確率モデルを用いた読み及びアクセント推定

長野 徹　森 信介　西村 雅史
日本アイ・ビー・エム東京基礎研究所
〒 242-8502 神奈川県大和市下鶴間 1623-14
{tohru3,forest,nisimura}@jp.ibm.com

あ ら ま し

本論文では、規則音声合成における読みとアクセントを、確率モデルに基づき同時に推定する手法を提案し、その実験結果を報告する。規則音声合成において、任意の入力テキストに対し、正しい音韻情報と韻律情報を生成することは、自然な合成音声を得るために重要な要件である。本研究では、入力テキストに対し、最も基本的な音韻情報と韻律情報である読みとアクセントを付与する問題を取り扱う。日本語の場合、入力テキストは一般的に漢字仮名交じり文であり、複数の読み候補から正しい読みを推定する必要があるとともに、その読みに対して正しいアクセントを推定する必要がある。従来、日本語テキストに対して、形態素解析・読み付与・アクセント句決定・アクセント核決定、という手順を段階的に行うことで、読みとアクセントを決定することが多かったが、本研究では、表記（単語境界）・品詞・読み・アクセントを 1 つの単位とみなし、n-gram モデルを用いて同時に推定する。実験では、ルールに基づきアクセント句およびアクセント核を決定する逐次的な手法との比較を行った。その結果、確率モデルに基づく手法の精度がルールに基づく手法の精度を上回ることを確認した。

キーワード　音声合成　言語モデル　読み　アクセント　未知語

A Stochastic Approach to Phoneme and Accent Estimation

Tohru NAGANO, Shinsuke MORI, Masafumi NISHIMURA
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma Yamatoshi Kanagawaken 242-8502 Japan
{tohru3,forest,nisimura}@jp.ibm.com

Abstract

We present a new stochastic approach to estimate accurately phonemes and accents for Japanese TTS (Text-to-Speech) systems. Front-end process of TTS system assigns phonemes and accents to an input plain text, which is critical for creating intelligible and natural speech. Rule-based approaches that build hierarchical structures are widely used for this purpose. However, considering scalability and the ease of domain adaptation, rule-based approaches have well-known limitations. In this paper, we present a stochastic method based on an n-gram model for phonemes and accents estimation. The proposed method estimates not only phonemes and accents but word segmentation and part-of-speech (POS) simultaneously. We implemented a system for Japanese which solves tokenization, linguistic annotation, text-to-phonemes conversion, homograph disambiguation, and accents generation at the same time, and observed promising results.

Key Words　Text-to-Speech, Language Model, Phonemes, Accents, Out-Of-Vocabulary

1　はじめに

語彙の制限のない任意のテキストを入力として、人間の発する音声と同様の音声を出力することが、規則音声合成の一つの最終目標である。一般的に、テキストつまり文字列を入力として音声を合成するためには、主に自然言語処理を用いたテキスト処理技術と、信号処理を用いた音声合成技術を組み合わせて実現される。テキスト処理においては、任意の入力テキストに対し、正しい音韻情報と韻律情報を生成することが、自然な合成音声を得るために必要な要件である。このテキスト処理で生成された音韻情報と韻律情報を元に、音声合成技術によって音声素片を組み合わせることで、音声が合成される。

テキストのみを入力として、正しい音声を生成するためには、テキストの構成要素である単語だけでなく、単語列として表される文全体が、言語的・音声的にどのような性質であるかを知る必要がある。本研究では、入力テキストに対し、最も基本的な音韻情報と韻律情報である読み仮名とアクセントを付与する問題を取り扱う。日本語の場合、入力テキストは一般的に漢字仮名交じり文であり、複数の読み候補から正しい読み仮名を推定する必要があるとともに、その読み仮名に対して正しいアクセントを推定する必要がある。日本語テキストを発話する際に、読み仮名が重要であることは言うまでもないが、アクセントも同様に重要な要素である。例えば、『庭には二羽、鶏がいる』という文には「にわ」という音が 3 回出現するが、このいずれかのアクセントを誤ると、単に発話の自然さに欠けるというだけではなく、文として意味が通じない、または違う意味になる、といったことが容易に推測できる。

