社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－SLP－63（4）
2006／10／20

最大エントロピー法による単語境界確率の推定

森 信介† 倉田 岳人† 小田 裕樹
†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
{forest,gakuto}@jp.ibm.com, oda@fw.ipsj.or.jp

あ ら ま し
言語モデルの分野適応において、適応対象の分野の単語境界情報のない生コーパスの有効な利用方法として、確率的単語分割コーパスとしての利用が提案されている。この枠組では、生コーパス中の各文字間に単語境界が存在する確率を付与し、それを用いて単語 n-gram 確率などが計算される。本論文では、この単語境界確率の新しい推定方法を提案し、これによってより良い言語モデルが構築できることを実験的に示す。加えて、確率的単語分割コーパスを従来の決定的に単語に分割されたコーパスで模擬する方法を提案し、言語モデルの能力を下げることなく計算コストが削減できることを示す。

キーワード　言語モデル　確率的単語分割　最大エントロピー法　音声認識

Word Boundary Probability Estimation by a Maximum Entropy Model

Shinsuke MORI†, Gakuto KURATA†, Hiroki ODA
†IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma Yamatoshi Kanagawaken 242-8502 Japan
{forest,gakuto}@jp.ibm.com, oda@fw.ipsj.or.jp

Abstract
Language model (LM) building needs a corpus whose sentences are segmented into words. For languages in which the words are not delimited by whitespace, an automatic word segmenter built from a general domain corpus is used. Automatically segmented sentences, however, contains many segmentation errors especially around words and expressions belonging to the target domain. To cope with segmentation errors, the concept of stochastic segmentation has been proposed. In this framework, a corpus is annotated with word boundary probabilities that a word boundary exists between two characters. In this paper, first we propose a novel method to estimate word boundary probabilities to obtain a better LM. Next we propose a method for simulating a stochastically segmented corpus by a segmented corpus and show that the computational cost is reduced without a performance degradation.

Key Words Language Modeling, Stochastic Segmentation, Maximum Entropy Method, Speech Recognition

1 はじめに

一般的な分野において精度の高い単語分割済みコーパスが利用可能になってきた現在、言語モデルの課題は、言語モデルを利用する分野への適応、すなわち、適応対象分野に特有の単語や表現の統計的振る舞いを的確に捉えることに移ってきている。この際の標準的な方法では、適応対象のコーパスを自動的に単語分割し、単語 n-gram 頻度などが計数される。この際に用いられる自動単語分割器は、一般分野の単語分割済みコーパスから構築されており、分割誤りの混入が避けられない。特に、適切に単語分割される必要がある適応対象分野に特有の単語や表現やその近辺において誤る傾向があり、単語 n-gram 頻度などの信頼性を著しく損なう結果となる。

上述の単語分割誤りの問題に対処するため、確率的単語分割コーパスという概念が提案されている [1]。この枠組では、適応対象の生コーパスは、各文字の間に単語境界が存在する確率が付与された確率的単語分割コーパスとみなされ、単語 n-gram 確率が計算される。従来の決定的に自動単語分割された結果を用いるより予測力の高い言語モデルが構築できることが確認されている。また、仮名漢字変換 [2] や音声認識 [3] においても、従来手法に対する優位性が示されている。

確率的単語分割コーパスの初期の論文では、単語境界確率は、自動分割により単語境界と推定された箇所では単語分割の精度 α (例えば 0.95) とし、そうでない箇所では $1-\alpha$ とする単純な方法により与えられている。実際には、単語境界が存在すると推定される確率は、文脈に応じて幅広い値を取ると考えられる。例えば、学習コーパスからはどちらとも判断できない箇所では 1/2 に近い値となるべきであるが、既存手法では 1 に近い α か、0 に近い $1-\alpha$ とする他ない。この問題に加えて、既存手法よりも計算コストが高いことが挙げられる。ある単語 n-gram の頻度を確率的単語分割コーパスに対して計算するためには、その単語の文字列としてのすべての出現に対して、頻度のインクリメントではなく、複数回の浮動小数点演算を実行しなければならない。この計算コストにより、単語クラスタリングによる言語モデルの改善 [4] や文脈に応じた参照履歴の伸長 [5] などの過去に提案された様々な言語モデルの改良を試みることが困難になっている。

