音声とテキストからの語彙獲得による読み推定精度の向上

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、音声言語処理技術を用いた研究が盛んに行われている。その中の代表的なものの 1 つとして音声合成によるテキスト読み上げ (TTS, Text-To-Speech) がある。TTS は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して出力する。

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いて特に問題となるのは未知語の存在であり、これは読み上げ対象のテキストが新聞やニュースなど、内容に日々変化のあるものであれば必ず発生する。この問題に対処する最も確実な方法は、読み上げ対象の分野適応コーパスから適切な単位で未知語を切り出し、読みを与えて辞書に追加していくというものである。しかしこの作業を継続的に人手で行うには多くのコストを要するので、自動化が強く望まれている。

未知語への対処を自動化する方法には、大きく分けて (a) 未知語モデルによってテキスト中の未知語の範囲を同定する、(b) 未知語を自動獲得して辞書に登録する、という 2 種類があり、このいずれかを行うことはテキスト解析システムを用いる上で必須であるといえる。このような研究としては、テキスト単語分割結果の N-best 候補内に存在する単語から文字種を考慮して未知語候補を選び、テキスト内での出現頻度の期待値を単語らしさとして未知語を自動獲得する手法 [2] や、最大エントロピーモデルを用いた形態素解析 [3] といった手法が提案されている。しかしこのような手法では単語表記の獲得はできても、TTS や音声認識などの音声言語処理アプリケーションを用いるために必要な読みの情報が獲得できない。

そこで、本論文ではテキストの読み上げを行う際に出現すると予想される未知語とその読みを自動獲得し、辞書に登録してテキストの読み推定精度の向上を実現する手法を提案する。未知語とその読みの自動獲得については、音声認識を用いた手法 [4] が提案されており、獲得した未知語による音声認識精度の向上が報告されている。

本論文においては、ウェブニューステキストの要約や全文の読み上げを行うことを想定し、(a) ある一定期間内のニュースから、その後のニュースに出現することが期待される未知語の候補を抽出する、(b) 未知語候補の読みを複数推定してその中から音声認識で正しい読みを選択する、という手法で未知語と読みを獲得する実験を行った。獲得された単語と読みを辞書に追加し、未知語候補の獲得対象としたニュースの翌日に取得されたウェブニューステキストを対象に読み推定を行ったところ、読み推定精度の向上を確認した。

[1] 合成音声の客観的評価には、音節明瞭度や単語了解度 [1] といった指標が用いられる。

2　N-gram モデルとその応用

N-gram モデルは、確率的言語モデルの中で最も一般的なものであり、様々な応用法が提案されている。本章では、N-gram モデルとその応用について、本論文で取り扱うものに絞って述べる。

2.1　N-gram モデルを用いた自動単語分割

日本語は分かち書きがされていないため、テキストを用いた言語処理においては、多くの場合まず単語分割[2]が必要となる。本論文では、単語 N-gram モデルに基づく単語分割器 [5] を紹介する。この方法では、以下の式で示されるように、与えられた文字列 $\boldsymbol{c}_1^h = c_1 c_2 \cdots c_h$ から得られるあらゆる単語列 $\boldsymbol{w}_1^n = w_1 w_2 \cdots w_n$ のうち、生成確率 $P_w(\boldsymbol{w}_1^n)$ が最大となる単語列 $\hat{\boldsymbol{w}}$ を単語分割結果とする。

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}_1^n = \boldsymbol{c}_1^h}{\mathbf{argmax}} P_w(\boldsymbol{w}_1^n)$$

ここで $P_w(\boldsymbol{w}_1^n)$ を単語 N-gram モデル $P_{w,N}(\boldsymbol{w}_1^n)$ とすると、

$$P_{w,N}(\boldsymbol{w}_1^n) = \prod_{i=1}^{n+1} P(w_i | \boldsymbol{w}_{i-N+1}^{i-1})$$

となる。上式における $w_i (i \leq 0)$ と w_{n+1} は、それぞれ文頭と文末を示す特別な記号である。

2.2　未知語モデル

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、任意のテキストを単語分割するためには未知語モデルが必要となる。単語列中のある単語 $w_i = \boldsymbol{c}_1^h$ が未知語であった場合、まず未知語全体を表す特殊記号 `UW` を含む単語 N-gram 生成確率 $P_{w,N}(\texttt{UW} | \boldsymbol{w}_{i-N+1}^{i-1})$ を与え、その中の $\boldsymbol{c}_1^h$ の生成確率を以下の文字 N-gram モデルで計算する。

$$P_{c,N}(\boldsymbol{c}_1^h) = \prod_{i=1}^{h+1} P(c_i | \boldsymbol{c}_{i-N+1}^{i-1})$$