従来、日本語テキストに対して、韻律・音韻情報を付与する手法として、形態素解析・読み付与・アクセント句決定・アクセント核決定、という手順を段階的に行うことで、読み仮名とアクセントを決定することが多かった。また、他言語においても、テキスト処理部に関しては、イントネーション句・アクセント句、といった階層構造を仮定し、各階層構造を決定木等の統計的な手法を用いてトップダウンに決定する手法が一般的である [2]。しかし、読み及びアクセントが発話順に順次決定するモデルだとすると、前者の場合、本来事後的に決定すべきアクセント句境界を先に決定していることから、最適性が保障されない、という問題がある。最適性については、各段階で N-best の解を出力して、順次、逐次的な処理を行うことで同じ効果を期待出来るが、組み合わせのモデルが複雑になる上、やはり、アクセント句を、各単語のアクセントに先行して求めることになる。

上記のことから、本研究では、表記（単語境界）・品詞・読み・アクセントを 1 つの単位とみなし、n-gram モデルを用いて同時に推定する手法を提案する。つまり、逐次的な処理ではなく、1 つの確率モデルで 4 つの値を同時に推定する。実験では、ルールに基づきアクセント句及びアクセント核を逐次的に決定する手法との比較を行った。その結果、確率モデルに基づく手法の精度がルールに基づく手法の精度を上回ることを確認した。また、同じ枠組みで品詞情報を用いないモデルに関しても併せて実験を行った。その結果、読み仮名とアクセントを決定するために品詞は有効な素性であるが、品詞を用いなくても、ある程度の精度が得られることを確認した。

2　日本語における読みとアクセント

問題は、任意の入力テキストに対して、単語分割を行い、読みとアクセントを割り当てることである。読みに関しては、他の言語も日本語と同様に、同じ表記を持ち、読みの異なる単語が存在するが、他の言語（例えば、英語や中国語）に比べて、圧倒的に読みの種類が多い。本章では、特に日本語のアクセントの特性について説明するとともに、日本語のアクセント付与に関しての関連研究について説明する。

2.1　読みとアクセント

多くの場合、日本語のアクセントは高低アクセント要素 (H)igh 及び (L)ow の列で表され、各モーラに付与される。例えば、3 モーラの単語『京都』(kyo,:,to) に対しては、(H,L,L) という 3 つのアクセント列が付与される。本研究においても、この 2 値表現のアクセントを用いる。

表 1: 辞書中の読みとアクセント

表層 w	品詞 t	読み $\boldsymbol{s}$	アクセント $\boldsymbol{a}$
京都	固有名詞	kyo : to	H L L
タワー	一般名詞	ta wa :	H L L
ホテル	一般名詞	ho te ru	H L L

一般的に、各単語の標準語（東京弁）におけるアクセントは、表記・品詞・読み、の組が定まれば一意に定まり、例えば、$\langle$『京都』, 名詞, $(kyo, :, to)\rangle$ という組に対しては上述したように、(H,L,L) というアクセントが定まる。ここで注意すべきことは、辞書に記されている、単語に対するアクセントは、単語が単独で出現したと仮定したときのアクセントである。つまり、前後に文脈が存在しないと仮定した場合のアクセントを示している。実際には、文脈によって単語のアクセントは変化し、表層・品詞・読み、の組からは一意に定まらない。例えば、名詞『タワー』が『京都』に後続した単語『京都 タワー』という複合名詞中では、『京都』のアクセントは (L,H,H) に変わり、全体のアクセントとしては、(L,H,H)(H,L,L) になる（表 2)。さらに、名詞『ホテル』が後続した単語『京都 タワー ホテル』の場合、『タワー』は (H,L,L) から (H,H,H) に変わり、全体のアクセントとしては、(L,H,H)(H,H,H)(H,L,L) になる（表 3)。合成音声のみが出力となる音声合成では、読みと同様にアクセントも、内容を正確に伝えるという点で重要な要素であり、例えば、『京都タワー』が (H,L,L)(H,L,L) というアクセントで読まれてしまうと、正確な読み (kyo,:,to) (ta,wa,:) が与えられたとしても、『今日とタワー』と解釈されてしまう可能性が非常に高い。このように、アクセント付与を誤ると、単にイントネーションが不自然になる、というだけでなく、文として意味が通じない、または違う意味に解釈される可能性が高い。

