自然言語処理 116－9
（1996. 11. 18）

人の発声単位を考慮した日本語言語モデルの検討 — 日本語における単語とは

伊東 伸泰, 西村 雅史, 荻野 紫穂, 山崎一孝
日本アイ・ビー・エム 東京基礎研究所
〒 242 神奈川県大和市下鶴間 1623-14
e-mail: iton@trl.ibm.co.jp

日本語では単語の境界があいまいで、文法的に定義された単位は必ずしも人が認知している単語単位と一致しない。本研究では人の発声単位を考慮した単語単位の構成方法とそれに基づいた日本語言語モデルを提案する。本手法では人が単語境界と考える点で分割した比較的少量のテキストデータと形態素解析による分割結果とを照合することにより、人が潜在意識としてもつ単語単位を形態素レベルのパラメータでモデル化した。そして多量のテキストを同モデルにしたがった疑似乱数で分割することにより、単語単位のセットと言語モデルの訓練用データを自動生成した。

A Japanese Language Model Based on Human Utterance Units — word in Japanese

Nobuyasu ITOH, Masafumi NISHIMURA, Shiho OGINO, and Kazutaka YAMASAKI
Tokyo Research Laboratory, IBM Japan Ltd.
1623-14, Shimotsuruma Yamato-shi Kanagawa, 242 Japan

Abstract

This article deals with a method for segmenting a text into words on the basis of human utterance units. In Japanese, word boundaries are not stable and grammatical units do not necessarily coincide with human intuition. For accurate segmentation it is therefore necessary to create a vocabulary set that covers human utterance units. In our method, a model of word boundary is decsribed by morphological parameters (i.e. part of speech), which are learned by comparing results of human segmentation with those of Japanese morphological analyzer. Then by using pseudo-random number and the model, it is determined whether each morpheme transition is a word boundary. As a result, we obtain a vocabulary set and learning data for Japanese language model automatically.

1 はじめに

音声認識技術はその発達にともなって、その適用分野を広げ、日本語においても新聞など一般の文章を認識対象とした研究が行なわれるようになった [1][7]。日本語をはじめとして単語の概念が明確ではない言語における音声認識を実現する場合、どのような単位を発声単位および認識単位として採用するかが大きな問題の1つとなる。この問題はユーザーの発声単位に制約を課す離散発声の認識システムの場合に限らず、連続音声の認識においても、ユーザーが適時ポーズを置くことを許容しなければならないため、やはり発声単位を考慮して認識単位を決める必要がある。従来発声単位としては孤立単語 [3]、文節区切り発声 [4] などが試みられているが、前者は人間にとってきわめて困難な分割を強いるものであるし、後者も安定なものとは言えない [5][6]。このように発声単位と認識単位を決めるさいの条件としては以下のことが考えられる。

- 認識単位は発声単位よりも細かい単位でなければならない。日本語では形態素解析の結果得られるトークンを認識単位とすることが比較的容易な方法であるが、形態素として慣用句のような長い単位も採用している場合は、この要請が満たされているとは限らない。
- 長い認識単位を採用する方が、音響上の識別能力という観点からは望ましい。つまり連続して発声される可能性が高い部分については、それ自身を認識単位としてもっておく方がよい。
- 言語モデルを構築するためには、多量のテキストを認識単位に分割する必要があり、処理の多くが自動化できなければ実用的ではない。

これらは、言い換えれば人間が発声のさいに分割する (可能性がある) 単位の Minimum Cover Set を求めることに帰着する。西村 [6] は、人の発声単位をルールベースにより求め、認識単位とする手法を提案しているが、その中に提示されている方法だけでは分割の揺れに対応することが困難である。またきわめて単純な方法としては、多量のテキストを多くの人手で分割し、頻度が高いトークンを採用することが考えられるが、上記の要請に反し、現実的ではないことは明らかである。

2 発声単位に基づく分割

日本語を分割して発声する場合、その分割点はきわめて安定している点と、人、または時によって分割されたりされなかったりする不安定な点がある。例として「私は音声認識器のテストを行なっています。」という文章を考えよう。これは形態素解析により、たとえば