本論文では、まず、確率的単語分割コーパスにおける新しい単語境界確率の推定方法を提案する。実験の結果、提案手法により約 16%のパープレキシティの減少と約 4.6%の文字誤り率の削減が確認された。さらに、確率的単語分割コーパスを通常の決定的に単語に分割されたコーパスにより模擬する方法を提案する。実験の結果、言語モデルの能力を下げることなく、確率的単語分割コーパスの利用において必要となる計算コストが削減可能であることを示した。これにより、高い性能の言語モデルを基礎として、既存の言語モデルの改良法を試みることが容易になる。

2 確率的単語分割コーパスからの言語モデルの推定

確率的言語モデルを新たな分野に適応する一般的な方法は、適応分野のコーパスを用意し、それを自動的に単語分割し、単語の頻度統計を計算することである。この方法では、単語分割誤りにより適応分野のコーパスにのみ出現する単語が適切に扱えないという問題が起こる。この解決方法として、適応分野のコーパスを確率的単語分割コーパスとして用いることが提案されている [1]。この節では、確率的単語分割コーパスからの確率的言語モデルの推定方法について概説する。

2.1 確率的単語分割コーパス

確率的単語分割コーパスは、生コーパス C_r(以下、文字列 $\boldsymbol{x}_1^{n_r}$ として参照) とその連続する各 2 文字 x_i, x_{i+1} の間に単語境界が存在する確率 P_i の組として定義される。最初の文字の前と最後の文字の後には単語境界が存在するとみなせるので、$i=0$, $i=n_r$ の時は便宜的に $P_i=1$ とされる。確率変数 X_i を

$$X_i = \begin{cases} 1 & x_i, x_{i+1} \text{ の間に単語境界が存在する場合} \\ 0 & x_i, x_{i+1} \text{ が同じ単語に属する場合} \end{cases}$$

とし ($P(X_i=1)=P_i, P(X_i=0)=1-P_i$)、各 $X_0, X_1, \ldots, X_{n_r}$ は独立であることが仮定される。

文献 [1] の実験で用いられている単語境界確率の推定方法は次の通りである。まず、単語に分割されたコーパスに対して自動単語分割システムの境界推定精度 α を計算しておく。次に、適応分野のコーパスを自動単語分割し、その出力において単語境界であると判定された点では $P_i=\alpha$ とし、単語境界でないと判定された点では $P_i=1-\alpha$ とする。後述する実験の従来手法としてこの方法を採用した。

2.2 単語 n-gram 頻度

確率的単語分割コーパスに対して単語 n-gram 頻度が以下のように定義される。

単語 0-gram 頻度 確率的単語分割コーパスの期待単語数として以下のように定義される。

$$f(\cdot) = 1 + \sum_{i=1}^{n_r-1} P_i \tag{1}$$

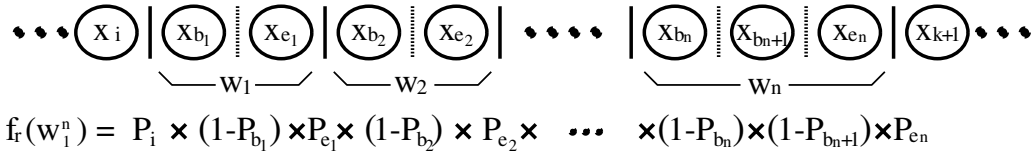

図 1: 確率的単語分割コーパスにおける単語 n-gram 頻度

単語 1-gram 頻度 確率的単語分割コーパスに出現する文字列 $\boldsymbol{x}_{i+1}^{k}$ が $l = k - i$ 文字からなる単語 $w = \boldsymbol{x'}_1^l$ である必要十分条件は以下の 4 つである。

1. 文字列 $\boldsymbol{x}_{i+1}^{k}$ が単語 w に等しい。
2. 文字 x_{i+1} の直前に単語境界がある。
3. 単語境界が文字列中にない。
4. 文字 x_k の直後に単語境界がある。