表 2: 複合名詞『京都タワー』における『京都』及び『タワー』のアクセント

表層	w	京都			タワー		
品詞	t	固有名詞			一般名詞		
読み	$\boldsymbol{s}$	kyo	:	to	ta	wa	:
アクセント	$\boldsymbol{a}$	L	H	H	H	L	L

表 3: 複合名詞『京都タワーホテル』における『京都』・『タワー』及び『ホテル』のアクセント

表層 w	京都			タワー			ホテル		
品詞 t	固有名詞			一般名詞			一般名詞		
読み $\boldsymbol{s}$	kyo	:	to	ta	wa	:	ho	te	ru
アクセント $\boldsymbol{a}$	L	H	H	H	H	H	H	L	L

以上のことから、本研究の課題は、文脈内に現れる単語に対して、正しい読みとアクセントを付与することである。言い換えると、入力文字列 $\boldsymbol{x}$ から、正しい単語境界・読み・アクセントの組 $\langle w, \boldsymbol{s}, \boldsymbol{a}\rangle$ を推定することである。ここで注意することは、読みを正しく推定しても、アクセントが正しく推定出来なければ、文脈として正しく伝わらない。読みだけでなく、アクセントも同時に推定することが重要な課題である。

2.2 関連研究

日本語の音声合成を目的とした読み付与及びアクセント付与についての研究はいくつか行われている。アクセント付与については、匂坂 [3] らにより、日本語の単語連鎖におけるアクセント核の移動規則に関して体系化が行われており、単独で出現する時の単語のアクセント型と、アクセント移動型によって、アクセント核を決定している。文をアクセント句の列 $\boldsymbol{v} = (v_1 v_2 \cdots v_l)$ とみなし、下記の手法でアクセントを生成する。

1. まず、形態素解析器等を用いて単語を分割し、単語境界・品詞・読み $\langle w, t, \boldsymbol{s}\rangle$ を決定する。
2. この形態素列 $\boldsymbol{w} = (w_1 w_2 \cdots w_h)$ を、アクセント句決定ルールを用いて、文を 1 つ以上のアクセント句の列 $\boldsymbol{w}_1^h \mapsto \boldsymbol{v}_1^l$ に分割する。
3. 各アクセント句 $\boldsymbol{v}_i (1 \leq i \leq l)$ に対して、アクセント句内の各形態素の単独でのアクセント型（アクセント核の位置）及び、アクセント移動型を辞書を参照して取得し、アクセント核をアクセント句内で移動する。この結果各アクセント句に対して 1 つのアクセント型が割り当てられる。

この手法は、標準語において、辞書が十分に整備されている状況であれば、比較的高精度が期待出来る。しかし、辞書の各見出し語に対して、アクセント型・アクセント移動型を必要とするため、新たに語を追加する場合、この両方を登録する必要がある。また、アクセント句の決定ルールにも追加する必要があり、副作用を避けながらこれらの辞書及びルールをメンテナンスしていく必要がある。また、これらのルールは形態素解析器の品詞体系に大きく依存するため、汎用性という面で不利である。

3 確率モデル

前章で、読みとアクセントの特徴について説明した。特に、アクセントに関しては、文脈によって大きく変わることを説明した。本章では、確率モデルを用いた、読み及びアクセント付与の枠組みを提案する。本モデルでは、単に読みとアクセントを推定するのみでなく、単語境界及び品詞も同時に推定する。

