自然言語処理 122－9
（1997. 11. 20）

N-gramを用いた日本語テキストの単語単位への分割

伊東 伸泰　　西村 雅史
日本アイ・ビー・エム 東京基礎研究所

概 要

われわれは音声認識の統計的言語モデル構築を目的として、日本語テキストを人が潜在意識的にもつ単語単位で処理するアプローチを試みている。最初の確率推定の学習データは形態素解析した結果を人の単語分割モデルに基づいた乱数により分割・統合して作成した。本研究ではそうして得られた *N*-gram モデルを用いて、日本語テキストを直接分割する試みを行った。そのさい未知語を *N*-gram モデルにどう取り込むかという点が問題となる。本研究では既存の辞書から、文字種や出現する位置に基づいた文字クラスを作成し、各表記をその文字が属するクラスに置き換えたパターンによる未知語モデルを提案する。未知語に対する生起確率は、これらのパターンの出現確率および文字クラスから各文字への条件付き確率を用いて計算される。新聞やパソコン通信の電子会議室中の文章を用いた実験によれば、人が分割した結果と 92%が一致し、未知語抽出部分については再現率 40-50%、適合率 80-85%が得られた。

A Method for Segmenting Japanese Text into Words by Using *N*-gram Model

Nobuyasu ITOH and Masafumi NISHIMURA
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.

This paper presents a method for segmenting Japanese text into words that coincide with human intuition by using an *N*-gram model, focusing on how to handle unknown words. The initial set of learning data was created by morphological analysis and split-and-merge based on a human word-boundary model. Unknown word model is described with patterns of macro character (character classes) strings. Probabilities of unknown strings are calculated from those of the patterns and the conditional probability of each character in a class. According to our test, 92% of all words coincided with those segmented by human subjects. The rate of recall for unknown words is 40-50% and the precision is 80-85%.

1 はじめに

音声認識技術の発達にともない大語彙の連続音声認識が、日本語でも試みられるようになってきた[1][2]。この要因は、音素環境依存型HMMによる音声信号のモデル化に加え、多量の言語コーパスが入手可能になった結果、文の出現確率を単語N個組の生起確率から推定するN-gramモデルが実現できるようになったことが大きく貢献している。日本語などわかち書きされていない言語においてN-gramモデルを作成するには、テキストを所定の単位に分割することが基本的な作業の1つとなる。筆者らは、人が潜在的にもつ単語を認識単位とした日本語音声認識を提案し、その有効性を実験により確かめた[3]。そのさいには、少量のテキストについて形態素解析の結果と被験者が分割したものとを比較して品詞にもとづく分割モデルを作成し、そのモデル(分割確率)にもとづいた乱数を使用してコーパスの各文章を分割した[11]。一方形態素解析自身を、文法に基づくルールではなく統計的アプローチで行おうとする試みが報告されている。そこで単語単位への分割においてもベースとなる言語モデルのN-gram統計を用いてテキスト分割することが考えられる。その場合もっとも問題となるのは未知語でありこれをいかに精度よく抽出できるかが、全体としての精度に大きく影響する。これはわれわれの単語分割に限らず、自然言語処理のベースとなるテキストの分割処理では必ず遭遇する問題であり、N-gramによる形態素解析においても、未知語の取り扱いが重要であると指摘されている[8]。

2 N-gramを用いた日本語テキスト分割

N-gramによる単語分割モデルとは文字列$S = c_1c_2,\ldots,c_n$を単語列$W = w_1w_2,\ldots,w_m$に分割する基準として、「そのときの条件付き確率$Pr(W \mid S)$が最大となる」を採用したものである[9]。つまり$\mathbf{argmax}_W\ Pr(W \mid S)$である$W$を最適な分割と定義する。ここでベイズ則から

$$Pr(W \mid S) = Pr(S \mid W)Pr(W)/Pr(S)$$

であり、かつ$Pr(S)$はWに依存しないので、

$$\mathbf{argmax}_W\ Pr(W \mid S) = \mathbf{argmax}_W\ Pr(S \mid W)Pr(W)$$

である。音声認識とは異なり、各単語の並びから入力である文字列が生成される確率は明らかに0か1であり(マッチするもの以外は考慮対象ではない)、通常言語モデルと呼ばれる$Pr(W)$のみを考慮すればよい。さらに$Pr(W)$をすべてのテキストについてもとめておくのは、事実上不可能なので、単語のN個組で近似する。たとえば単語trigramでは

$$\begin{aligned} Pr(W) &= \prod_{i=1}^{m} Pr(w_i \mid w_1w_2,\ldots,w_{i-1}) \\ &\approx \prod_{i=1}^{m} Pr(w_i \mid w_{i-1}w_{i-2}) \end{aligned}$$

