確率的単語分割コーパスからの単語 N-gram 確率の計算

森 信介[†] 宅間大介[†] 倉田岳人[†]

確率的言語モデルは、音声認識やスペルチェッカーなどの言語処理において重要な役割を担っている。最も一般的な確率的言語モデルは単語 n-gram モデルであるが、実用的な予測力を実現するには、正しく単語に分割された対象分野のコーパスが大量に必要である。日本語では単語境界は明示されないので、自動単語分割による推定結果を人手で修正する。これには、対象分野の語彙の知識を有する作業者があたる必要があり、多大な時間とコストがかかる。この問題を解決するために、本論文では、文字列である生コーパスに各文字間に単語境界が存在する確率を付与した「確率的単語分割コーパス」という概念を提案し、確率的単語分割コーパスからの単語 n-gram 確率の計算について述べる。この方法の有用性を評価するために、確率的言語モデルにおける昨今の課題である分野適応の実験を行ない、既存手法に対する優位性を示した。

Word N-gram Probability Calculation from a Stochastically Segmented Corpus

Shinsuke Mori[†] and Daisuke Takuma[†]

Statistical language modeling plays an important role in a state-of-the-art language processing system, such as speech recognizer, spelling checker, etc. The most popular language model (LM) is word n-gram model, which needs sentences annotated with word boundary information. In various Asian languages, however, words are not delimited by whitespace, so we need to annotate sentences with word boundary information to prepare a statistically reliable large corpus. In this paper, we present the concept of a stochastically segmented corpus, which consists of a raw corpus and word boundary probabilities, and a method for calculating word n-gram probabilities from a stochastically segmented corpus. In the experiment, our method is applied to a LM adaptation problem and showed an advantage to an existing method.

1. はじめに

実用化に至っている言語処理技術の多くは、確率的言語モデルに基づいている。音声認識システム[1)]の多くが、音響モデルとともに確率的言語モデルを参照し、複数の候補の中から最尤の単語列を選択する。文字誤り訂正[2)]では、確率的言語モデルの尤度に基づいて不自然な文字列とその訂正候補を列挙する。単語分割や形態素解析[3)]に代表される言語解析においても、ユーザーがその出力を直接必要としないので、本質的に必要かという疑問はあるが、少なくとも研究者や開発者の道具として確率的言語モデルを利用する方法が用いられている。一般的な分野に対しては、十分な量の学習コーパスが確保できるので、上述の確率的言語モデルの応用の精度は実用上十分に高い。しかしながら、利用可能な学習コーパスの量が乏しい新規の分野においては、正確な確率的言語モデルの構築が困難であり、これを用いる応用の精度の低下を招いている。したがって、確率的言語モデルを用いる言語処理の実用化における課題は確率的言語モデルの分野適応である。

確率的言語モデルは単語や文字の頻度に基づいており、確率的言語モデルの分野適応には適用分野の大量の例文 (コーパス) が必要不可欠である。実用的なモデルは単語を単位としているので、コーパスには単語境界の情報が付与されている必要がある。したがって、日本語などの単語境界が明示されない言語のモデル構築には、コーパス中の大量の文に単語境界を正しく付与する作業が不可欠である。例えば、音声認識を医療分野に応用する場合には、カルテや医療所見の機械可読の例文を収集し、それら例文を、新聞などの既存のコーパスから構築された自動単語分割システムにより単語に分割し、その結果を一文ずつ人手で修正する。

† 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.

このようにして得られた単語分割済みコーパスから認識語彙を選択し、コーパスにおける頻度を計数することで医療分野向けの音声認識システムの言語モデルが出来上がる。実用に耐える認識精度を確保するためには、一般的に最低数万文からなる単語分割済みコーパスが必要であり、この準備にかかる時間とコストは開発の大半を占める。医療などのように、専門用語や特殊な表現を多く含む場合には、新聞などの一般的な分野のコーパスから構築された既存の自動単語分割システムの精度は低く、このことが開発コストをさらに押し上げ、短時間での開発を阻害している。

前述の問題の解決を主な目的として、本論文では、確実な単語境界情報を含まないコーパスから確率的言語モデルを構築する方法を提案する。提案手法では、まず、各文字間に単語境界が存在する確率を計算する。次に、これを用いて生コーパスを確率的に単語に分割されたコーパスとみなし、単語 n-gram 確率を計算する。実験では、十分な量の主に新聞記事や辞書の例文からなる一般的な分野の単語分割済みのコーパスと、適応対象 (交通に関する対話) の例文があるという前提で、これらから複数の言語モデルを構築し、適応対象の分野での予測精度を評価基準としてその効果を検証する。