3.1 N-gram モデルに基づく形態素解析

確率的な言語モデルである n-gram モデルは、英語や他のヨーロッパ言語のような空白で分かち書きされた文に対する品詞タグ付けのモデルとして用いられており、永田 [7] によって、n-gram モデルを日本語や中国語のような分かち書きされない言語に対しての形態素解析のモデルとして一般化され、表層 w と品詞 t の組を一つの単位として、形態素解析のモデルとなっている。

$$\begin{aligned} & P(\langle w_1, t_1\rangle\langle w_2, t_2\rangle \cdots \langle w_h, t_h\rangle) \\ = & \prod_{i=1}^{h+1} P(\langle w_i, t_i\rangle | \langle w_{i-k}, t_{i-k}\rangle \cdots \langle w_{i-1}, t_{i-1}\rangle) \end{aligned}$$

ここで、 $k = n-1$, $\langle w_i, t_i\rangle (i \leq 0)$ は、文頭に対応する特別な記号であり、$\langle w_{h+1}, t_{h+1}\rangle$ は文末に対応する特別な記号を表す。

確率モデルを用いた形態素解析器は、形態素の表層を連結した文字列が文の文字列に等しい $\boldsymbol{x} = x_1 x_2 \cdots x_h = \boldsymbol{w}$ という制約条件下で、最も確率値の高い品詞と表層の組の列を出力する。

$$\begin{aligned} & (\langle w_1, t_1\rangle\langle w_2, t_2\rangle \cdots \langle w_h, t_h\rangle) \\ = & \ \mathbf{argmax}\, P(\langle w_1, t_1\rangle\langle w_2, t_2\rangle \cdots \langle w_q, t_q\rangle | x_1 x_2 \cdots x_h) \end{aligned}$$

3.2 N-gram モデルに基づく読み及びアクセント付与

読み及びアクセント推定を目的として、形態素 n-gram モデルを拡張する方法を提案する。まず、表層 w ・品詞 t ・読み $\boldsymbol{s}$ ・アクセント $\boldsymbol{a}$ の四つ組を一つの単位 u とした n-gram モデルの拡張を考える。つまり $u = \langle w, t, \boldsymbol{s}, \boldsymbol{a}\rangle$ となる。この 四つ組 n-gram モデル M_u による、四つ組列 $u_1, u_2 \cdots u_h$ の生成確率は以下の式で表される。

$$M_u(u_1 u_2 \cdots u_h) = \prod_{i=1}^{h+1} P(u_i | u_{i-k} \cdots u_{i-2} u_{i-1}) \qquad (1)$$

形態素解析と同様に、四つ組 n-gram モデルの確率値も、コーパスの頻度から最尤推定される。

形態素列の表層を結合させた文字列は、元の文の文字列と一致している $\boldsymbol{x} = x_1x_2\cdots x_h = \boldsymbol{w}$ という制約条件の元で下記の式において解探索を行う。

$$\hat{\boldsymbol{u}} = \mathbf{argmax}\, M_u(u_1u_2\cdots u_q|x_1x_2\cdots x_h) \qquad (2)$$

解探索に関しては、動的計画法を用いて効率的に解けることが分かっており [8]、解探索の計算量は入力文字列長に比例する。

3.3 未知語モデル

四つ組 n-gram モデルは、文を四つ組の列 $\boldsymbol{u} = u_1u_2\cdots u_h$ の表層を連結したものと見なし、各形態素を文の先頭から順に予測する。しかし、日本語の形態素を全て列挙することは出来ないので、未知形態素の扱いが避けられない問題となる。通常は、 式 (1) の確率値は、コーパスの頻度から最尤推定されるが、文に未知語（コーパスに出現しない語）を含む場合、 M_u の確率値は 0 となり、未知語を含む形態素列が、式 (2) によって選ばれることは無い。ただ、実際の問題として、入力テキスト中に出現する可能性のある全ての形態素が、学習コーパス中に出現することは望めない。

