. 1383–1394.

伊藤彰則，好田正紀 (2000). “N-gram 出現回数の混合によるタスク適応の性能解析.” 電子情報通信学会論文誌, **J83-D-II** (11), pp. 2418–2427.

永田昌明 (1999a). “統計的言語モデルと N-best 探索を用いた日本語形態素解析法.” 情報処理学会論文誌, **40** (9), pp. 3420–3431.

永田昌明 (1999b). “未知語の確率モデルと単語の出現頻度の期待値に基づくテキストからの語彙獲得.” 情報処理学会論文誌, **40** (9), pp. 3373–3386.

森信介，土屋雅稔，山地治，長尾真 (1999). “確率的モデルによる仮名漢字変換.” 情報処理学会論

文誌, **40** (7), pp. 2946–2953.

森信介，小田裕樹 (2007). “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 回年次大会発表論文集, pp. 340–343.

森信介，小田裕樹 (2009). “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), pp. 7–21.

長野徹，森信介，西村雅史 (2006). “N-gram モデルを用いた音声合成のための読み及びアクセントの同時推定.” 情報処理学会論文誌, **47** (6), pp. 1793–1801.

略歴

笹田 鉄郎：2007 年京都大学工学部電気電子工学科卒業．2009 年同大学院情報学研究科修士課程修了．同年， 同大学院博士後期課程に入学，現在に至る．

森 信介：1993 年京都大学工学部電気工学第二学科卒業．1995 年京都大学大学院工学研究科電気工学第二専攻修士課程修了．1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム（株）入社．2007 年日本アイ・ビー・エム（株）退社．同年より京都大学学術情報メディアセンター准教授．現在に至る．自然言語処理ならびに音声言語処理，特に確率的言語モデルに関する研究に従事．京都大学工学博士．1997 年情報処理学会山下記念研究賞受賞．

河原 達也：1987 年 京都大学工学部情報工学科卒業．1989 年 同大学院修士課程修了．1990 年 同博士後期課程退学．同年 京都大学工学部助手．1995 年 同助教授．1998 年 同大学情報学研究科助教授．2003 年 同大学学術情報メディアセンター教授．現在に至る．この間，1995 年から 1996 年まで 米国・ベル研究所客員研究員．1998 年から 2006 年まで ＡＴＲ客員研究員．1999 年から 2004 年まで 国立国語研究所非常勤研究員．2001 年から 2005 年まで 科学技術振興事業団さきがけ研究 21 研究者．2006 年から 情報通信研究機構短時間研究員．音声言語処理，特に音声認識及び対話システムに関する研究に従事．京大博士（工学）．1997 年度 日本音響学会粟屋潔学術奨励賞受賞．2000 年度 情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表，IEEE SPS Speech TC 委員，IEEE ASRU 2007 General Chair，言語処理学会理事，を歴任．情報処理学会音声言語情報処理研究会主査．日本音響学会，情報処理学会 各代議員．電子情報通信学会，人工知能学会，言語処理学会，IEEE 各会員.