2. 確率的言語モデルとその応用

自然言語処理における確率的言語モデルの役割は、与えられた文字列がある言語の文である尤度を数値化することである。確率的言語モデルに基づく言語処理は、候補から解を選択する際にこの尤度を参照する。形態素解析器は解析系の一例であり、文字列を与えられると尤度が最大になる品詞と表記の組の列を計算する。認識系の代表例の音声認識器では、音響信号列を入力として、尤度が最大となる文字列を算出する際に、音響モデルと併せて確率的言語モデルを参照する。

2.1 確率的言語モデル

確率的言語モデルは、ある言語の文字集合を $\mathcal{X}$ として、その言語の文 $\mathcal{X}^*$ が出現する確率値を記述する。これは、以下のように表される。

$$P : \mathcal{X}^* \mapsto [0,\ 1]$$

確率的モデルであるので、確率値をすべての文字列に渡って合計すると 1 以下になる必要がある。

$$\sum_{\boldsymbol{x} \in \mathcal{X}^*} P(\boldsymbol{x}) \leq 1$$

最も一般的な言語モデルは単語 n-gram モデルである。このモデルは、文を単語列 $\boldsymbol{w}_1^h = w_1 w_2 \cdots w_h$ とみなし、これらを文頭から順に予測する。

$$M_{w,n}(\boldsymbol{w}_1^h) = \prod_{i=1}^{h+1} P(w_i | \boldsymbol{w}_{i-n+1}^{i-1})$$

この式の中の w_i $(i \leq 0)$ は、文頭に対応する特別な記号であり、w_{h+1} は、文末に対応する特別な記号である。完全な語彙を定義することは不可能であるから、未知語を表わす特別な記号 `UW` を用意する。未知語の予測の際は、まず、単語 n-gram モデルにより `UW` を予測し、さらにその表記 $\boldsymbol{x}_1^{h'}$ を以下の文字 n-gram モデルにより予測する。

$$M_{x,n}(\boldsymbol{x}_1^{h'}) = \prod_{i=1}^{h'+1} P(x_i | \boldsymbol{x}_{i-n+1}^{i-1}) \qquad (1)$$

この式の中の x_i $(i \leq 0)$ は、語頭に対応する特別な記号であり、$x_{h'+1}$ は、語末に対応する特別な記号である。したがって、未知語は以下のように予測される。

$$P(w_i | \boldsymbol{w}_{i-n+1}^{i-1}) = M_{x,n}(w_i) P(\texttt{UW} | \boldsymbol{w}_{i-n+1}^{i-1})$$

2.2 自動単語分割

日本語においては、単語単位への分割が自然言語処理における最初の曖昧性解消の問題である。この問題を解決するために、単語 n-gram モデルに基づく自動単語分割器が提案されている[4]。この方法では、以下の式で表されるように、文の生成確率が最大となる単語列を自動分割結果とする。

$$\hat{\boldsymbol{w}} = \mathop{\mathbf{argmax}}_{\boldsymbol{w}=\boldsymbol{x}} M_{w,n}(\boldsymbol{w})$$

永田[4] は、パラメータ推定に 10,945 文を用いて、約 97%の単語を単位とする精度を報告している。

確率的言語モデルを用いる言語処理を新しい分野に適用する際には、その分野特有の単語の統計的振舞をモデルに反映させるために、正しく単語に分割されたその分野の例文が大量に必要となる。これを生成するために、自動単語分割システムが利用されるが、一般的な分野のコーパスから構築された自動単語分割システムは、新しい分野特有の単語や表現の近辺で誤る傾向が強く、一意に自動分割した結果に対して単語の統計をとることには問題がある。

3. 生コーパスからの言語モデルの計算

この節では、生コーパスの各文字間に単語境界が存在する確率を付与し、生コーパスとこの情報から単語 n-gram 確率を計算する方法を提案する。この方法は、決定的に単語に分割されているコーパスを利用する場合を単語境界確率が 1 または 0 となる特殊な場合として包含する。

3.1 生コーパスの確率的単語分割

生コーパス C_r(以下、文字列 $\boldsymbol{x}_1^{n_r}$ として参照) を所与として、連続する 2 文字 x_i, x_{i+1} の間に単語境界が存在する確率 P_i を付与したものを考える。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 $P_i=1$ とする。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 $X_0, X_1, \ldots, X_{n_r}$ を独立とした単語分割を確率的単語分割と呼ぶことにする (図 1 参照)☆。