したがって、確率的単語分割コーパスの単語 1-gram 頻度 f_r は、単語 w の表記の全ての出現 $O_1 = \{(i,k) \mid \boldsymbol{x}_{i+1}^{k} = w\}$ に対する期待頻度の和として以下のように定義される。

$$f_r(w) = \sum_{(i,k)\in O_1} P_i \left[\prod_{j=i+1}^{k-1} (1 - P_j) \right] P_k \qquad (2)$$

単語 n-gram 頻度 ($n \geq 2$) L 文字からなる単語列 $\boldsymbol{w}_1^n = \boldsymbol{x'}_1^L$ の確率的単語分割コーパス $\boldsymbol{x}_1^{n_r}$ における頻度、すなわち単語 n-gram 頻度について考える。このような単語列に相当する文字列が確率的単語分割コーパスの $(i+1)$ 文字目から始まり $k = i + L$ 文字目で終る文字列と等しく ($\boldsymbol{x}_{i+1}^{k} = \boldsymbol{x'}_1^L$)、単語列に含まれる各単語 w_m に相当する文字列が確率的単語分割コーパスの b_m 文字目から始まり e_m 文字目で終る文字列と等しい ($\boldsymbol{x}_{b_m}^{e_m} = w_m,\ 1 \leq \forall m \leq n;\ e_m + 1 = b_{m+1},\ 1 \leq \forall m \leq n-1;\ b_1 = i+1;\ e_n = k$) 状況を考える (図 1 参照)。確率的単語分割コーパスに出現する文字列 $\boldsymbol{x}_{i+1}^{k}$ が単語列 $\boldsymbol{w}_1^n = \boldsymbol{x'}_1^L$ である必要十分条件は以下の 4 つである。

1. 文字列 $\boldsymbol{x}_{i+1}^{k}$ が単語列 $\boldsymbol{w}_1^n$ に等しい。
2. 文字 x_{i+1} の直前に単語境界がある。
3. 単語境界が各単語に対応する文字列中にない。
4. 単語境界が各単語に対応する文字列の後にある。

確率的単語分割コーパスにおける単語 n-gram 頻度は以下のように定義される。

$$f_r(\boldsymbol{w}_1^n) = \sum_{(i,e_1^n)\in O_n} P_i \left[\prod_{m=1}^{n} \left\{ \prod_{j=b_m}^{e_m-1} (1 - P_j) \right\} P_{e_m} \right]$$

ここで

$$\begin{aligned} e_1^n &= (e_1, e_2, \cdots, e_n) \\ O_n &= \{(i, e_1^n) | \boldsymbol{x}_{b_m}^{e_m} = w_m, 1 \leq m \leq n\} \end{aligned}$$

である。

2.3 単語 n-gram 確率

確率的単語分割コーパスにおける単語 n-gram 確率は、単語 n-gram 頻度の相対値として計算される。

単語 1-gram 確率 以下のように単語 1-gram 頻度を単語 0-gram 頻度で除することで計算される。

$$P_r(w) = \frac{f_r(w)}{f_r(\cdot)} \qquad (3)$$

単語 n-gram 確率 ($n \geq 2$) 以下のように単語 n-gram 頻度を単語 $(n-1)$-gram 頻度で除することで計算される。

$$P_r(w_n | \boldsymbol{w}_1^{n-1}) = \frac{f_r(\boldsymbol{w}_1^n)}{f_r(\boldsymbol{w}_1^{n-1})} \qquad (4)$$

3 最大エントロピー法による単語境界確率の推定

この節では、最大エントロピー法による単語分割器を単語境界確率の推定に用いる方法について述べる。

3.1 単語境界確率の推定

日本語の単語分割の問題は、入力文の各文字間に単語境界が発生するか否かを予測する問題とみなせる。つまり、文 $\boldsymbol{x} = x_1 x_2 \cdots x_m$ の各文字 x_i に対して、その文字が単語の末尾であるか否かを表す POC (position of character) タグ t_i を付与する問題とみなすのである [6]。POC タグは、単語の末尾文字であることを表すタグ **E** と、それ以外の文字であることを表すタグ **N** の 2 つのタグからなる。各文字の POC タグがこのいずれかであるかは、単語境界が明