私は ＋ は ＋ 音声 ＋ 認識 ＋ 器 ＋ の ＋ テスト ＋ を ＋ 行な ＋ っ ＋ て ＋ い ＋ ます ＋ 。

と分割されるが、動詞の活用語尾である「っ」や接続助詞の「て」はほぼ確実に「行」と結合して「行なって」と発声されるのに対し、接辞である「器」は分割して発声する場合もあれば、結合されることもあるだろう。そこで文章がある位置で「分割」される確率を形態素のレベルでモデル化することを考える。そして人が分割した学習用テキストと同じテキストを形態素解析により分割した結果を照合し、各形態素の遷移ごとに当該点で分割される確率を得る。その後、より大量のテキストをそのモデルに基づいて分割すれば (このプログラムを以後 Segment simulator と呼ぶ)、人が分割した傾向をもったわかち書きテキストを容易に得られる。

「分割」される位置としては、形態素の境界 (形態素単位への分割) とさらに細かく形態素の途中 (文字単位への分割) がある。ここで分割記号として$\sharp$を使用し、「分割」は記号「$\sharp$」が生起し、「結合」は「NULL」が生起すると考えれば、前者はある形態素から別の形態素に遷移したときにその間に「$\sharp$」が生起する確率として

$$P(\sharp_i \mid Morpheme_i \rightarrow Morpheme_{i+1})$$

となる。後者のそれは$Morpheme$を文字列$C_1C_2,\ldots,C_n$で表すと、そのj番目の文字の後に$\sharp$が生起する確率と考えれば

$$P(\sharp_j \mid Morpheme,\ C_j \rightarrow C_{j+1})$$

と表現できる。モデルのパラメータ (形態素の属性) としては、形態素の品詞 (Part-of-Speech: POS) および単語種別 (Kind of Word: KoW)、そして表記 ($String$) を採用し、$(KoW[POS], String)$ と表現する。ここで品詞、単語種別とはわれわれの用いた形態素解析プログラム [2] の出力として得られるものであり、品詞は 119、単語種別は 81 に分類されている。したがって形態素単位の分割では 6 個、文字単位への分割では 4 個のパラメータで記述されることになるが、そうすると明らかに多量の学習用テキスト (人が分割したもの) が必要となる。そこでデータが閾値以下であるような場合については、パラメータを特定の順序で縮退させた確率値を用意し Segment simulator の実行時も、確率が記述されているレベルまで同様の順序で縮退し、当該確率値で代用することを考える。この順序と参照される確率値を木構造で表現したのが図 1である。各ノードには形態素の属性とその属性が満たされた場合に分割される確率が対応する。たとえば図 1中

$$P(\sharp \mid V.infl.[29] \rightarrow Conj.\ p.p.[69],\ \text{て})$$

は形態素単位への分割に対する記述例で、形態素の属性が動詞活用語尾 [29] から接続助詞 [69]「て」へ遷移したときに、その間で分割される確率を意味する[1]。1 つ上のレベルでは、表記「て」が省略され、他の接続助詞でかつ品詞番号が 69 であるとき (i.e. 「で」) に分割された場合のカウントもマージした上で計算された確率 ($P(\sharp \mid V.infl. \rightarrow Conj.\ p.p.$) をもつ。ただし単語種別が名詞の場合には文字数が分割確率を記述するパラメータとして有効と考えられるので[2]、表記を省略した場合、文字数をパラメータとして残した。さらに上位レベルでは、品詞番号も省略し、単語種別 $V.\ infl.$ から $Conj.\ p.p.$ への遷移に対して、人が分割する確率を記述する。たとえば、「残して」という文節を形態素に分割すると

$$\text{残}\ (Verb[4])\ +\ \text{し}\ (V.infl[29])\ +\ \text{て}\ (Conj.\ p.p.[69])$$

となるが、その中に表れる「し」と「て」の間で分割されたカウント等もマージした上で算出された確率となる。このように木はリーフから上位のノードに行くにしたがって縮退されたパラメータ、言い換えればより大まかなパラメータとなる。