以下では、確率的に単語分割された生コーパスにおける単語 n-gram 頻度を定義し、その頻度に基づいて単語 n-gram 確率を計算する方法を示す。

3.2 単語 0-gram 頻度

分割済みコーパスにおける単語 0-gram 頻度は、そのコーパス中に含まれる単語の数 (延べ) である。本論文では、生コーパスにおける単語 0-gram 頻度 $f_r(\cdot)$ をそのコーパス中の期待単語数として定義する。これは付録 A.1 の命題 1 から

$$f(\cdot) = 1 + \sum_{i=1}^{n_r-1} P_i \tag{2}$$

である。この式から、単語 0-gram 頻度が確率的単語分割コーパスの単語分割確率を一度ずつ読むだけで計算できることがわかる。

3.3 単語 1-gram 頻度

確率的に単語分割された生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。

$$\boldsymbol{x}_{i+1}^{k} = \boldsymbol{x'}_1^{l}$$

(2) 文字 x_{i+1} の直前に単語境界がある。

$$X_i = 1$$

(3) 単語境界が文字列中にない。

$$X_j = 0, \ \ i+1 \le \forall j \le k-1$$

(4) 文字 x_k の直後に単語境界がある。

$$X_k = 1$$

したがって、X_i の独立性から、単語 w の生コーパス中の単語 1-gram 頻度 f_r は、単語 w の表記の全ての

```
…し た 高 橋 是 清 ( 大…
… 0.84 0.98 0.58 0.58 0.58 0.99 0.99 …
```

図 1 確率的分割済みコーパス
Fig. 1 Stochastically segmented corpus.

出現 $O_1 = \{(i,k) \,|\, \boldsymbol{x}_{i+1}^{k} = w\}$ に対する期待頻度の和として以下のように定義される。

$$\begin{aligned} f_r(w) &= \sum_{(i,k)\in O_1} P(X_i=1)\left[\prod_{j=i+1}^{k-1} P(X_j=0)\right] P(X_k=1) \\ &= \sum_{(i,k)\in O_1} P_i \left[\prod_{j=i+1}^{k-1} (1-P_j)\right] P_k \end{aligned} \tag{3}$$

k は単語 w の文字数 l とその生コーパス中での開始位置 i から定まるため ($k=i+l$)、事実上 Σ の中は i のみの 1 変数と考えて良い。ここでも、付録 A.1 の単語 0-gram 頻度の証明 1 と同様に、$f_r(w)$ が生コーパス C_r における単語 w の期待頻度であることが示せる。

3.4 単語 1-gram 確率

決定的に単語に分割されたコーパスからの単語 1-gram 確率の最尤推定の場合と同様に、確率的に単語分割された生コーパスにおける単語 1-gram 確率を

$$P_r(w) = \frac{f_r(w)}{f_r(\cdot)} \tag{4}$$

と定義する。付録 A.2 の命題 2 により $P_r(w)$ が確率として正しく定義されていることが示される。

3.5 単語 n-gram 頻度

単語 1-gram 頻度と同様に、L 文字からなる単語列 $\boldsymbol{w}_1^n = \boldsymbol{x'}_1^L$ の生コーパス $\boldsymbol{x}_1^{n_r}$ 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生コーパスの $(i+1)$ 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生コーパスの 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 ($\boldsymbol{x}_{b_m}^{e_m} = w_m, \ 1 \le \forall m \le n$; $e_m+1 = b_{m+1}, \ 1 \le \forall m \le n-1$; $b_1 = i+1$; $e_n = k$) 状況を考える。単語 1-gram 頻度の計算の場合と同様に、単語列と生コーパスの部分文字列は、文字列として対応していることに加えて、各文字間における単語境界の有無も対応している場合にのみ単語列が出現していると考えられる。したがって、確率的に単語分割された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。

$$\boldsymbol{x}_{i+1}^{k} = \boldsymbol{x'}_1^L$$

☆ 本論文で提案する単語 n-gram 確率の計算においてはこの独立性が必要である。

表 1 単語分割済みコーパス
Table 1 Segmented Corpora

	用途	文数	単語数	文字数
一般 (L)	学習	52,787	1,250,638	1,838,399
交通 (L)	学習	2,523	39,172	59,176
交通 (T)	テスト	280	4,244	6,385