示されたコーパスから学習された最大エントロピーモデル (ME model; maximum entropy model) により推定する。その結果、より高い確率を与えられたタグをその文字のタグとし、単語境界を決定する。すなわち、以下の式が示すように、最大エントロピーモデルにより単語の末尾文字と推定される確率がそれ以外の文字であると推定される確率より高い場合に単語の末尾文字とする。

$$t_i = \begin{cases} \mathbf{E} & \text{if } P_{ME}(t_i = \mathbf{E}|\boldsymbol{x}) > P_{ME}(t_i = \mathbf{N}|\boldsymbol{x}) \\ \mathbf{N} & \text{otherwise} \end{cases}$$

これにより、入力文を単語に分割することができる。

本論文では、以下のように、POC タグの出現確率を確率的単語分割コーパスにおける単語境界確率 P_i として用いることを提案する。

$$P_i = P_{ME}(t_i = \mathbf{E}|\boldsymbol{x})$$

これにより、注目する文字の周辺のさまざまな素性を参照し、単語境界確率を適切に推定することが可能になる。

3.2 参照する素性

後述する実験においては、最大エントロピーモデルの素性としては、x_{i-2}^{i+2} の範囲の文字 n-gram$(n = 1, 2, 3)$ をすべて用いた。ただし、以下の点を考慮している。

- 素性として利用する文字は、その字種[1] が次の文字の字種と同じか否かの情報 $k_i \in \{+, -\}$ を付加した拡張文字 $x_i^{'} = (x_i, k_i)$ [7] として参照する。
- カタカナ、アルファベット、数字は、各々1 つのクラスとみなして同一視する。
- 学習データ中で出現頻度が 1 回の拡張文字は未知文字とみなされ、字種ごとに用意した未知文字クラスを学習に用いる。
- 各文の文頭と文末には、それを表す特殊記号 BT を必要に応じて付加する ($x_i = \mathtt{BT}$, where $i < 1 \vee m < i$)。

最大エントロピーモデルのパラメータ推定には、GIS アルゴリズム [8] を使用した。

4 疑似確率的単語分割コーパス

確率的単語分割コーパスに対する単語 n-gram 頻度は、高いコストの計算を要する。また、確率的単語分割コーパスは、頻度計算の対象となる単語や単語断片 (候補) を多数含む。ある単語 n-gram の頻度の計算に際しては、その単語の文字列としてのすべての出現に対して、頻度のインクリメントではなく、複数回の浮動小数点演算を実行しなければならない。この計算コストにより、より長い履歴を参照する単語 n-gram モデルや単語クラスタリングなどの言語モデルの改良が困難になっている。

上述の困難を容易に回避する方法として、単語分割済みコーパスで確率的単語分割コーパスを近似する方法を提案する。具体的には、確率的単語分割コーパスに対して以下の処理を最初の文字から最後の文字まで $(1 \leq i \leq n_r)$ 行なう。

1. 文字 x_i を出力する。
2. 0 以上 1 未満の乱数 r_i を発生させ P_i と比較する。$r_i < P_i$ の場合には単語境界記号を出力し、そうでない場合には何も出力しない。

これにより、確率的単語分割コーパスに近い単語分割済みコーパスを得ることができる。これを疑似確率的単語分割コーパスと呼ぶ。

上記の方法では、文字列としての出現頻度が低い単語 n-gram の頻度が確率的単語分割コーパスと疑似確率的単語分割コーパスにおいて大きく異なる可能性がある。そもそも、出現頻度が低い単語 n-gram の場合、単語分割が正しいとしても、その統計的振る舞いを適切に捉えるのは困難であるが、近似によって誤差が増大することは好ましくない。従って、この影響を軽減するために、上記の手続きを N 回行ない、その結果得られる N 倍の単語分割済みコーパスを単語 n-gram 頻度の計数の対象とすることとする。このときの N を本論文では倍率と呼ぶこととする。