これにより、未知語 $w_i = \boldsymbol{c}_1^h$ の生成確率は

$$P(w_i | \boldsymbol{w}_{i-N+1}^{i-1}) = P_{c,N}(\boldsymbol{c}_1^h) P_{w,N}(\texttt{UW} | \boldsymbol{w}_{i-N+1}^{i-1})$$

となる。このため、前もって学習コーパス中に出現する全単語を既知語集合と低頻度語の集合に分け、低頻

[2] 以後、「単語」や「未知語」はそれぞれの表記という意味で扱う。

度語の集合に与えられた確率の合計を未知語全体の生成確率とする。未知文字表記を予測する場合は、同様に未知文字を表す記号 UC を用いて未知文字全体の生成確率を $P_{c,N}(\mathtt{UC}|\boldsymbol{c}_{i-N+1}^{i-1})$ によって与え、日本語で用いられる全ての文字 (JIS X 0208 文字セットの 6,879 種類) から既知文字を除いたものを未知文字とし、一様分布によって各未知文字の生成確率を与える。

2.3 外部辞書

2.2 節で述べた文字 N-gram による未知語モデルは、未知語だけでなく既知語の生成確率も計算することができる。しかし既知語を予測する際は単語 N-gram モデルによってのみ確率が与えられるため、文字 N-gram モデルによる既知語の生成確率が計算されることはない。このため、既知語集合 $\mathbf{D}_{in}$ に含まれる単語の文字 N-gram モデルによる生成確率の和を、未知語のうち辞書などから単語と考えられる文字列の集合に等しく配分して未知語モデルの精度を上げる手法が提案されている [6]。このような文字列の集合を外部辞書 $\mathbf{D}_{ex}$ と定義し、2.2 節で述べた学習コーパス中の低頻度語の集合を外部辞書として用いることで解析精度の向上が報告されている。この外部辞書を用いた未知語モデルによる、未知語 $w_u = \boldsymbol{c}_1^h (w_u \notin \mathbf{D}_{in})$ の生成確率 $P'_{c,N}(w_u)$ は以下の式で表される。

$$P'_{c,N}(w_u) = \begin{cases} P_{c,N}(w_u) + \frac{1}{|\mathbf{D}_{ex}|} \sum_{w_k \in \mathbf{D}_{in}} P_{c,N}(w_k) & \text{if } w_u \in \mathbf{D}_{ex} \\ P_{c,N}(w_u) & \text{if } w_u \notin \mathbf{D}_{ex} \end{cases}$$

この外部辞書を用いた未知語モデルは (a) 単語を追加するだけで未知語の解析精度が向上する (b) 未知語が発生しない限り解析精度に影響が出ない、という特徴があるため、本論文で取り扱うような未知語を辞書に追加するタスクの評価に適している。

2.4 確率的単語分割による言語モデル構築

言語モデルを用いる対象となる分野が一般分野と異なっている場合、その分野の言語的特徴を反映しているような適応コーパスを解析して一般分野のコーパスとともに用いる必要がある。しかし適応コーパスを解析する際には、分野特有の単語 (未知語) の周辺で分割誤りを起こすことが多い。この問題に対処する方法の 1 つとして、単語境界確率の推定による確率的単語分割コーパスの作成を行う手法が提案されており [7]、決定的な単語分割を行ったコーパスを用いる場合よりも高い予測力を持つ N-gram 言語モデルの構築が可能であると報告されている。確率的単語分割コーパスはコーパス内の文字列 c_i, c_{i+1} の間に単語境界が存在する確率 P_i を与えたものとして定義される。確率的単語分割コーパスを用いることで、コーパス内の全ての部分文字列を単語として N-gram 確率を推定することが可能になる。しかし確率的単語分割コーパスには N-gram 確率の計算に多くの計算量を要するという問題があるため、これを近似するための方法として疑似確率的単語分割コーパスが提案されている [7]。具体的には、確率的単語分割コーパスの文字境界に付与された単語境界確率 P_i と乱数 $r_i (0 \leq r_i < 1)$ を比較し、$r_i < P_i$ ならば文字列 c_i, c_{i+1} の間を単語境界とすることで決定的な単語分割を行ったものを疑似確率的単語分割コーパスと呼ぶ。この単語分割を M 回行うことで、単語境界の曖昧な部分 (P_i が 0.5 に近いほど曖昧といえる) で分割位置が異なる M 個のコーパスが作成される。これらを別々のコーパスとして扱うことで、通常の N-gram 確率の計算方法で確率的単語分割の特徴を近似した N-gram 確率を求めることができる。この N-gram 確率は疑似確率分割の回数 M を大きくするほど良い近似となる。

