確率モデルによる自由発話の形態素解析
A Stochastic Morphological Analyzer for Spontaneously Spoken Languages

永田昌明
Masaaki NAGATA
NTT 情報通信研究所
NTT Information and Communication Systems Laboratories

Abstract: We present a morphological analysis method for the phonetic transcription of spontaneous speech using a stochastic language modeling technique and an efficient two-pass N-best search strategy. It can segment a phonetically transcribed utterance into word, assign parts of speech to each segmented word, and convert the phonetic transcription into an orthographic transcription, which, in the case of Japanese, means the conversion from "hiragana" (phonogram) to "kanji" (ideogram). The morphological analyzer can handle pauses, interjections, restatements and chimings, all which are characteristics of spontaneous speech, by learning the parameters of the language model directly from the phonetic transcription. The proposed morphological analyzer achieves 95.0% recall and 95.3% precision on closed text when it was trained and tested on a portion (containing 172,826 words) of the ATR Corpus, telephone dialogues in the conference registration domain.

1 はじめに

自由発話 (spontaneous speech) とは何かを演繹的に定義することは難しい。従って、知識工学的な手法 (knowledge-based approach) を用いて自由発話を扱う音声言語システムを構築することは困難である。そこで、本論文では、統計的な手法で求めた確率的言語モデルを用いて自由発話の形態素解析を行なう方法を提案する。

既に、確率的言語モデルは英語の形態素解析において広く用いられている [2,3]。これらの確率的形態素解析法は単語を分かち書きする (単語と単語の間に空白を入れる) 英語の正書法を前提としている。しかし、音声言語では句境界におけるポーズやイントネーション変化の他に音響的な区切りがないので、確率的形態素解析と音声認識 (音韻照合) を音声言語システムの中で密結合することは難しかった。

永田 [7] は、最近、確率的言語モデルを用いて、単語の間に区切り記号 (空白) がない入力文を形態素解析する方法を提案した。本来、この方法は分かち書きの習慣がない日本語の書き言葉の形態素解析用に開発されたものだが、本報告では、自由発話を発音表記で文字化したテキスト (phonetic transcription) から言語モデルのパラメタを学習することによって、この方法で話し言葉も形態素解析できることを示す。

2 対話コーパスの文字化

我々は、ATR 対話データベースを対象として実験を行なった。ATR 対話データベースは、会議予約および旅行相談における電話およびキーボードを介した模擬対話を収録したものである [4]。

図 1に ATR コーパスに収録された発話の例を示す。図 1の最上段のように、ATR コーパスでは発話はかな漢字混じり表記で文字化され、ポーズは長さに応じて、読点 (、)、句点 (。)、またはリーダー (. . .) で表現している。また、間投詞は [] で囲まれ、言い淀みは () で囲まれている。さらに、発話は単語に分割され、各単語には品詞と読みが与えられている。

自由発話の音声認識結果をシミュレーションす

- かな漢字混じり表記 (間投詞や言い淀みを括弧で囲む):
  ［えーっと］今手元に［あの］登録用紙があるんですけれども、［えーっと］（その）その中でちょっと［あの］クレジットカードをね、［あのー］クレジットカードの名前となんかナンバーを書く所があるんですが、
- ひらがな表記 (間投詞や言い淀みを括弧で囲まない):
  えーっといまてもとにあのとうろくようしがあるんですけれども、えーっとそのそのなかで ...
- ローマ字表記 (間投詞や言い淀みを括弧で囲まない):
  eettoimatemotonianotourokuyoushigaarundesukeredomo,eettosonosononakade ...