となる。ただし$i-1$等が負になる場合は文の始まりである特別のシンボルが付いていると解釈し、それからの遷移確率を用いる。この値を最大化するような単語分割を求めればよい。

3 未知語の検出

未知語の検出問題については、過去多くの手法が提案されているが、そのほとんどはヒューリスティックに基づいたルールを構築するものである。たとえば連続した片仮名列やアルファベット列、1文字の漢字を自立語と同様に扱うことは、非常によく知られたヒューリスティックである[4]。西野[5]はさらに柔軟な対応を可能にするため正規表現をベースとした未知語テンプレートを登録し、これを単語と同様に扱って解析するという手法を提案した。文字種の変化点による区切りをベースにしながら、付属語辞書を援用し切り出された部分の後ろから最長一致で2次分割を行うものも存在する[6]。山田[7]は未知語そのものの文字列としての特徴ばかりではなく、その前後の文字列も含めてルールパターンとしている。

統計を用いたものとしては、永田[9]、森[10]が発表しており、いずれも文字のbigramで未知語文字列の生起確率を推定している。

ルールベースを構築するアプローチは言語エキスパートの先験的知識を反映しやすく、またアプリケー

ション固有の未知語に対応しやすいというメリットがある反面、ルールパターンが増加してくるとそれらの副作用も考慮しながらより詳細化することが著しく困難になる。かつ N-gram に基づく単語分割モデルに組み込むことは難しい。一方文字 N-gram を用いる方法は、その状態数から考えて (そもそも学習という *Closed*な世界では未知語は存在しない) きわめて多量のコーパスを集める必要がある[1]。そこで本研究では文字種（クラス）に基づくパターンで未知語を表現するという手法は踏襲しながら、ヒューリスティックではなく統計を基盤としたパターンの作成、および N-gram モデルへの導入を試みる。

4 パターンの生起確率にもとづく未知語モデル

現在まで文字列パターン (ルール) の選択は人手 (エキスパート) の経験にたよっていた。しかしながら、未知語を収集する場合の多くはベースとなる辞書やその統計量が得られていることが多く、それらからパターンを作成することが可能である。そして学習データの量に応じて逐次的により詳細化できることが望ましい。

文字列パターンを記述するとき文字種が大きな情報量をもっていることは明らかなので、まずベースの辞書の各文字を文字種を代表するマクロ文字（クラス) に置き換えたパターンを作成する。たとえば「サンプル」という単語からは"$CKCKCKCK$"というパターンが作成され、「取り出す」という単語からは"$CJCHCJCH$"というパターンが作成される。ただし CK, CJ, CH はそれぞれ片仮名、漢字、平仮名を代表するマクロ文字である。さてクラスで表現された単語は当該クラスに属する文字すべてを代表しているのできわめて多くの「単語でないもの」にマッチする。そこで各パターンで単語をより詳細に分類できる特定位置の文字をキー文字として選択する。たとえば"$CJCHCJCH$"というパターンは、その元になった単語を観察すると多くが「組み込む」「引き出し」などの複合動詞やそれが名詞化したものであることがわかる。つまり 3 文字目が「出」や「込」であることが多い。このような文字を以下のような統計量により選択した。

$$Pr(M_w, < c, pos >)\, Pr(M_w, < \bar{c}, pos >)$$

ただし M_wはある文字列パターン (マクロ文字で表現されたもの)、$< c, pos >$は位置 p の文字が c であること、$< \bar{c}, pos >$は逆に位置 p の文字が c ではないことを意味する。この値はある文字列パターンにおいてエントロピー最大となるような文字$< c, pos >$を求めることに相当し、この値が大きいほど位置 p の文字 c が当該文字列の詳細分類に役立つことが期待できる。したがって閾値δを設定して選択すればよい。

選択された文字および位置を用いて、パターンを分類する。パターン"$CJCHCJCH$"において、$< 2, 出 >$がキーとして選択されたならば"$CJCHCJCH$"は

$$CJCH\ CJ[\hat{}\ 出]CH\ \text{と}\ CJCH\ 出CH$$

に詳細分類される。ただし $CJ[\hat{}\ 出]$ は「出」ではない漢字を表すクラスである。先に述べたδの値を$-\infty$にするとすべての文字がクラスから分離し、元の文字に戻る。言い換えれば未知語を発生する確率が 0 になる。したがってこの値で各パターンが未知語を発生する余地の大きさを指定することができる。