3 N-gram モデルを用いた読み推定

本章では、N-gram モデルを用いて文の読み推定を行う方法について述べる。

3.1 文の読み推定

文の読み推定は、単語分割と単語ごとの読みタグ付けを同時に行う問題とみなすことができる。これは 2.1〜2.3 節で述べた手法を拡張することで実現できる。具体的には、N-gram モデルの単位を単語 w から単語 w と読み y の組 $\langle w, y \rangle$ として N-gram モデルを構築すればよい。ある 1 文とその読みの組を $\langle \boldsymbol{w}, \boldsymbol{y} \rangle_1^n = \langle w, y \rangle_1 \langle w, y \rangle_2 \cdots \langle w, y \rangle_n$ とすると、その生成確率は

$$P_{\langle w,y \rangle,N}(\langle \boldsymbol{w}, \boldsymbol{y} \rangle_1^n) = \prod_{i=1}^{n+1} P(\langle w, y \rangle_i | \langle \boldsymbol{w}, \boldsymbol{y} \rangle_{i-N+1}^{i-1})$$

で与えられる。未知の $\langle w, y \rangle$ を予測する際は、2.3 節で延べた未知語モデル中の c を文字 c と読み y の組 $\langle c, y \rangle$ に拡張したものを用いる。

3.2 単語の読み推定

3.1 節で述べた文の読み推定と同様の手法で単語の読みを推定することができる。これは文字ごとの読みタグ付けを行う問題とみなすことができ、このとき N-gram モデルの単位は文字 c と読み y の組 $\langle c, y \rangle$ となる。N-gram モデル学習の際は、単語を 1 文字ごとに分割して読みを与えたコーパスを用いればよい[3]。文の場合と同様に、単語と読みの組を $\langle \boldsymbol{c}, \boldsymbol{y} \rangle_1^h = \langle c, y \rangle_1 \langle c, y \rangle_2 \cdots \langle c, y \rangle_h$ とすると、その生成確率は

$$P_{\langle c,y \rangle,N}(\langle \boldsymbol{c}, \boldsymbol{y} \rangle_1^h) = \prod_{i=1}^{h+1} P(\langle c, y \rangle_i | \langle \boldsymbol{c}, \boldsymbol{y} \rangle_{i-N+1}^{i-1})$$

となる。単語の読み推定中に学習コーパス中に存在しない未知文字表記が発生した場合は単漢字辞書を参照し、一様分布によって $\langle c, y \rangle$ の生成確率を割り当てる (5 章の実験で用いた単漢字辞書中における $\langle c, y \rangle$ の異なり総数は 13,463 であった)。