一方、前節で述べたように人は形態素として定義されたトークンをさらに文字単位で分割する場合もある。これは形態素解析の都合上慣用句を 1 つの形態素としてとり扱うことが行なわれるためである。そこで形態素レベルの分割よりもさらに詳細なレベルとして、文字レベルの分割をモデル化した[3]。

---

[1] $V.\ infl.$ は Verb inflection、 $Conj.\ p.p.$ は Conjunctive post-positional particle の略。

[2] 「誤認識」が「誤」「認識」と分割されるよりは「音声認識」が「音声」「認識」となりやすいなど。

[3] たとえば $P(\sharp \mid Adjective[17], \text{言うまでもな}, \text{も} \rightarrow \text{な})$ は形態素「言うまでもな」の 5 文字目と 6 文字目の間で分割され「言うまでも$\sharp$な」となる確率を意味する。

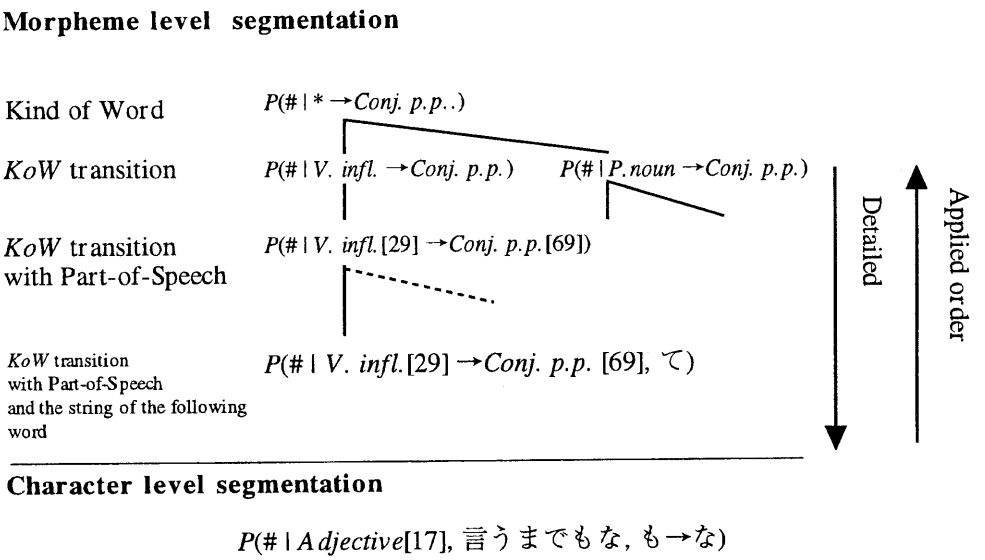

図 1: Segment Simulator におけるパラメータ縮退の順序

このような確率木はつぎのように構成することができる。つまりもっとも細かい分類における各パラメータについて、人が分割した結果と形態素解析の結果を照合してカウントし、その値をリーフから上位ノードに伝搬させた後、確率値に正規化すればよい。全カウント数が少ないと当該確率 (推定値) の信頼性が低いので、カウント、マージ作業を行なって、頻度がある閾値以上のノードを最終的なノードとして採用することにする。

このモデル化では学習データの量に応じて、そのデータから得られる情報を最大限に利用することができる。たとえば、2 文字漢語から接尾辞への遷移には、非常に多くのものがあるが、その分割されやすさは接尾辞の種類によって異り、それらを捨象してモデル化したのでは、あいまいさが大きくなってしまう。しかし逆にそのすべてを細分化したのでは、頻度が低い接尾辞に対するルールが得られないか、または信頼性の低い確率推定値となってしまう。本手法によれば学習データ中に頻度が高いものについてはより細かい分類でモデル化され、頻度が下るにしたがって統計として信頼にたる単位まで縮退されたパラメータによる確率値が得られることになる。

3 実験

本手法により、どの程度の数の (人が単語として考える) トークンが得られるであろうか。このことを検証するため 2 つの実験を行なった。

3.1 Closed data

最初に分割ルールを学習するためのデータを得るため、被験者4人により、新聞5ヵ月分および日本語用例集から選択した[4]計約26,000文を分割する作業を行なった。被験者には