未知語モデルを trigram を用いた単語分割モデルの枠組で記述すると、未知語 $c_1c_2,\ldots,c_k$に対する trigram 確率は

$$\begin{aligned} & Pr(c_1c_2,\ldots,c_k \mid w_{i-1}w_{i-2}) \\ \simeq\ & Pr(unk \mid w_{i-1}w_{i-2}) \\ & \times Pr(c_1c_2,\ldots,c_k \mid unk) \ + \\ & Pr(known \mid w_{i-1}w_{i-2}) \\ & \times Pr(c_1c_2,\ldots,c_k \mid known) \end{aligned}$$

である。ただし $unk, known$はある辞書にとって $c_1c_2,\ldots,c_k$ が未知または既知である状態を示す。$c_1c_2,\ldots,c_k$が未知語であるという条件の元では、明らかに $Pr(c_1c_2,\ldots,c_k \mid known)$ の値は 0 であり第 1 項のみを考慮すればよいことになる。したがって任意の文字列 $w_{unk} = c_1c_2,\ldots,c_k \notin VS$($VS$は語彙セット) に対してその生起確率

$$Pr(c_1c_2,\ldots,c_k \notin VS)$$

言い換えれば未知語であることがわかったという条件での文字列$c_1c_2,\ldots,c_k$の条件付き確率が求まれば

[1] 頻度カウントが 1 のものを仮の未知語として扱う手法や、コーパスを N 個に分割し、個々に語彙セットを作成し、重複していないものを未知語とするといった手法が提案されている。

よい。いま文字列 $c_1c_2,\ldots,c_k$ が未知語パターン T_p から生成されたものと考えると

$$\begin{aligned} & Pr(c_1c_2,\ldots,c_k \mid unk) \\ = & \sum_{T_p}(Pr(c_1c_2,\ldots,c_k \mid T_p,unk) \times Pr(T_p \mid unk)) \end{aligned}$$

である。同一文字が複数のクラスに属さないという条件の元では、該当する未知語パターンは1個 $T_{p'}$ に限られ、$\sum$ をはずすことができる。また $Pr(c_1c_2,\ldots,c_k \mid T_p,unk)$ を構成する各クラス (g_j) が当該文字 (c_j) である条件付き確率の積で近似することにすると上式はさらに

$$\begin{aligned} & Pr(c_1c_2,\ldots,c_k \mid unk) \\ = & Pr(c_1c_2,\ldots,c_k \mid T_{p'},unk) \times Pr(T_{p'} \mid unk) \\ \simeq & \prod_{j=1}^{k} Pr(c_j \mid g_j,unk) \times Pr(T_{p'} \mid unk) \end{aligned}$$

と展開できる。したがって $Pr(unk \mid w_{i-1}w_{i-2})$、$Pr(T_p \mid unk)$、および $Pr(c_j \mid g_j,unk)$ を学習データから求めればよいことがわかる。

本手法は分割の単位が何であるかには依存していないため、単語単位への分割ばかりでなく、形態素への分割においても適用可能であることは言うまでもない。

5 確率の推定

「はじめに」で述べたように、われわれは人が分割した学習テキストから単語分割を品詞レベルでモデル化し、形態素解析を行った後、そのモデルに基づいた乱数を用いて分割・統合を行って分割コーパス (N-gram の学習データ) を作成した [11]。原理的には同じデータから $Pr(T_p \mid unk)$ 等が学習できるはずであるが、現実には形態素解析のエラーのため未知語部分についてはよい学習データとはならない。少量のデータについて予備実験を行ったところベース辞書 (約 42,000 語) で未知語と判定された箇所の内、約 30%で形態素解析エラーが生じていた。そこで辞書に含まれている単語の unigram 確率が θ 以下のもの (約 36,000 語[2]) を仮の未知語 (w_{unk}) として $Pr(T_p \mid unk)$、および $Pr(c_j \mid g_j,unk)$ を学習した。具体的には T_p が生成し得る文字列の集合を $W_{set}(T_p)$ として

$$\begin{aligned} & Pr(T_p \mid unk) \\ = & \sum_{w_{unk}} Pr(w_{unk} \in W_{set}(T_p)) / \sum_{w_{unk}} Pr(w_{unk}) \end{aligned}$$

$$\begin{aligned} & Pr(c_j \mid g_j,unk) \\ = & \sum_{w_{unk}} Pr(w_{unk},g_j,c_j) / \sum_{w_{unk}} Pr(w_{unk},g_j) \end{aligned}$$