[3] 英単語など、文字ごとの読みを与えられないものはひとまとまりの読みに対応する文字列を 1 文字とみなす必要がある。

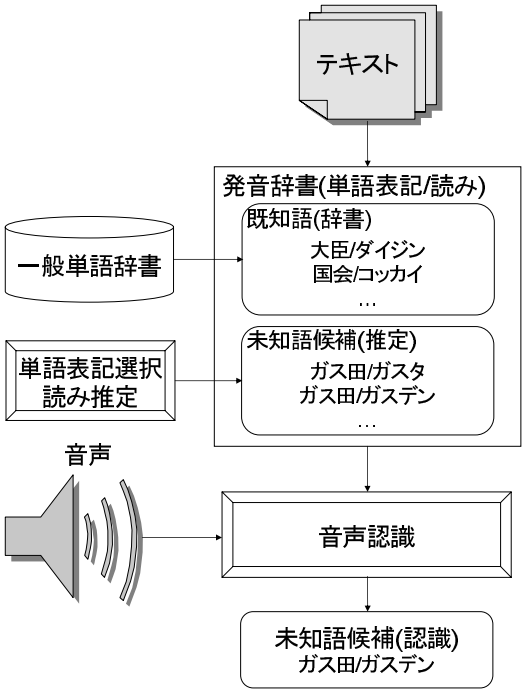

図 1: 音声とテキストからの未知語候補と読みの獲得

4 音声認識による未知語候補と読みの獲得

本章では、テキスト内に出現する未知語候補を音声認識システムによって正しい読みとともに獲得する方法について述べる。処理全体の流れを図 1 に示す。

4.1 テキストの単語分割

2.4 節で述べた確率的単語分割を用いると、テキスト内に出現する全ての部分文字列に対する N-gram 確率の計算を行うことができる。実際には、(a)N-gram 確率の計算に多くの時間が必要、(b) 全ての部分文字列を獲得対象とする必要はない、という 2 点を考慮し、疑似確率的単語分割 [7] を用いて確率的単語分割の近似を行う。疑似確率的単語分割を適応対象のテキストに対して複数回行うことで、異なった複数の単語分割済みテキストを言語モデル適応に用いることができる。

4.2 未知語候補の選択

音声認識用言語モデルの語彙、ならびに未知語候補を決定するため、一般単語辞書を参照してテキスト内の全単語から以下の 2 種類の単語を抽出する。

既知語 一般単語辞書に登録されている単語

未知語候補 一般単語辞書に未登録の単語のうち、出現頻度が F_{th} 以上のもの

言語モデルの単位は単語表記のみとして、既知語と未知語候補を合わせたものを用い、テキスト内の全単語から上記の 2 種類を除いたものは未知語クラスに割り当てられる。

表 1: 単語分割、読み付与済みのテキスト

文数	単語数	文字数
52,955	1,254,867	1,844,106

表 2: 言語モデル学習用テキスト

種類	文数	文字数
一般 (新聞)	3,671,344	152,293,814
適応 (ウェブニュース)	33,336	2,550,120

4.3 発音辞書の作成と音声認識による読みの選択

4.2 節で選択した未知語候補に正しい読みが付与されていれば、音声認識用言語モデルの学習テキストに近い分野の音声から正しい読みを獲得することができる。未知語候補の読みは 3.2 節の方法で推定することができるが、生成確率最大の $\langle \boldsymbol{c}, \boldsymbol{y} \rangle_1^h$ によって与えられる読みが正しいとは限らない。そこで、$\langle \boldsymbol{c}, \boldsymbol{y} \rangle_1^h$ の生成確率上位にある読み推定結果から得られる未知語候補とその読みの中に正しいものがあると仮定し、それらを発音辞書のエントリとして複数加えておく (既知語に関しては、一般単語辞書の読みを与える)。この発音辞書を用いて音声認識を行うことにより、誤った音素列を出現させることなく正しい音素列から未知語候補と読みを取得することが期待される。

5 実験

本章では、未知語獲得ならびにテキストの読み推定を行った際の実験条件とその結果について述べる。

5.1 音声認識による未知語候補と読みの自動獲得

4 章で述べた手法を用いて未知語候補とその正しい読みを獲得する実験を行った。音声認識には julius3.5.3 を使用した。確率的単語分割のモデル学習には、辞書の例文や新聞からなる人手で単語分割が行われたテキストを用いた。その内容は表 1 の通りである。このテキストには読みも付与されており、5.2 節で述べる読み推定のモデル学習にも用いた。

4.1 節で述べた、乱数を分割の閾値とする疑似確率分割はウェブニューステキストに対してのみ行い、新聞テキストを分割する際には単語境界確率が 0.5 よりも大きいか小さいかで決定的に分割を行う。これは読み上げ対象であるウェブニューステキストのみ単語分割位置に曖昧性を持たせ、多くの未知語候補が得られることを期待しての処理である。疑似確率的単語分割の回数は $M = 10$ とした。これにより、言語モデル適応用ウェブニューステキストのサイズは表 2 で示した数の 10 倍 (333,360 文、25,501,200 文字) となる。

未知語候補選択の際、言語モデル学習用テキスト中の出現頻度の閾値は、(a) 疑似確率分割の性質上、低頻度語には分割誤りの文字列が多い、(b) 後に複数の読みを推定することで発音辞書エントリが増加する、という 2 点を考慮して、$F_{th} = 200$ とした。

表 3: 言語モデルの語彙数と発音辞書のエントリ数

種類	語彙	発音辞書
既知語	32,114	34,338
未知語候補	2,999	7,721

未知語候補の読み推定には $\langle c, y\rangle$ を単位とする 2-gram モデルを用いた。2-gram モデルの作成には、読みの与えられたコーパスから単語を抽出し、文字ごとの分割と読みの付与を行って $\langle c, y\rangle$ ごとに区切られたコーパスを作成する必要がある。本実験では EDR コーパス [8] の一部 (187,022 文) をもとに、単漢字辞書を引いて文字と読みごとのアラインメントを取ることができる自立語を抽出し、文字と読みごとに分割されたコーパスを作成した (抽出された自立語の総数は 255,767 個、文字と読みの組 $\langle c, y\rangle$ の総数は 793,918 個であった)。