5 評価

単語境界確率の推定方法の評価として、言語モデルの適応の実験を行なった。まず、適応対象文野の大きな生コーパスに既存手法と提案手法のそれぞれで単語境界確率を付与した。次に、その結果得られる確率的単語分割コーパスから単語 2-gram モデルを推定し、これを一般分野の単語分割済みコーパスから推定された単語 2-gram モデルと補間した。最後に、適応分野のテストコーパスに対して、予測力と仮名漢字変換 [2] の精度の評価を行なった。後者は、完璧な理想的なモデルを用いた場合の音声認識と考えることも可能である。この節では、実験の結果を提示し、評価を行なう。

[1] 字種は、漢字、ひらがな、カタカナ、アルファベット、数字、記号の 6 つとした。

表 1: 一般コーパス (単語分割済み)

用 途	文 数	単語数	文字数
学 習	20,808	406,021	598,264
テスト	2,311	45,180	66,874

主に新聞記事や辞書の例文からなる。

表 2: 適応対象コーパス (単語境界情報なし)

用 途	文 数	単語数	文字数
学 習	797,345	—	17,645,920
テスト	1,000	—	20,935

主に業務日報からなる。

5.1 実験の条件

実験に用いたコーパスは、主に新聞記事や辞書の例文からなる一般コーパスと業務日報からなる適応対象のコーパスである。一般コーパスの各文は正しく単語に分割され、各単語に入力記号列 (読み) が付与されている。これを 10 個に分割し、この内の 9 個を学習コーパスとし、残りの 1 個をテストコーパスとした (表 1 参照)。自動単語分割器や単語境界確率の推定のための最大エントロピーモデルはこの学習コーパスから構築される。一方、適応対象のコーパスは大量にあるが、単語境界情報を持たない。この内の 1,000 文に入力記号列 (読み) を付与しテストコーパスとし、残りを確率的単語分割コーパスとして言語モデルの学習に用いた (表 2 参照)。

5.2 評価基準

確率的言語モデルの予測力の評価に用いた基準は、テストコーパスにおける単語あたりのパープレキシティである。まず、テストコーパス C_t に対して未知語の予測も含む文字単位のエントロピー H を以下の式で計算する [9]。

$$H = -\frac{1}{|C_t|}\log_2 \prod_{\boldsymbol{w}\in C_t} M_{w,n}(\boldsymbol{w})$$

ここで、$M_{w,n}(\boldsymbol{w})$ は単語 n-gram モデルによる単語列 $\boldsymbol{w}$ の生成確率を、$|C_t|$ はテストコーパス C_t の文字数を表す。次に、単語単位のパープレキシティを以下の式で計算する。

$$PP = 2^{H\times\overline{|\boldsymbol{w}|}}$$

ここで $\overline{|\boldsymbol{w}|}$ は平均単語長 (文字数) である。

表 3: 単語境界確率の推定方法と言語モデルの能力の関係

	単語境界確率の推定方法	PP	CER
BL	単語自動分割器の精度	57.80	2.85%
ME	最大エントロピーモデル	48.53	2.72%

PP: パープレキシティ, CER: 文字誤り率

仮名漢字変換の評価基準は、文字誤り率である。文字誤り率は $\mathrm{CER} = 1 - N_{LCS}/N_{COR}$ と定義される。ここで、N_{COR} は正解に含まれる文字数であり、N_{LCS} は各文を一括変換することで得られる最尤解と正解との最長共通部分列 (LCS; longest common subsequence)[10] の文字数である。

5.3 単語境界確率の推定方法の評価

単語境界確率の推定方法の差異を調べるために、以下の 2 つの確率的単語分割コーパスを作成しそれらから推定された単語 2-gram モデルの能力を調べた。

BL: 従来手法
各単語境界確率は、単語 2-gram モデルに基づく自動単語分割器の判断に応じて α 又は $1-\alpha$ とする。ここで、$\alpha = 0.9852$ は一般分野のテストコーパスにおける単語境界推定精度である (第 2.1 項参照)。