により推定した。これらは仮の未知語について unigram 確率がわかっていれば計算できる。ただし、われわれの近似モデルでは各未知語テンプレートが未知語をどの程度生成しやすいかというパラメータが過小評価されやすい。実際未知語テンプレートはその名前に反して、既知の単語 (文字列) も生成するわけで、そのための補正として実験では $Pr(w_{unk} \in W_{set}(T_p)/(Pr(w_{unk} \in W_{set}(T_p)) + Pr(w_{known} \in W_{set}(T_p)))$ を掛けて $Pr(T_p \mid unk)$ とした。

$Pr(W)$ を最大化する w 系列は A*探索により N-best 解を求めた。

[2]残り、高頻度単語約 6,000 語で Coverage は 90%にのぼる。

6 実験と考察

ベースとなる N-gram モデルに用いた学習テキストは、日経新聞、産経新聞、毎日新聞、EDR コーパス、パソコン通信 People の電子会議室からとったもので、それぞれの文章数を表 1 に示す。

データはそれぞれ全体を 95%と 5%に 2 分し、前者を N-gram カウントに、後者を Held-out 補間に用いた。つまり trigram をそれ自身と bigram、unigram により線形補間している。

表 1: ソース別のテキスト数

日経新聞	715K
産経新聞	1,837K
毎日新聞	1,401K
EDR コーパス	169K
People	1,391K

一方、ベースとして用いた辞書は約 42,000 語を含んでおり、学習データに対して約 95%の Coverage を

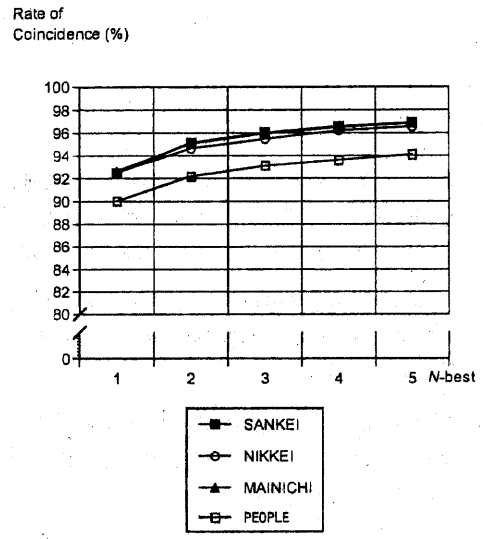

図 1: 人の分割結果との一致率

もつように設定したものである。これから上記の手法により未知語パターンを作成したところ、全部で290個のパターンが得られた。そして、新聞および電子会議室 People から採った文章 (600 文-1000 文、Open data) について分割した結果を人が分割した結果と比較したのが図 1 である。比較には以下のように定義される一致率 (Rate of Coincidence: ROC %) を用いた。

$$ROC = \frac{(\text{人による分割数}) - Ins. - Del.}{\text{人による分割数}} \times 100$$

ただし $Ins.$は人が分割しなかった箇所での分割数、$Del.$は人が分割した箇所での非分割数を意味する。図から平均 90-92%、5 位までの best 解の累積で 94-96%の一致率が得られることがわかる。新聞 3 紙はほとんど同じ一致率を示し、People のみがやや低い。

次に未知語箇所のみに注目して、人による分割と比較してみる。エキスパートが指摘した未知語箇所に対して、本手法の結果得られた未知語を比較してみると表 2 が得られた。

ただし n 位の再現率とは n 位までの分割候補の中から ROCがもっとも高いものを選び、その結果に対して再現率を求めたものである。この数字自体はそれほど高いとは言えない。しかしながらここで対象としているのは、人が直観的に「単語」と感じるものであり、それ自身正解が存在しないか、あるにしてもしばしば意見が分かれるものである。そこで 5 位までに人と一致する未知語文字列が得られなかった場合について候補内で、ROCがもっとも高かった解と人による分割結果と比較し A. いずれの分割も可能、B. いずれも可能だが人による分割の方がより自然、C. エラー [3]に分類したのが図 2 である。

表 2: 未知語の抽出精度

テキスト	再現率 (1 位)	再現率 (5 位)	適合率
日経新聞	37.6%	51.5%	84.5%
産経新聞	46.8%	65.7%	79.2%
毎日新聞	46.5%	62.5%	79.7%
People	52.0%	60.2%	80.7%