- 不自然にならない限り、より細かく分割すること。
- 書かれた文章ではなく発声する場合の分割点を回答すること。

という指示を与えた。

その結果に基づいて、確率木を構成したところ、計2,829個のノードからなる木が得られた。表1に一例を示す。ただしノードとして採用するか否かの閾値は当該ノードの出現回数(カウント)が50以上のものとした。このモデルに基づいて以下のように多量の(形態素解析された)テキス

表1: 生成された木に記述された分割確率の例

パラメータ値	分割確率
$Noun$[19] → $Noun$[19], 者	0.33
$Noun$[19] → $Noun$[19], 人	0.71
$Noun$[19] → $Adj.$[18], 的	0.36
$V.infl.$[29] → $Conj.p.p.$[69]	0.03
$Noun$[19] → $P.p.$, を	1.0

トを分割する。

1. 各形態素およびその遷移について、品詞、単語種別、形態素の表記を得て、確率木のリーフに記述があるかどうかを調べる。
2. なければ、木作成の説明で述べた順にパラメータ値を縮退させ、確率木に記述があるかどうかを調べる。
   - 記述があれば、0から1の範囲の乱数を発生し、その値がノードに付随する確率以下であれば当該位置で分割し、そうでない場合は分割しない。
   - 記述がなければ、縮退を繰り返す。
3. もっとも上位のノードにも該当しない場合、形態素の分割点であれば当該位置で分割し、それ以外は分割しない。

新聞3ヵ月分(合計446,079文)に上記の手続きを適用して分割、連結を行なった。その結果を図2に示す。合計で約10^7個、のべ216,904種類のトークンが生成された。図はそれらを頻度の高いのものから順にとった場合のカバレージを示している。ただし数字表現、姓名はカウントから除いている。これによれば上位約25,000個(種類)のトークンで全トークンの約95%がカバーできることがわかる。

[4]選択は文の長さが一定の範囲に入っていることを除けば無作為に行なった。

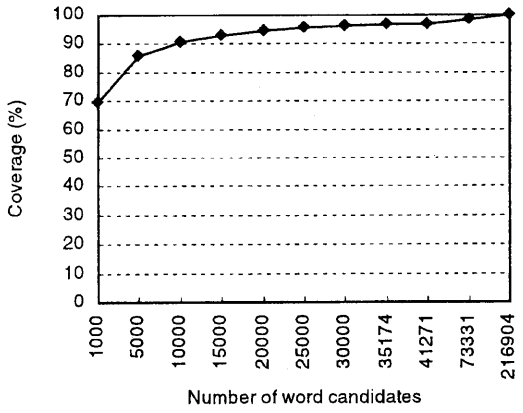

図 2: 新聞 3ヵ月のテキストに対するトークン数とカバレージ

3.2 Open data

前節の実験は、分野、データとも Closed なものであった。Open なテキストや分野について同様の考察を行なうため、さらに別のデータについて実験を行なった。まず語彙セットを決めるため、新聞 2 種類 (各 3ヵ月と 2ヵ月分)、EDR コーパス [8]、およびパソコン通信ピープルの電子会議室 (約 30 種類) に投稿されたテキストを用意し、上記と同様の手法を用いてトークンを生成した後、頻度の高いものから選択し、これらに対して 95%をカバーする語彙辞書を作成した[5]。電子会議室のテキストは他のテキスト群と著しくスタイルが異るので Segment simulator が用いる確率木も当該分野のテキスト (8,500 文を計 13 人で分割したもの) から新たに作成した。これに数字表現に必要な用語 (「一」「二百」など) を加えた結果 39,295 語の辞書が得られた。

この語彙セットと上記で分割されたテキストから単語の Trigram モデルを学習した。各データの文章数を表 2に示す。文章は全体の 95%を n-gram カウントに、残り 5%を Held-out 補間のパラメータ学習用に用いた。なお n-gram モデル作成には、乱数による分割処理 (Segment simulator) は必ずしも必要ではなく、形態素解析の結果と分割確率を使って直接各 n-gram の生起確率を推定することも可能である。