ME: 提案手法
各単語境界確率は、最大エントロピーモデルを用いて文脈に応じて推定される (第 3.1 項参照)。

適応対象分野のテストコーパスにおける予測力と文字誤り率を表 3 に示す。この結果から、本論文で提案する最大エントロピー法による単語境界確率の推定方法により約 16%のパープレキシティの削減が実現されている。この結果から、最大エントロピー法により推定された単語境界確率を持つ確率的単語分割コーパスを用いることで適応対象分野における単語 2-gram 確率がより正確に推定されていることがわかる。文字誤り率の比較から、提案手法により、従来手法の文字誤りの約 4.6%が削減され、この点からも言語モデルが改善されていることが確認される。従来手法の文字正解率は 97.15%と高いので、提案手法により実現された誤りの削減は十分有意義であろう。

5.4 疑似確率的単語分割コーパスの評価

本論文のもう一つの論点は、単語分割済みコーパスによる確率的単語分割コーパスの近似である。この評価として、

表 4: 疑似確率的単語分割コーパスから推定された言語モデルの能力

方法	学習コーパス	倍率	PP	CER
ME	確率的単語分割	–	48.53	2.72%
ME	疑似確率的単語分割	×1	50.69	2.85%
ME	疑似確率的単語分割	×10	48.99	2.81%
ME	疑似確率的単語分割	×100	48.68	2.75%

PP: パープレキシティ, CER: 文字誤り率

疑似確率的単語分割コーパスから推定した言語モデルのテストコーパスに対するパープレキシティと文字誤り率を複数の倍率 ($N = 1, 10, 100$) に対して計算した。表 4 はその結果である。倍率が 1 の場合は、パープレキシティや文字誤り率は、確率的単語分割コーパスから推定された言語モデルに対して少し高く、倍率を上げることによりこれらは確率的単語分割コーパスによる結果に近づいていくことがわかる。このことから、疑似確率的単語分割コーパスは、確率的単語分割コーパスのよい近似となっているといえる。倍率が 100 の場合は、単語に分割された 79,734,500 文から言語モデルを推定することになる。現在の計算機はこの大きさのコーパスを処理する能力が十分ある。したがって、単語 3-gram モデルや可変長記憶マルコフモデル、あるいは言語モデルのための単語クラスタリングなどさらなる言語モデルの改善を容易に試みることが可能となる。

6 おわりに

本論文では、確率的単語分割コーパスにおける新しい単語境界確率の推定方法を提案した。実験の結果、提案手法により約 16%のパープレキシティの減少と約 4.6%の文字誤りの削減が確認された。さらに、確率的単語分割コーパスを通常の決定的単語分割コーパスにより模擬する方法を提案した。実験の結果、言語モデルの能力を下げることなく、確率的単語分割コーパスの利用において必要となる計算コストが削減可能であることを示した。

参考文献

[1] Shinsuke Mori and Daisuke Takuma. Word n-gram probability estimation from a Japanese raw corpus. In *Proc. of the ICSLP2004*, 2004.

[2] 森信介. 無限語彙の仮名漢字変換. 情報処理学会研究報告, 第 NL172 巻, 2006.

[3] Gakuto Kurata, Shinsuke Mori, and Masafumi Nishimura. Unsupervised adaptation of a stochastic language model using a Japanese raw corpus. In *Proc. of the ICASSP2006*, 2006.

[4] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[5] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, Vol. 25, pp. 117–149, 1996.

[6] N. Xue. Chinese word segmentation as character tagging. *International Journal of Computational Linguistics and Chinese*, Vol. 8, No. 1, pp. 29–48, 2003.

[7] 風間淳一, 宮尾祐介, 辻井潤一. 教師なし隠れマルコフモデルを利用した最大エントロピータグ付けモデル. 自然言語処理, Vol. 11, No. 4, pp. 3–24, 2004.

[8] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The annuals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1479–1480, 1972.

[9] 森信介, 山地治. 日本語の情報量の上限の推定. 情処論, Vol. 38, No. 11, pp. 2191–2199, 1997.

[10] Alfred V. Aho. 文字列中のパターン照合のためのアルゴリズム. コンピュータ基礎理論ハンドブック, I: 形式的モデルと意味論, pp. 263–304. Elseveir Science Publishers, 1990.