図 1: ATR 対話データベースの例 (正書法表記と発音表記)

るために、我々は、図1の二段目の例ように、ATR コーパスの発話を全てひらがなで表記し、間投詞や言い淀みを示す括弧を取り除いたテキストを作成した。括弧の代わりに、言い淀みの最後に編集信号 (edit signal) に相当する記号を挿入する表記法も考えられるが、我々は、音響・韻律的手段で編集信号を検出することは現状では難しいと判断した。ただし、ポーズを表す記号はそのまま表記に残した。これは、ポーズは音響的手段で比較的容易に検出できると判断したためである。

どのような形式を自由発話の形態素解析の入力として仮定するかは議論の余地がある。音声認識出力をより現実的にシミュレートするためには、図1の最下段のように、ローマ字または音素記号で発話を表記し、confusion matrix などを使って音声認識誤りをシミュレートすべきだろう。これらは今後の課題としたい。

我々の形態素解析法は、音素記号の系列を正書法に従って文字化する一般的な方法を提供するので、以後では、ひらがな表記のテキストを発音表記テキスト (phonetic transcription)、かな漢字混じり表記のテキストを正書法表記テキスト (orthographic transcription) と呼ぶことにする。

3 統計的言語モデル

我々は、品詞付けモデル (tagging model) として品詞三つ組モデル (tri-POS model)、または、二次隠れマルコフモデル (2nd-order HMM) と呼ばれるものを用いる。

入力文が単語列 $W = w_1 w_2 \ldots w_n$ および品詞列 $T = t_1 t_2 \ldots t_n$ から構成されるとすれば、形態素解析は、単語列と品詞列の同時確率 $P(W,T)$ を最大化する単語分割と品詞付与の組を求めるという数学的問題に帰着される。品詞三つ組モデルでは、単語列と品詞列の同時確率 $P(W,T)$ を品詞三つ組確率 $P(t_i|t_{i-2},t_{i-1})$ と品詞別単語出現確率 $P(w_i|t_i)$ の積で近似する[1]。

$$P(W,T) = \prod_{i=1}^{n} P(t_i|t_{i-2},t_{i-1})P(w_i|t_i) \qquad (2)$$

もし品詞タグ付きコーパスがあれば、パラメタ $P(t_i|t_{i-2},t_{i-1})$ と $P(w_i|t_i)$ は、対象とする事象の頻度から求めることができる。

我々は、ATR コーパスから作成した間投詞や言い淀みを含む発音表記テキスト上で言語モデルの学習を行なった。間投詞や言い淀みは、単語列の中の一種のきょう雑物と考えられので、その直前と直後の単語の間の依存関係が品詞三つ組モデルによって捉えられる可能性は高い。

4 N-Best 探索アルゴリズム

ここで用いる形態素解析法は「前向き DP 後向き A^* アルゴリズム」と呼ばれる。これは、動的計画法 (Dynamic Programming) を用いた前向き探索と A^* アルゴリズムを用いた後向き探索の2つのパスから構成される。以下では、このアルゴリズムの概要を説明する。詳細は [7] を参照して頂きたい。

[1]実際には, 文頭と文末を表す特別な記号 "#" を考え、次式を使用する。

$$P(W,T) = P(t_1|\#)P(w_1|t_1)P(t_2|\#,t_1)P(w_2|t_2) \prod_{i=3}^{n} P(t_i|t_{i-2},t_{i-1})P(w_i|t_i)P(\#|t_{n-1},t_n) \qquad (1)$$