Acceptable で多いものは 3 文字漢語の接辞を分割する・しないの違い (i.e.「積極策」と「積極 + 策」)、助数詞を分割する・しないなどである。助詞「が」や「て」などの直前で分割されなかったもの (「とどまったが」) もめだった。これらは自然な単語単位とも見ることができるが人の単語分割モデルの学習データに、そのような例がほとんど含まれていないことを考慮し、Error に分類した。Unnatural に分類されたものでは、複合動詞が (結合する方が自然であるところを) 分割された例 (i.e.「払い + 出し」) が多かった。Error になったものとしては会話体特有の言い回しである「って」が分割されたもの[4]、平仮名書きされたために別の意味の単語列になってしまった例 (「なおりました」が「な + おりました」と分割) などがある。電子会議室のデータは Error が多いが、これはベースの学習データを作成したさいの精度、とくに形態素解析のそれが影響し、誤った単語列を学習した可能性が考えられる。

さらにパープレキシティ (Perp.)、カバレージ (Cov.) を、人が分割したものと比較すると表 3 のようになった。

[3]「意味がとおらないか、異なってしまっている」という基準で判断。

[4]分割すると発声できなくなるため、被験者はほとんどすべて前の形態素と結合していた。

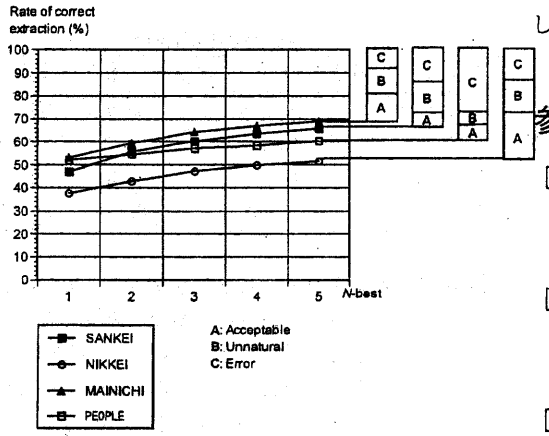

図 2: 未知語の抽出率

本手法は単語列の並びとしての生起確率を最大化するというものであり、パープレキシティについては必ずしも明らかではない。上記実験では産経新聞、毎日新聞がやや増加しており、他2データは逆に下がるなど、一貫した傾向は見られなかった。一方カバレージは一様に上がり98%から99%に達している。

以上述べたように、単語分割精度としてはほぼ満足すべき結果が得られたが、未知語抽出率という点ではさらなる改善が望まれる。とくに口語体を多く含む電子会議室 (People) のデータでは形態素解析の精度がまだ十分ではないことが懸念され、この観点からの検討も行いたい。

謝辞

データを利用させていただいた産経新聞社、日本経済新聞社、毎日新聞社 (CD-毎日新聞94) ならびに (株) ピープルワールドカンパニーに感謝します。また同じグループの荻野紫穂研究員にはルールに基づく未知語抽出法の論文をご教示いただいた。ここに感謝したい。

表 3: Perplexity と Coverage

テキスト	人による分割		本手法による分割	
	Perp.	Cov.(%)	Perp.	Cov.(%)
日経新聞	105.7	98.0	86.3	98.7
産経新聞	168.5	95.5	178.1	98.7
毎日新聞	136.5	96.1	147.5	98.1
People	207.1	95.2	197.5	97.9

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兼用, 音響学会講演論文集, 平成8年度春, pp.159-160, (1996).

[2] 西村他: 単語を認識単位とした日本語大語彙連続音声認識, 音響学会講演論文集, 平成9年度秋季, pp. 95-96, (1997).

[3] 西村, 伊東: 単語を認識単位とした日本語ディクテーションシステム, 電子通信学会論文誌, D-II, 1998 1月号 掲載予定.

[4] 吉村他: 未登録語を含む日本語文の形態素解析, 情報処理学会論文誌, Vol. 30, No. 3, pp. 294-301, (1989).

[5] 西野: 未登録語テンプレートを用いた日本語形態素解析, 情処第39回全国大会, 2F-2, pp. 594-595, (1989).

[6] 朴, 筧: 語の連接関係を利用した未知語の形態素辞書情報の獲得手法, 自然言語処理, Vol. 4, No. 1, pp. 71-86, (1997).

[7] 山田: 統計情報を用いた日本語形態素解析, 言語処理学会第3回年次大会発表論文集, pp. 417-420, (1997).

[8] M. Nagata: Automatic Extraction of New Words from Japanese Texts using Generalized Forward-Backward Search, EMNLP, pp. 48-59, (1996).

[9] 永田: 単語頻度の期待値に基づく未知語の自動収集, 情処研究会, NL 116-3, pp. 13-20, (1996).

[10] 森, 山路: 日本語の情報量の上限の推定, 情報処理学会論文誌, 1997 11月号掲載予定.

[11] 伊東他: 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, 情処研究会 , NL 116-9, pp. 57-64, (1996).