音声認識言語モデル作成用の語彙 (単語表記のみ) とそれに読みを付与した後の発音辞書エントリ数を表 3 に示す。未知語候補に関しては読み推定の生成確率上位 5 個[4]をとって発音辞書に追加し、それぞれの読み推定結果 5 個の中での相対確率を発音辞書に与える確率とした。既知語の読みが一般単語辞書中に複数ある場合にはそれぞれの読みごとに等しい確率を与えた。なお、今回用いた一般単語辞書には表 1 のテキスト中に出現する全ての単語と読みの組 (エントリ数 45,282) 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) より学習した 3000 状態、64 混合の状態共有 triphone HMM を用いた。

以上で述べた言語モデル、発音辞書、音響モデルを用いてニュース音声 (30 分 ×34) [5]から未知語候補を認識した。音声認識結果として得られた単語列の中にある未知語候補と、そのときの音素列を読みに変換したものの組を計数して、2 回以上出現したものを最終的に獲得した。これにより、281 個の未知語候補と読みの組が得られた。その例を以下に示す。

{ガス田/ガスデン, 和解案/ワカイアン, サブプライムローン/サブプライムローン, $\cdots$} [6]

音声中で実際に発音されている単語と読みが獲得されたかを上位の 100 個について検証したところ、80 個が正しい読みであった。また、発音はされていない (音声認識誤りを起こしている) が一般的に正しい読みであり、かつ読みに曖昧性のないものを含めると 95 個が正しい読みであった。

5.2 テキストの読み推定実験

単語分割、読み付与済みのテキスト (表 1) を学習コーパスとして、単語と読みの組 $\langle w, y\rangle$ を単位とする 2-gram による読み推定を行った。学習コーパス中に 1 回のみ出現した $\langle w, y\rangle$ の集合を外部辞書 (2.3 節を参照) とした場合をベースラインとして、獲得した未知語候補とその読みを上記の外部辞書に加えた場合との読み推定精度を比較した。テストセットには、未知語候補と読みを獲得した最後のニュース音声の翌日に取得したウェブニューステキスト (2008 年 1 月 9 日の記事中 250 文、読みの全文字数 29,339) を用い、評価指標として

$$\text{再現率} = \frac{\text{読み推定による正解文字数}}{\text{テストセットに付与された読みの全文字数}}$$

$$\text{適合率} = \frac{\text{読み推定による正解文字数}}{\text{読み推定によって得られた読みの全文字数}}$$

を用いた。読み推定の実験結果を表 4 に示す。

表 4: テストセットに対する読み推定の精度

	再現率 (%)	適合率 (%)
ベースライン	99.29	99.13
+ 獲得未知語候補, 読み	99.36	99.22

獲得した未知語候補と読みの追加により、読みの誤り数 207 個のうち 19 個 (9.2%) が削減された。具体的には、主に難しい読みをもつ人名[7]が出現する部分で誤りが改善されており、期待通りの結果が出ているといえる。

6 おわりに

音声合成システムにおいては、テキスト中に現れる単語、特に未知語の読みを誤らないことが重要である。本論文では、ウェブニューステキストの読み上げを想定して、未知語候補の読みを複数推定した中から音声認識で正しい読みを獲得する手法について述べた。また、未知語候補獲得対象として用いたニュースの翌日のウェブニューステキストをテストセットとし、獲得した未知語候補と読みを辞書に追加して文の読み推定を行った結果、精度の改善が確認された。

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] 倉田岳人, 森信介, 西村雅史: 日本語生コーパスから自動獲得した未知語と言語モデルによる大語彙連続音声認識, 情処研報, 2005-SLP-57 (2005).

[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] 森信介, 長尾眞: 形態素クラスタリングによる形態素解析精度の向上, 自然言語処理, Vol. 5, No.2 (1998).

[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] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (1993).

[4]実際には仮名のみの単語など、候補が 5 個以下のものもある。

[5]07/12/23 を除く 07/12/05〜08/01/08 の 34 日分を用いた。

[6]その他、ニュースで頻出する人名などが得られた。

[7]3.2 節の単語読み推定では正しい読みが生成確率最大とならなかった。