この問題に対処するため、未知形態素に対応する特別な記号 UNK を用意し、既知の形態素以外はこの記号を用いた未知語モデルにより、与えられる確率で生成されることとする。未知形態素に対応する特別な記号は、かならずしも唯一である必要はなく、品詞などの情報を用いて区別される複数の記号であってもよい。以下の説明では、各品詞 t に対して未知形態素に対応する記号 $\mathtt{UNK}_t$ を設ける。式 (1) における確率値 P は 未知語モデル M_x を含み、以下のように表される。

$$P(u_i|u_{i-k}\cdots u_{i-2}u_{i-1}) \qquad (3)$$
$$= \begin{cases} P(u_i|u_{i-k}\cdots u_{i-2}u_{i-1}) & \text{if } u_i \in \mathcal{V} \\ P(\mathtt{UNK}_{t_i}|u_{i-k}\cdots u_{i-2}u_{i-1})M_x(u_i|t_i) & \text{if } u_i \notin \mathcal{V} \end{cases}$$

任意の入力テキストに対して未知語モデルが適用可能であるためには、下記の条件が成り立つ必要がある。

- 任意の文字に対して、出現確率が 0 より大きいこと[1]
- 未知入力文字列に対して 最も確率値の高い品詞・読み・アクセントを推定すること

このような制約を満たすモデルとして、表層と読みの組 $\langle x, \boldsymbol{s}\rangle$ を単位とした未知語読み n-gram モデルを各品詞毎に定義する。

$$M_x(\langle x_1, \boldsymbol{s}_1\rangle\langle x_2, \boldsymbol{s}_2\rangle\cdots\langle x_{h'}, \boldsymbol{s}_{h'}\rangle|t) \qquad (4)$$
$$= \prod_{i=1}^{h'+1} P(\langle x_i, \boldsymbol{s}_i\rangle|\langle x_{i-k}, \boldsymbol{s}_{i-k}\rangle\cdots\langle x_{i-1}, \boldsymbol{s}_{i-1}\rangle, t)$$

[1] この条件は、任意の文字列の入力に対して、解が得られることを保障する。

また、未知語のアクセントに関してはコーパス中の全単語で最も頻度の高いアクセント LHH...H を用いた。これは先頭のモーラのアクセント要素のみが L で、2 モーラ目以降のアクセント要素が H であるアクセントである。このアクセントは学習コーパス中に現れる全形態素のアクセントのうち全体の 37.28% を占める[2]。

以上から、未知語モデルによって推定される四つ組の出現確率 $u = \langle w, t, \boldsymbol{s}, \boldsymbol{a}\rangle$ は、品詞 t 毎に、下記の式で与えられる。

$$M_x(u|t) = \begin{cases} M_x(\langle x_1, \boldsymbol{s}_1\rangle\langle x_2, \boldsymbol{s}_2\rangle\cdots\langle x_{h'}, \boldsymbol{s}_{h'}\rangle|t) & \\ \qquad \text{if } \boldsymbol{a} = \mathtt{LHH...} & (5) \\ 0 \qquad \text{それ以外} & \end{cases}$$

ここで、未知形態素の表層文字列は、各文字を結合した結果に等しく ($w = x_1x_2\cdots x_{h'}$) 、音素列の長さはアクセント要素の長さに等しくなければならない ($|\boldsymbol{s}| = |\boldsymbol{a}|$)。

最終的に、我々の提案する確率的モデルに基づく処理部は、式 (1)、(3)、(5) の生成確率を計算し、最終的に最も生成確率の高い解が式 (2) によって与えられる。

3.4 パラメータ推定

式 (3) のパラメータは、コーパスの頻度から最尤推定される。コーパスの各文は予め形態素に区切られており、各形態素には、品詞・読み・アクセントが付与されている。学習コーパスは 9 個に分割され、四つ組が分割されたコーパスの 1 個のみに出現する場合、コーパス中の四つ組を品詞付きの未知形態素に対応する特別な記号 $\mathtt{UNK}_t$ に置き換える。式 (3) における n-gram 確率値はより低次のモデルとの補間を行う。補間係数は、削除補間法 [9] によって推定される。式 (4) は四つ組の n-gram 確率値を計算する際に、未知形態素に対応する特別な記号によって置き換えられた低頻度の形態素から計算される。