表 2: ソース別のテキスト数

日経新聞	415,071
サンケイ新聞	1,291,845
EDR コーパス	169,382
電子会議室	173,649

一方、新聞 2 種類、電子会議室、ビジネス上のスピーチ 2 種類、小説のテキストを別に用意し、

[5]ただし、電子会議室では製品名やハンドル名が頻出するが姓名、数字表現に加えこれらも統計から除いている。

被験者により分割を行なった。それらのテキストをテストデータとして上記の語彙セット、および言語モデル (Trigram) の効率を調べたものが表3である。

この結果、新聞や、ビジネストークにおいては Open Data に対しても 95% 程度のカバレージが得られ、かつパープレキシティも現在の音響識別能力で十分対応可能な値 (100-250 程度) になっていることがわかる。

Segment simulator が表3の日経新聞テキストを分割するさいに使用したルールのレベル別比率を表4に示す。表から明らかなように全体の約 60% の場合には、一番詳細なレベルのルールが適用されていることがわかる。

表 3: テストデータにおける単語カバレージとパープレキシティ(イタリックは姓名を除いた場合)

	文章数	単語数	カバレージ (%)	パープレキシティ
日経新聞	775	19.210	97.5	143.1
		19,063	*97.9*	
サンケイ新聞	600	18,495	95.8	253.0
		18,339	*96.2*	
小説	993	18,842	95.0	297.6
		18,757	*95.3*	
ビジネストーク	679	17,505	97.8	171.5
		17,433	*98.0*	
電子会議室	1,043	18,021	95.9	377.2
		17,881	*96.3*	

表 4: 適用されたノードの比率 (階層別)

パラメータ値	比率 (%)
$P(\sharp \mid KoW_1[POS_1] \rightarrow KoW_2[POS_2], String)$	59.6
$P(\sharp \mid KoW_1[POS_1] \rightarrow KoW_2[POS_2])$	29.2
$P(\sharp \mid KoW_1 \rightarrow KoW_2)$	3.9
$P(\sharp \mid KoW_2)$	6.6
該当なし	0.7

4 おわりに

このようにして得られたトークンは、日本語で人がある単位だと感覚的に思う単語候補を示していると考えられる。たとえば「行う」という動詞とその後続の付属語列からは

行い	行いたい	行う
行うべき	行え	行えば
行える	行った	行ったら

行って　　行っても

の計 11 トークンが生成された。また「たい」や「べき」といったトークンも生成されており、分割に揺れがある部分では複数の分割に対応したトークンが得られることがわかる。

このように、本手法は比較的少量の人による分割データから揺らぎを含めた分割傾向を推定し、それに対応したトークン を自動生成するものである。また学習された確率木をみると $P(\natural \mid Unknown$, んですね, す → ね) など形態素解析が対応できず未知語となった形態素に対するノードも 59 個生成されており、本手法は形態素解析の誤りにもある程度対応できることがわかった。

本研究にテキストデータ使用を許諾していただいた、産経新聞社、日本経済新聞社、そして (株) ピープルワールドカンパニーに感謝いたします。

参考文献

[1] 森 他: 新聞読み上げタスクを用いた大語彙連続音声認識における言語モデルの検討, 音響学会講演論文集, 3-8-7, (1996-3).

[2] 丸山, 荻野: 正規文法に基づく日本語形態素解析, 情報処理学会論文誌, Vol. 35, No. 7, pp. 1293-1299, (1994).

[3] 斗谷 他：音節をベースとした大語彙単語認識装置における認識方式, 音響学会講演論文集 2-3-10, (1988-10).

[4] 神谷 他: 日本語文節音声の認識, シャープ技報, Vol. 49, No. 6, pp. 23-26, (1991).

[5] 原田: 日本語テキストにおける認知的単位, 情処文書処理研究会, DPHI22-3, (1988).

[6] 西村, 大嶋, 野崎: 日本語 Dictation System のための統計的言語モデルに関する一考察, 情処 51 全国大会, 3R-7, (1995).

[7] 西村, 伊東: 離散単語発声による日本語ディクテーションシステムについて, 音響学会講演論文集 3-3-9, (1996-9).

[8] EDR 電子化辞書仕様説明書, (株) 日本電子化辞書研究所.