社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－SLP－72 (1)
2008／7／18

テキストと音声を用いた単語と読みの自動獲得

笹田 鉄郎† 森 信介†‡ 河原 達也†‡
†京都大学 情報学研究科
‡京都大学 学術情報メディアセンター
〒606-8501 京都市左京区吉田本町

あらまし
音声合成や音声認識における課題の一つとして未知語の問題があるが、テキストから得られる情報だけで未知語の読みを推定することは難しい。本論文では、テキストと同様の話題を扱った音声を用いた音声認識によって未知語とその読みを自動獲得する手法を提案する。未知語を含む音声認識結果を言語モデル適応に利用し、テキストの読み推定による評価実験を行った結果、読み推定精度が向上することを確認した。

キーワード 未知語 読み 音声合成

Extracting Words and Pronunciations from a Set of Text and Speech

Tetsuro SASADA†, Shinsuke MORI†‡, Tatsuya KAWAHARA†‡
†Graduate School of Informatics, Kyoto University
‡Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

Abstract
One of the problems in text-to-speech (TTS) systems and automatic speech recognition (ASR) systems is pronunciation estimation of unknown words. Usually a TTS system is equipped with a module which estimates a pronunciation of an unknown word from its spelling. The accuracy of this module is, however, not sufficiently high. In this paper, we propose a method for extracting unknown words and their pronunciations from a comparable set of text data and speech data. We tested the TTS front-end enhanced with the ASR outputs on other web news articles, and observed an improvement in the pronunciation estimation.

Key Words Unknown Words, Pronunciations, Text-to-Speech

1 はじめに

近年では、音声言語処理技術が発展してきたことにより、実用を見据えたアプリケーションについての研究が多く進んでいる。その中の一つに、音声合成を用いたテキスト読み上げ（TTS, Text-To-Speech）がある。TTSはテキストから音声を生成、出力するシステムで、まず言語処理部分でテキストの読みとアクセントの推定を行い、それを音声波形に変換する。

TTSによってテキストの内容を伝えるためには、テキストの読み推定を正しく行うことが重要である[1]。読み推定の精度向上における最も大きな問題の一つとしては未知語の存在があげられる。一般にTTSでは未知語の表記からその読みを推定するが、日本語を構成する文字の多くは読みに曖昧性があるため、テキストのみから読み推定を行うことは難しい。これは音声合成に限らず、音声認識や仮名漢字変換など、単語表記と読みを辞書中に必要とするアプリケーションにおいて共通の問題である。

本論文では、テキストと音声から未知語の表記と読みを自動獲得する手法を提案する。これによって単語とその読み、さらにその周辺の文脈情報を獲得することができ、各種のアプリケーションの精度向上に役立つことが期待される。以下に本論文で用いる手法の基本的な手順を示す。

1. テキストから未知語候補となる単語を抽出する。
2. それぞれの未知語候補に対し、複数の読みを推定して列挙する。
3. 推定した読みを音声と比較し、正しいものを選択する。

手順1の時点では読みの情報が不要であるため、問題はテキストの単語分割のみに絞られる。未知語の抽出を目的とした研究、もしくは未知語に対して頑健な解析システムの先行研究としては[2][3]などがある。本論文では確率的単語分割コーパス[4]の考え方を用いて単語分割ならびに未知語候補の抽出を行う。手順2は文字と読みの組を単位としたn-gramモデルを用いることで実現できる。手順3を実現するための手段としては音声認識システムを用いるが、類似した読みを持つ別の単語が認識誤りを起こすことのないようにする必要がある。本論文では未知語候補の抽出元となったテキストデータを音声認識システムの言語モデル推定にも用い、かつテキストと同じ話題を扱った音声を用意することで、この問題に対処する。

同時期のウェブニューステキストとニュース音声を用いた実験の結果、文字列のみを用いた推定では正しい読みが獲得できなかったものを含め、読みの与えられた学習コーパスには現れなかった未知語とその読みを適切に取得できた。また、提案手法の評価実験としてテキスト読み推定を行った。このとき、未知語候補と読みの獲得実験で得られた音声認識結果のうち適切な部分を選択して言語モデル適応を行うことにより、未知語の周辺を含む全体の推定精度が向上することを確認した。

2 n-gramモデルに基づく読み推定

TTSを用いるためにはまず何らかの方法で文の読みを推定する必要がある。日本語のように分かち書きされていない言語を解析するための確率的モデルとして、永田[5]により形態素解析のためのn-gramモデルが提案されている。また、長野ら[6]はこのモデルを拡張し、読みとアクセントの推定を行っている。本節では、文ならびに文中の未知語の読み推定を行うためのn-gramモデルについて述べる。