これらの低頻度の形態素からは品詞毎に、文字列と音素列の組が取り出される。この文字列と音素列の組のアラインメントは、辞書を用いて自動的に行われる。この辞書には、全ての文字 0-gram に対してあり得る読みが記述されている。式 (4) の各品詞に対するパラメータは、このアラインされた形態素集合から推定される [3]。式 (4) の n-gram 確率も同様に低次のモデルと補間が行われる。

4 評価

3 章において提案した手法に対して、ルールを用いた手法と比較して評価を行った。さらに、3 章において提案した手法から品詞情報を取り除いたモデルに関しても評価を行った。

[2] アクセントの推定も同様に未知語モデルによって推定可能であると考えられ、さらなる精度の向上が期待出来る。

[3] 幾つかの例において、アラインメントに複数の候補が存在する可能性がある。

4.1 コーパス

実験に用いたのは、新聞記事・テレビニュースの書き起こし・電話応答文など、雑多な内容を含むコーパスである。したがって、含まれる語彙も多岐にわたり、書き言葉だけでなく、話し言葉も含まれる。各文は、予め、人手によって形態素列に分割されており、各形態素 w には、品詞 t・読み $\boldsymbol{s}$・アクセント $\boldsymbol{a}$ が付与されている。ここで読みは音素アルファベット s の列であり、アクセントは各音素アルファベットに対するアクセント要素 $a = \{\ \mathtt{H}\ ,\ \mathtt{L}\ \}$ の列である。1 文の長さ平均は 21.6 語で、各語の平均文字列長は 1.91 である。学習用のコーパスは 8,800 文で、テスト用のコーパスは 150 文である（表 4）。

表 4: コーパスのサイズ

	文数	形態素数	文字数
学習	8,800	190,318	285,082
テスト	150	2,130	3,170

4.2 比較

本提案手法の有効性を調べるため、ルールを用いた手法との比較を行った。

WTS+A　確率モデル＋ルールによるアクセント付与

1. 文を表層・品詞・読みの三つ組 $\langle w, t, \boldsymbol{s}\rangle$ の列とみなし、n-gram モデルを用いて、単語境界・品詞及び読みを推定する。
2. アクセント句境界を予め作成しておいたルール（約 1,000 ルール）を用いて、順次適用することによって決定する。アクセント句境界を決定するためのルールには主に、品詞が用いられる。
3. アクセント句内でアクセント核を 2.2 節で説明した方法に基づき決定し、各単語のアクセントを決定する。

WTSA　確率モデル

- 文を表層・品詞・読み・アクセントの四つ組 $\langle w, t, \boldsymbol{s}, \boldsymbol{a}\rangle$ の列とみなし、n-gram モデルを用いて、単語境界・品詞・読み及びアクセントを同時に推定する。

WSA　確率モデル（品詞なし）

- 文を表層・読み・アクセントの三つ組 $\langle w, \boldsymbol{s}, \boldsymbol{a}\rangle$ の列とみなし、n-gram モデルを用いて、単語境界・読み及びアクセントを同時に推定する。

今回用いた 3 つのモデルでは、いずれも読みに関しては n-gram モデルによって推定する。**WTSA** と **WSA** の違いは品詞情報を用いるか否かという点のみである。つまり、**WTSA** において品詞を 1 種類に固定したものと同じである。これはコーパスを作る際に、品詞情報が必要かどうか調べるためである。

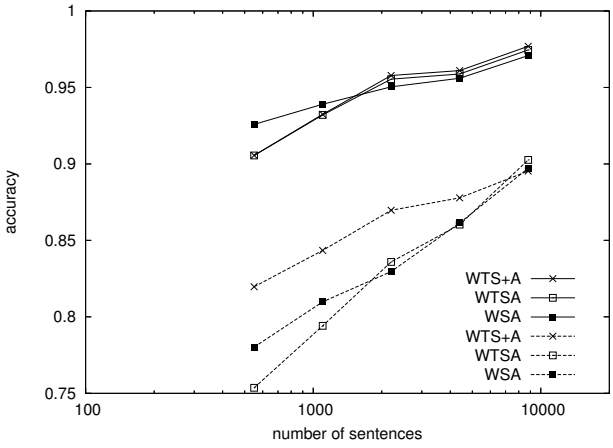

図 1: 読みとアクセントに関する学習曲線。実線は読み $\langle w, \boldsymbol{s}\rangle$ の精度。破線は読み＋アクセントの精度 $\langle w, \boldsymbol{s}, \boldsymbol{a}\rangle$。