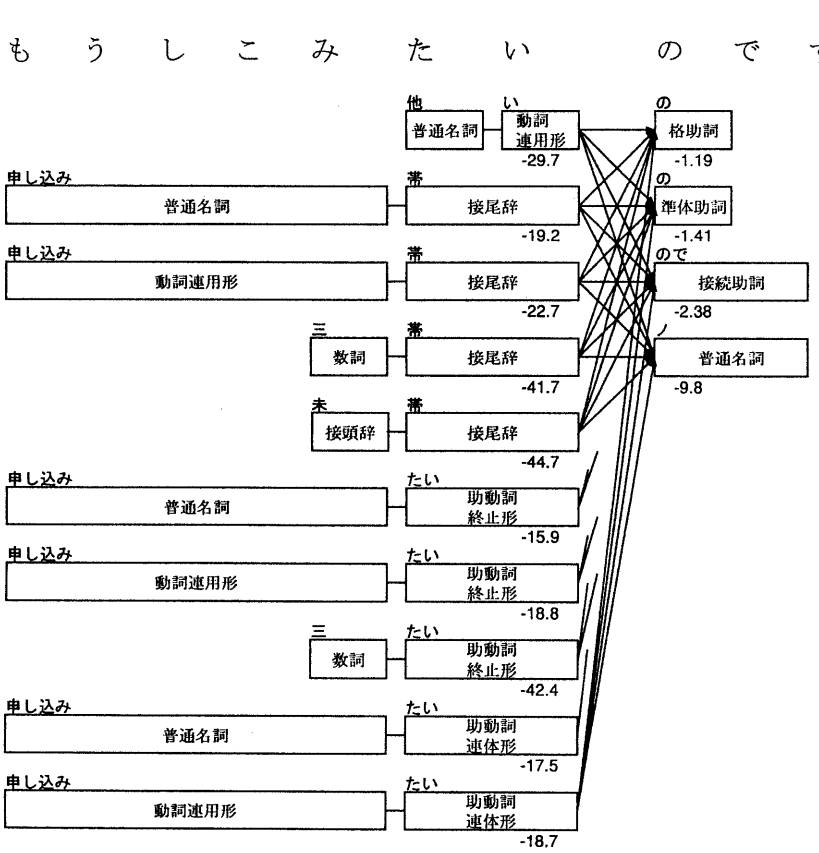

図 2: 前向き DP 探索

4.1 前向き DP 探索

前向き DP 探索は、文頭から一文字ずつ文末方向へ進む動的計画法を用いて、解析候補の確率を計算する。各文字位置では、その文字位置で終る最適な部分解析と、その文字位置から始まる単語仮説の全ての組合せが調べられる。もし部分解析と単語仮説の接続が品詞付けモデルによって許されるならば、新しい部分解析が作られ最適部分解析テーブルに記憶される。新しい部分解析の確率は、元の部分解析の確率と品詞三つ組確率と品詞別単語出力確率の積である。

図2は、“もうしこみたい”の終りの文字位置における前向き探索の様子を示している。この例では、最後の単語の区切りと表記、および、最後の二つの単語の品詞が異なる部分解析が 10 個あり、“のですが”の最左部分文字列に照合する辞書項目が 4 個ある。全ての組合せを調べ、最後の単語 (の区切りと表記) と最後の二つの品詞が同じ部分解析の中で最大の確率を持つ部分解析が最後の単語の終了位置の最適部分解析テーブルに記憶される。

4.2 後向き A^* 探索

後向き A^* 探索は、文末から文頭方向へ進む A^* アルゴリズムを用いて、解析候補を確率が大きい順に一つずつ取り出す。後向き探索では、前向き探索で最適部分解析テーブルに記憶された部分解析を A^* アルゴリズムの状態と考える。最後の単語の区切りと表記が同じで、最後の二つの単語の品詞が同じ時、二つの状態 (部分解析) は等しい。

初期状態は、文末における最適部分解析テーブルの項目から得られる。次状態は、まず、現在の状態に対応する部分解析の最後の単語の開始位置における最適部分解析テーブルの項目を検索し、次に、この遷移が品詞付けモデルによって許