2.1 n-gramモデルによる文の読み推定

日本語における文の読み推定は、単語分割と読みのタグ付けを同時に行う問題とみなすことができる。このとき、文は単語の表記wと読み$\boldsymbol{y}$の組の列となる。この組$u = \langle w, \boldsymbol{y} \rangle$を単位とする$n$-gramモデルを用いることでテキストの読み推定を行うことができる[2]。uを単位とするn-gramモデル$M_{u,n}$は以下の式によって表される。

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

ここで$u_i\,(i \leq 0)$とu_{h+1}はそれぞれ文頭と文末を表す特別な記号である。

読み推定の結果は文字列$\boldsymbol{x}$を入力とした場合、以下の式(1)で示されるように、生成確率が最大となる単語表記と読みの組の列$\hat{\boldsymbol{u}}$で与えられる。

$$\hat{\boldsymbol{u}} = \underset{\boldsymbol{x}=w_1 w_2 \cdots w_h}{\mathbf{argmax}} \; M_{u,n}(u_1 u_2 \cdots u_h) \qquad (1)$$

式(1)において、w_iはu_iの単語表記である。

2.2 未知語の読み推定

学習コーパスに現れない単語が含まれるn-gram確率は計算できないため、上で述べたモデルにおいて未知語とその読みを予測するために、未知の単語と読みの組を

[1] 合成音声の評価指標として、音節明瞭度や単語了解度[1]が用いられている

[2] 文献[6]においては、uを単語表記、読み、品詞、アクセントの4つ組としている。

表す特別な記号 UU を導入する。本論文では学習コーパス中に 1 回のみ出現する u の集合を $\mathcal{U}$ と定義する。未知語を予測する際はまず $M_{u,n}$ によって UU を予測し、文字とその読みのペア $v = \langle x, \boldsymbol{y} \rangle$ を単位とした n-gram モデル $M_{v,n}$ を用いることで未知語とその読みを予測する。これは以下の式で表される。

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

ここで $v_i\,(i \leq 0)$ と $v_{h'+1}$ は単語の始点と終点を表す特別な記号である。

未知語の読み推定結果は式 (1) と同様に、生成確率が最大となる文字表記と読みの組の列によって与えられる。

3 確率的単語分割コーパスからの言語モデル推定

本論文で提案する手法を用いて未知語とその読みを獲得するためには、何らかの手段でテキスト中に存在する未知語候補の検出を行う必要がある。一般的な形態素解析システムによってテキストを解析した場合、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可能性が高い。この問題に対処するため、本論文では森ら [4] により提案されている確率的単語分割コーパスを用いる。本節では、まず確率的単語分割コーパスを用いた n-gram 言語モデルの推定方法について述べる。次に、これを決定的に単語分割されたコーパスで近似する疑似確率的単語分割コーパスについて説明する。

3.1 確率的単語分割コーパス

日本語の単語分割は、入力文（文字列）の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。確率的単語分割コーパスは、それを構成する長さ n_r の文字列 $\boldsymbol{x}_1^{n_r}$ において、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。本論文では、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 [7]。最大エントロピーモデルの素性としては、x_{i-2}^{i+2} の範囲の文字 n-gram（$n = 1, 2, 3$）に文字種の情報を付加したものを用いる。

決定的に自動単語分割を行うシステムを用いると、未知語の多くは過分割もしくは過併合を起こすこととなる。しかし確率的単語分割コーパスを用いると、コーパス中に存在する全ての部分文字列を単語として扱うことができる。したがってそれぞれの部分文字列に対して、

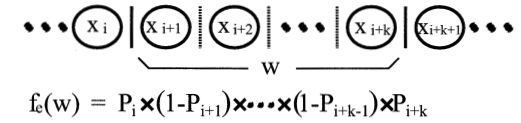

図 1: 確率的単語分割コーパスにおける期待頻度

後述する期待頻度を計算し、それを手がかりに未知語を検出することが可能である。

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 $f_r(\cdot)$ 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。

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

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 $\boldsymbol{w}$ の頻度 $f_e(\boldsymbol{w})$ は、文字列 $x_{i+1}x_{i+2}\cdots x_{i+k}$ が単語 $\boldsymbol{w}$ である確率を以下に示す式で計算することで得られる。

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

確率的単語分割コーパス中における単語 $\boldsymbol{w}$ の扱いを図 1 に示す。

$f_e(\boldsymbol{w})$ は 1 箇所の $\boldsymbol{w}$ に対する期待頻度なので、単語 1-gram 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 頻度（$n \geq 2$）についても同様の方法で計算を行うことができる。