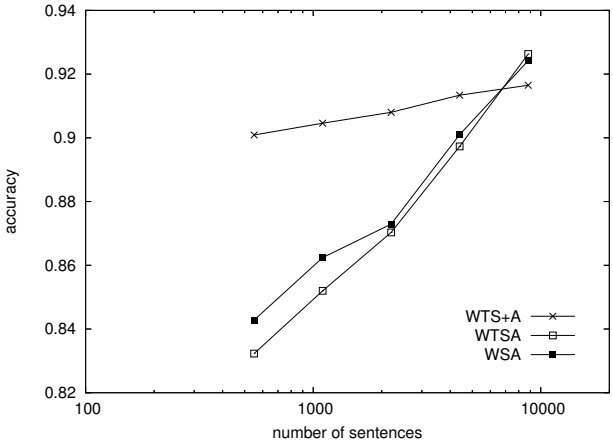

図 2: アクセントに関する学習曲線。実線は近似されたアクセントの精度 $\langle w, \boldsymbol{s}, \boldsymbol{a}\rangle|\langle w, \boldsymbol{s}\rangle$。

4.3 評価

表 5 は各モデルに対する読みとアクセントの精度を示してある。また、参考として、単語分割の精度と、品詞付与の精度も記す。表によると **WTSA** の精度が 3 モデルの中で最も高く、 90.26% となっている。**WTSA** から品詞を取り除いた **WSA** の精度は 89.72% であり、これは **WTSA** に比べて 0.54% 低い。**WTSA** と **WSA** の比較から、品詞情報は精度向上に寄与することがわかるが、その差はそれほど大きくない。また、この確率モデルのみを用いた両モデルのいずれも、ルールを用いた **WTS+**A よりも精度が高い。また、表 5 の一番右のカラムはアクセント付与単体の精度を示してある。読みとアクセントは同時に決定するため、アクセント付与単体の精度を正確に求めることは出来ないが、およその値として、読みが正解と一致した形態素に対してアクセントが一致した確率 $accuracy(\langle w, \boldsymbol{s}, \boldsymbol{a}\rangle|\langle w, \boldsymbol{s}\rangle)$ で近似してある。その結果、**WTSA** の精度が最も高く、 92.63% となっている。

図 1 及び図 2 は、学習コーパスのサイズと精度の関係を

表 5: モデル毎の精度（単語境界・品詞付与・読み付与・アクセント付与）

		異なり語数	単語境界 $\langle w \rangle$	単語境界 & 品詞 $\langle w, t \rangle$	単語境界 & 読み $\langle w, \boldsymbol{s} \rangle$	**単語境界 & 読み&アクセント** $\langle w, \boldsymbol{s}, \boldsymbol{a} \rangle$	アクセント $\langle \boldsymbol{a} \rangle \sim \langle w, \boldsymbol{s}, \boldsymbol{a} \rangle \vert \langle w, \boldsymbol{s} \rangle$
WTS+A	$\langle w, t, \boldsymbol{s} \rangle$	15,723	*97.61*	*96.08%*	97.69%	**89.53%**	91.65%
WTSA	$\langle w, t, \boldsymbol{s}, \boldsymbol{a} \rangle$	21,164	*97.87*	*95.79%*	97.45%	**90.26%**	92.63%
WSA	$\langle w, \boldsymbol{s}, \boldsymbol{a} \rangle$	19,560	*97.64*	*N/A*	97.08%	**89.72%**	92.42%