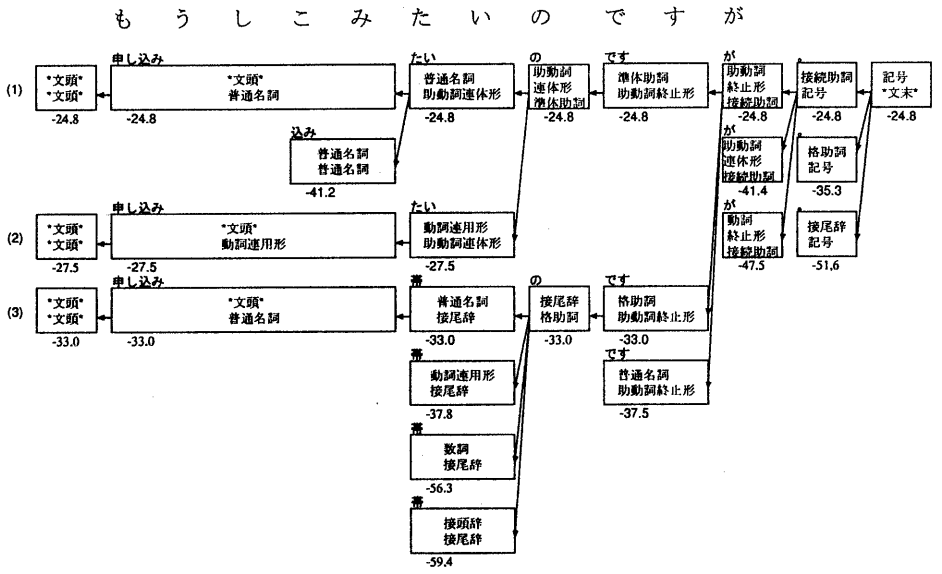

図 3: 後向き A^* 探索

され、かつ、現在の状態の最初の品詞と次状態候補の最後の品詞が等しいか (2 次 HMM の制約) をチェックすることにより得られる。

後向き探索の部分経路は全経路のコスト (形態素解析候補の対数確率の絶対値) に基づいて展開される。全経路のコストは、後向き探索の部分経路のコストと、前向き探索で記憶された残りの部分の最適経路のコスト (最適部分解析の対数確率の絶対値) の和として求められる。残りの部分の最適経路のコストが分かっているので、後向き探索は admissible である。すなわち、正確に上位 N 個の形態素解析候補が求まる。

図3に、“もうしこみたいのですが” という文の後向き探索の様子を示す。この図の各箱は、文頭からある単語までの最適部分解析に対応する。箱の中の一行目は一つ前の単語の品詞、二行目はその単語の品詞を表す。箱の下の数字は全経路の対数確率である。この図より、後向き探索の計算量は、前向き探索の計算量よりずっと少ないことが分かる。

5 評価尺度

ここでは、日本語形態素解析の評価尺度 [7] を拡張して、発音表記テキストの形態素解析の評価尺度を以下のように定義する。

まず、入力文に対する形態素解析候補を、各単語に対する区切り・品詞・表記の三つ組 (3-tuple) の集合と考え、次に、正解データに含まれる三つ組の集合と、システム出力に含まれる三つ組の集合を比較する。N-Best 候補の精度の計算では、上位 N 個の候補に含まれる三つ組の集合和を計算し、これを正解データに含まれる三つ組の集合と比較する。

三つ組の集合の比較は、まず、正解データに含まれる三つ組の数 (Std)、システム出力に含まれる三つ組の数 (Sys)、および、照合した三つ組の数 (M) を数え、次に、評価尺度として再現率 (recall = M/Std) および適合率 (precision = M/Sys) を計算する。

我々は、二つの三つ組 (形態素候補) を照合する際の基準として、次の三つの等価性を定義する。

- 音素的に等価: 単語の区切りが同じ。
- 文法的に等価: 単語の区切り・品詞が同じ。
- 正書法的に等価: 単語の区切り・品詞・表記が同じ。

発音表記テキストの形態素解析の例を図 4 に示す。入力発話は、間投詞 (“えーっと”, “あのー”)、

```
> (morph-n-best "えーっとあのー、かいぎにですね、もうしこみたいんですけど。")
-56.10057291587962  ; 第一候補の対数確率
えーっと / えーっと / 間投詞 あのー / あのー / 間投詞 、/、 / 記号 会議 / かいぎ / 普通名詞
 に / に / 格助詞 です / です / 助動詞・終止 ね / ね / 終助詞 、/、 / 記号
 申し込み / もうしこみ / 普通名詞 たい / たい / 助動詞・連体 ん / ん / 準体助詞
 です / です / 助動詞・終止 けど / けど / 接続助詞 。/。 / 記号
-59.03105372275291  ; 第二候補の対数確率
えーっと / えーっと / 間投詞 あのー / あのー / 間投詞 、/、 / 記号 会議 / かいぎ / 普通名詞
 に / に / 格助詞 です / です / 助動詞・終止 ね / ね / 終助詞 、/、 / 記号
 申し込み / もうしこみ / 本動詞・連用・五段 たい / たい / 助動詞・連体 ん / ん / 準体助詞
 です / です / 助動詞・終止 けど / けど / 接続助詞 。/。 / 記号
```