単語 n-gram 確率の推定方法は決定的に分割されたコーパスの場合と同様に行うことができ、以下のようにして計算できる。

$$P(w) = f_r(w)/f_r(\cdot) \quad (n = 1)$$
$$P(w_n | \boldsymbol{w}_1^{n-1}) = f_r(\boldsymbol{w}_1^n)/f_r(\boldsymbol{w}_1^{n-1}) \quad (n \geq 2)$$

3.2 疑似確率単語分割コーパス

前項で紹介した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には、n-gram 期待頻度の計算に多くの計算時間が必要となる。また、多くの n-gram 言語モデル学習ツールは、決定的に単語分割されたコーパスの利用を前提としている。提案手法においては、未知語候補の抽出と分野適応のための音声認識用言語モデル構築を同時に行うため、既存のツールを用いることが

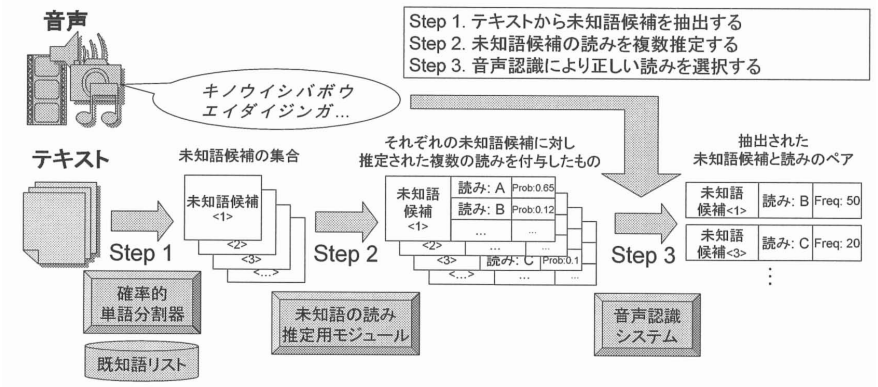

図 2: 提案手法の概要図

できるほうが都合が良い。そこで、本節では疑似確率的単語分割コーパス [7] による確率的単語分割コーパスの近似について述べる。

擬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た単語分割済みコーパスを得ることができる。これを疑似確率的単語分割コーパスと呼ぶ。

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 N 回行って得られる N 倍の単語分割済みコーパスを単語 n-gram 頻度の計数の対象とすればよい。

4 未知語候補と読みの自動獲得

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について述べる。図 2 に本節で述べる手法の概要を示す。

4.1 未知語候補の抽出

テキストから擬似確率的単語分割コーパスを作成し、テキストから未知語候補の表記を抽出する。具体的な手続きを以下に示す。

1. テキストに単語境界確率を付与し、確率的単語分割コーパスを作成する。
2. 3.2 節で紹介した方法を用いて、N 個の疑似確率的単語分割コーパスを作成する。
3. 作成したコーパスにおいて既知語リストにない単語のうち、F_{th} 回以上出現したものを未知語候補として抽出する。

これにより抽出した未知語候補の集合が、本論文において獲得対象になる単語集合である。

4.2 未知語候補の読み推定

抽出した未知語候補の読みを 2.2 節で述べた文字 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.) $P(\text{イシヤブ}, \text{石破}) = 0.53$
   $P(\text{イシバ}, \text{石破}) = 0.22$
   ⋮

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式 (2) によって正しく読みを推

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
44,561	1,072,356	1,571,285

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。

4.3 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 [8] の音声認識結果には、単語ごとに信頼度 [9] が付与されている。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が期待される。

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. 既知語リストから作成される音声認識システムの発音辞書に未知語候補とその読み推定結果を追加する。
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. 用意した音響モデル、作成した言語モデルと発音辞書を用いて 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} 以上の未知語候補と読みを抽出する。

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が得られ、またそれぞれの出現頻度が得られる。また、未知語以外の部分に対しても単語信頼度でフィルタリングをかけることによって、一部もしくは全ての音声認識結果を読みの付与された単語分割済みコーパスとして獲得できる。

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
新聞	4,334,748	165,905,535
ウェブニュース	30,552	2,528,722

表 3: 音声認識用の発音辞書

	単語数	エントリ数
既知語	29,777	31,640
未知語候補	6,830	18,547

5 評価

本論文で提案した手法の評価として、ウェブニュースを対象とした読み推定実験を行った。本章ではその詳細を述べる。

5.1 実験の条件

実験とその結果に対する評価を行う際に必要なリソース、設定する必要のあるパラメータを以下に示す。

単語境界確率の推定に用いる最大エントロピーモデルの学習のために、あらかじめ単語分割が行われ、人手によって修正された辞書の例文と新聞からなるコーパスを用意した。このコーパスの各単語には読みも付与されており、後述するテキストの読み推定においても用いられる。また、既知語リストとしてこのコーパス内に出現する全ての単語と読みのエントリの集合を用いた。このコーパスの詳細を表 1 に示す。