示してある。図 1 の各線の最も右の点は、表 5 の左から 5 番目の列 $\langle w, \boldsymbol{s} \rangle$ と同 6 番目の列 $\langle w, \boldsymbol{s}, \boldsymbol{a} \rangle$ と一致する。図 2 の各線の最も右の点は、表 2 の一番右の列 $\langle w, \boldsymbol{s}, \boldsymbol{a} \rangle | \langle w, \boldsymbol{s} \rangle$ と一致する。読み及びアクセントの精度 $\langle w, \boldsymbol{s}, \boldsymbol{a} \rangle$ に関する **WTSA** および **WSA** の学習曲線に注目すると、 8,000 文あたりで、 **WTS+A** の精度を超える。学習曲線の傾きから推測すると、8,000 文以降も、コーパスのサイズを増やせば増やすほど、確率モデルによる手法がルールに基づく手法との差を大きくすることが予想出来る。また、図 2 によると、**WTSA** と **WSA** の精度は 1.0 に近づきつつある。これはコーパスの量が増えるにつれ、読み＋アクセントの精度が、読みの精度に近づいていることを示す。それに対して、 **WTS+A** は徐々には上がっているが、前 2 モデルと比較すると平坦に近い。

品詞に関しては、品詞を用いた場合の **WTSA** に比べ **WSA** の精度は、読みに関しては 0.37% 、アクセントに関しては 0.54% 精度が低い。実用上、この差を大きいと見るか小さいと見るかは難しいが、コーパスを用意する際、品詞情報が与えられないとしても、少し大きめのコーパスを用意してやることで、品詞を付与したときと同じ精度を得ることが期待出来る。

5 おわりに

本論文では、確率的な手法を用い、入力テキストに対し、読み仮名及びアクセントを付与する手法について、述べた。このモデルでは、単語境界・品詞・読み・アクセントの四つ組を 1 つの単位と捉え、 n-gram モデルを用いて推定を行う。言い換えると、単語境界・品詞・読み・アクセントの 4 つの素性を同時に推定する。実験の結果、確率モデルに基づく手法が、ルールを用いた手法を上回る精度を獲得した。また、品詞を用いない三つ組（単語境界・読み・アクセント）を単位として n-gram モデルを生成し実験を行ったが、四つ組みに対して精度は低かった。ただし、三つ組でもコーパスを補えば、四つ組と同程度の精度が得られる。

基本的な韻律情報及び音韻情報である、読み仮名及びアクセントの付与に関しては 1 つの枠組みで学習できる可能性が高いことを示した。このことは、コーパスのみ与えらることができれば、日本語の標準語のみでなく、方言や、他言語でも同じ枠組みで読み及びアクセントが付与出来ると考える。

参考文献

[1] Pan, S. and Hirschberg, J., “Modeling local context for pitch accent prediction,” Proceedings of ACL, pp. 233-240, 2000.

[2] Shi, Q. and Fischer, V., “A comparison of statistical methods and features for the prediction of prosody prosodic structures,” Proceedings of ICSLP, ThA1404p, 2004.

[3] 匂坂., 佐藤., “日本語単語連鎖のアクセント規則,” 電子情報通信学会 技術研究報告, Vol. J66-D, No. 7, 1983.

[4] Beckman, M. and Pierrehumbert, J., “Japanese prosodic phrasing and intonation synthesis,” Proceedings of ACL, P86-1025, 1986.

[5] Klein, E., “A constraint-based approach to English prosodic constituents,” Proceedings of ACL, pp 217-224, 2000.

[6] Marsi, E., et al, “Learning to predict pitch accents and prosodic boundaries in Dutch,” Proceedings of ACL, pp 489-496, 2003.

[7] Nagata, M., “A stochastic Japanese morphological analyzer using a Forward-DP Backward-A* N-Best search algorithm,” Proceedings of Coling, pp 201-207, 1994.

[8] Cormen, T., Leiserson, C., and Rivest, R., “Introduction to algorithms,” The MIT Press, 1990.

[9] Jelinek, F., “Self-organized language modeling for speech recognition,” Technical report, IBM T. J. Watson Research Center, 1985.