図 4: 形態素解析の例

間投詞的な句(“ですね”)、終助詞的な句(“んですけど”)、ポーズ(“、”, “。”)を含んでいる。第一候補は“もうしこみ”の品詞が間違っており、第二候補が正解である。従って、第一候補は、音素的等価性に関して再現率 14/14、適合率 14/14 であり、文法的等価性(および正書法的等価性)に関して再現率 13/14、適合率 13/14 である。また、上位 2 候補の文法的等価性に関する再現率と適合率は、それぞれ 14/14 と 14/15 である。

6 実験

表 1: 訓練データと試験データの量

	訓練データ		試験データ (closed)	
	括弧あり	括弧なし	括弧あり	括弧なし
文	9276	9276	1000	1000
単語	172826	151380	18155	15949
文字	1653060	1524382	174203	160967

我々は形態素解析法の学習とテストのために ATR 対話データベース [4] を使用した。ATR コーパスは全体で約 80 万語あり、文字化・単語分割・品詞付与は人手で行なわれている。この実験では、ATR コーパスの約 1/4 に相当する、会議予約に関する電話会話の部分(約 20 万語)を用いた。まず、将来のオープンテストのために無作為に 1,000 文を選び、残りを訓練データとする。次に、訓練データの中から無作為に 1,000 文を選び、クローズドテストに用いる。訓練データと試験データに含まれる文・単語・文字の数を、間投詞や言い直しを示す括弧を含む場合と含まない場合について、表 1に示す。

区切り・品詞・表記が正しく求まった単語の割合を図5 に示す。本形態素解析法が単語の区切り・品詞・表記を全て正しく求める割合、すなわち正書法的等価性に関する再現率と適合率は、第一候補で 95.0% および 95.3%、上位五候補で 97.1% および 73.% であった。区切りと品詞の正しさ(文法的等価性)だけに着目すれば、第一候補の再現率と適合率は約 2% 高くなり、区切り(音素的等価性)だけに着目すれば、さらに 4% 高くなる。

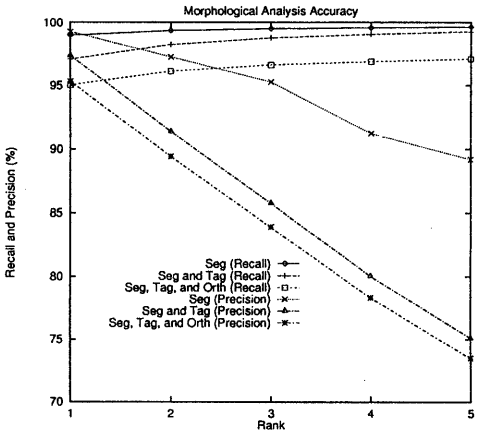

図 5: 区切り・品詞・表記が正しい単語の割合

7 考察

自由発話を対象とした音声言語システムでは、disfluency の扱いが大きな問題となる。近年、言い直し(repair, self-correction)を検出する研究、すなわち、言い直しにおいて、発話が中断する場

所 (interruption site)、言い直される区間 (reparandum interval)、言い直しを発話するまでの時間 (disfluency interval)、言い直しの区間 (repair interval) を同定する研究が幾つか行なわれている [1,5,8]。これらは、自由発話の局所的な構文解析と考えられ、システムが発話の意味を正確に把握するためには必須の技術である。