未知語候補の抽出元として、単語境界と読みの情報が共にないテキストを用意した。一般分野として 2002 年から 2006 年の 5 年間の新聞を用い、適応分野として 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用いた。実験では適応分野のテキストから未知語候補をより多く抽出するため、新聞に対して決定的な単語分割を行い、ウェブニュースから擬似確率的単語分割コーパスを $N = 10$ として作成した。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、複数の読み推定による発音辞書のサイズ増大を考慮して調節し、$F_{th} = 100$ とした。未知語候補の抽出に用いた新聞と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の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。音響モデルは、新聞記事読み上

表 4: 読み推定実験

	recall	relative
ベースライン	99.16(%)	-
	= 29,056/29,303	-
+ 音声認識結果	99.40(%)	29.1(%)
	= 29,128/29,303	= 72 / 247
+ 未知語候補と	99.31(%)	17.8(%)
前後の 1 単語	= 29,100/29,303	= 44 / 247

げ音声コーパス（JNAS）から学習した 3,000 状態、64 混合状態共有 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もの ($C_{th} > 0.1$) で、2 回以上認識された未知語候補を最終的な獲得単語とした。

5.2 テキスト読み推定による評価

テキスト読み推定の評価基準として、再現率 (recall) ならびにベースラインからの読み誤りの減少率 (relative) を用いる。

実験では、表 1 に示したコーパスから単語と読みを単位とした 2-gram モデルを学習した場合をベースラインとし、実験時の音声認識結果のうち一定値以上の単語信頼度が与えられている部分 ($C_{th} > 0.1$) の音素列を読みに変換して学習コーパスに追加した場合と比較した。また、このうち未知語周辺の文脈情報の影響を明らかにするために、信頼度の高い音声認識結果の中から未知語候補の前後 1 単語のみを抽出したものを学習コーパスに追加した場合についても実験を行った。

テストセットとして、2008 年 1 月 9 日のウェブニュース 250 文に読みを付与したものを用いた。各実験結果を表 4 に示す。表 4 からわかるように、本実験で用いた学習コーパスはテストセットに対して単独で高い読み推定精度を実現している。これは新聞が学習コーパスに含まれていることを考えると当然であるといえるが、その時期はテストセットのものとは離れているため、短期的に発生する人名、あるいはたまたま学習コーパスに含まれなかった単語などの未知語に対して読みを推定することは難しい。本論文で提案した手法を用いて得られる音声認識結果を適応コーパスとして用いることで、読み誤り文字数は 247 文字中 72 文字 (29.1%) が改善された。また、追加する音声認識結果を未知語の前後 1 単語に限定しても、44 文字 (17.8%) が改善されている。したがって本手法を用いて未知語とその周辺の文脈を獲得することは、読み推定精度の改善を行う上で有効な手段であると言える。

6 おわりに

本論文では類似した話題を扱っているテキストと音声を利用し、音声認識を用いて未知語とその読みを自動獲得する手法を提案した。実験の結果、音声認識の単語信頼度により獲得した単語をフィルタリングすることで、より頑健な未知語獲得が可能であることを確認した。また、本手法で得られる音声認識結果の一部を信頼度により適切に選択し、テキスト読み推定の適応用学習コーパスとして用いることで、未知語周辺の読み推定を正しく行うことが可能であることを確認した。

参考文献

[1] 広瀬啓吉. 音声合成技術. 情報処理, Vol. 38, No.11, 1997.

[2] 永田昌明. 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. 情処論, Vol. 40, No.9, 1999.

[3] 内元清貴, 関根聡, 伊佐原均. 最大エントロピーモデルに基づく形態素解析 未知語の問題の解決策. 自然言語処理, Vol. 8, No.1, 2001.

[4] 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. 情処論, Vol. 48, No.2 (2007).

[5] 永田昌明. 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. 情処論, Vol. 40, No.9 (1999).

[6] 長野徹, 森信介, 西村雅史. 確率モデルを用いた読み及びアクセント推定. 情処研報, 2005-SLP-57, 2005.

[7] 森信介, 倉田岳人, 小田祐樹. 最大エントロピー法による単語境界確率の推定. 情処研報, 2006-SLP-63, 2006.

[8] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius – an open source real-time large vocabulary recognition engine. In *Proc. of the EuroSpeech2001*, pp. 1691–1694, 2001.

[9] 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. 情処研報, 2003-SLP-49, 2003.