しかし、Bear ら [1] や Kikui ら [5] の方法は、音声から正書法表記テキストが得られることが前提になっている (text-first approach)。これに対して、Nakatani ら [8] は、音響・韻律的特徴 (例えば、編集信号が音声信号の突然の中断として現れる) のみから言い直しを検出しようとする (speech-first approach)。

本報告で述べた形態素解析法は、このような自由発話に特有の言語現象の解析技術と、自由発話の音声認識技術の間の「架け橋」となる。すなわち、発音表記テキストという音声認識出力に比較的近いものを入力として、単語の区切り・品詞・(正書法) 表記を出力する。本手法は編集信号を仮定していないが、編集信号が検出できれば更に精度が向上するだろう。

未知語の扱いは、今後の最大の課題である。自由発話の音声認識を想定した場合、(1) 辞書未登録語、(2) 言い誤りや言い淀み、(3) 音声認識誤り、などが未知語の原因となる。我々は、書き言葉における辞書未登録語に対しては、文字三つ組に基づく単語モデル (word model) を用いて単語の区切りと確率を推定する方法を用いた [7]。日本語の書き言葉の場合、未知語の大部分は名詞であり、表意文字である漢字で表記され、長さが比較的短いので、文字三つ組を使う方法は有効であった。しかし、話し言葉の場合には、言い誤りや言い淀みの量が無視できず、しかもこれらはオープンカテゴリなので、言い誤りや言い淀みの区切りと確率を推定する語片モデル (word fragment model) を考える必要がある。そして、本形態素解析法のオープンテキストに対する性能を評価しなければならない。

本報告の形態素解析システムは、自由発話の音声認識の中へ組み込むことができるはずである。ATR の HMM-LR 音声認識システム [6] は、統語的制約が音声認識精度の向上に非常に有効であることを示したが、このシステムでは予測型 LR パーザのための文脈自由文法規則を人手で記述しなければならない。これに対して、我々の確率的形態素解析システムは言語モデルをコーパスから自動的に学習する。これは、文法規則を記述すること自体が難しい自由発話を扱うのに適した特徴である。

中国語も日本語と同様に分かち書きしない。最近、WFST(Weighted Finite-State Transducer: 重み付き有限状態変換器) という確率モデルを用いる中国語の形態素解析法が Sproat らによって提案された [9]。WFST も HMM も正規文法に近いクラスの確率的言語モデルであるから、Sproat らの手法と本手法は潜在的には同程度の能力を持っていると思われる。両者の比較については今後の課題としたい。

8 Conclusion

本報告では、自由発話の発音表記テキストを確率的言語モデルに基づいて形態素解析する方法について述べた。この方法は、タグ付けされた自由発話のコーパスからパラメタを学習することにより、ポーズ・間投詞・言い淀みなどを含む発話を扱うことができる。このアルゴリズムは言語に依存しない一般的な方法であり、自由発話を扱う音声言語システムへの応用が期待できる。

参考文献

[1] J. Bear, J. Dowding, E. Shriberg, "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog", ACL-92, pp.56-63.

[2] K. Church, "A Stochastic Part of Speech Tagger and Noun Phrase Parser for English", Proc. ANLP-88, pp.136-143, 1988.

[3] D. Cutting, et al., "A Practical Part-of-Speech Tagger", Proc. ANLP-92, pp.133-140, 1992.

[4] T. Ehara, et al., "ATR Dialogue Database," Proc. ICSLP-90, pp.1093-1096, 1990.

[5] G. Kikui and T. Morimoto, "Similarity-Based Identification of Repairs in Japanese Spoken Language", ICSLP-94, pp.915-918, 1994.

[6] K. Kita, et al., "HMM continuous speech recognition using predictive LR parsing", Proc. ICASSP-89, pp.703-706, 1989.

[7] M. Nagata, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm", COLING-94, pp.201-207, 1994.

[8] C. Nakatani and J. Hirschberg, "A Speech-First Model for Repair Detection and Correction", ACL-93, pp.46-53, 1993.

[9] Sproat, R. et al.: "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", ACL-94, pp.66-73, 1994.