表 2 生コーパス
Table 2 Raw corpora

	用途	推定単語数	文字数
新聞 (1)	学習	40,367.5	59,339
新聞 (2)	学習	47,929,857.3	70,455,401

推定単語数は文字数を学習コーパスの平均単語長で除した値である。

(2) 文字 x_{i+1} の直前に単語境界がある。

$$X_i = 1$$

(3) 単語境界が各単語に対応する文字列中にない。

$$X_j = 0,\ \ b_m \leq \forall j \leq e_m - 1,\ \ 1 \leq \forall m \leq n$$

(4) 単語境界が各単語に対応する文字列の後にある。

$$X_{e_m} = 1,\ \ 1 \leq \forall m \leq n$$

単語 1-gram 頻度と同様に、生コーパスにおける単語 n-gram 頻度を以下のように定義することができる。

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

ここで

$$e_1^n = (e_1, e_2, \cdots, e_n)$$

$$O_n = \{(i, e_1^n) | \boldsymbol{x}_{b_m}^{e_m} = w_m, 1 \leq m \leq n\}$$

とした。

3.6 単語 n-gram 確率

各 $n = 1, 2, \cdots$ について、単語 n-gram の出現確率を以下で定義する。

$$P_r(\boldsymbol{w}_1^n) = \frac{f_r(\boldsymbol{w}_1^n)}{f_r(\cdot)} \tag{5}$$

このとき、P_r が確率として適正であることが付録 A.3 の命題 3 により確認できる。また $\boldsymbol{w}_1^{n-1}$ が既知のときに、それに単語 w_n が続く確率は、以下のように単語 n-gram の出現確率の条件付確率として定義される。

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

これを単語 n-gram 確率と呼ぶ。式 (5) から単語 n-gram 確率は、以下の式が示すように、期待頻度の比として計算してよい。

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

4. 評　　価

生コーパスから推定した言語モデルの評価のために、様々な言語モデルを構築し比較した。課題は、確率的言語モデルの分野適応であり、評価基準は適応対象の分野におけるモデルの予測力である。

4.1 実 験 条 件

実験には、主に新聞記事や辞書の例文からなる適応元である一般分野のコーパスと、適応分野のコーパスとしての交通に関する対話の例文集を用いた (表 1 参照)。両分野のコーパスの各文は人手で単語に分割されているが、適応分野の学習コーパスは、提案手法では生コーパスとして利用される。単語分割済みコーパスとしての利用は、正しく単語に分割されている理想的な状況を実現し、これと提案手法を比較するためである。さらに、生コーパスが適応対象外の文からなる場合の提案手法の効果を評価するために、生コーパスとして新聞記事を用いる実験も行なった (表 2 参照)。

4.2 評 価 基 準

言語モデルの予測力の評価に用いた基準は、文字単位のクロスエントロピーと単語あたりのテストセットパープレキシティーである。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[6)]。

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

ここで、$|C_t|$ はテストコーパス C_t の文字数を表す。次に、単語単位のテストセット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。

4.3 自動単語分割システム

実験に利用した自動単語分割システムは、単語分割済みの一般分野のコーパスから構築された単語 2-gram モデル (後述するモデル A) に基づいており、入力文に対して最大確率となる単語列を返す (第 2.2 項参照)。学習コーパスと同一の分野のテストコーパスに対する単語境界の推定精度は 98.70%であった☆。参考までに、対象分野のテストコーパス (表 1 の交通 (T)) に対する単語境界の推定精度が 97.85%であったことを付記しておく。

4.4 単語境界確率の推定

確率的単語分割コーパスの単語境界確率の推定方法としては、自動分割の結果を利用する方法を採用した。

☆ 単語境界の推定精度の定義は、各隣接文字間において単語境界が存在するか否かが正しく推定された割合である。自動単語分割の評価基準としてよく用いられる再現率と適合率はそれぞれ 97.74% と 97.46%であった。

表 3 各モデルの予測精度
Table 3 Predictive powers of each model

	生コーパス	利用方法	モデル	語彙数	未知語率	H	PP
A	なし	–	–	20,788	4.19%	5.273	218.0
B	交通 (L)	自動分割	2-gram	21,689	2.78%	4.978	161.4
C	交通 (L)	確率分割	2-gram	341,790	1.86%	4.926	152.9
D	交通 (L)	手動分割	2-gram	21,888	1.93%	4.872	144.8
E	交通 (L)	確率分割	1-gram	341,790	1.86%	5.156	193.5
F	新聞 (1)	確率分割	2-gram	570,056	3.93%	5.234	209.5
G	新聞 (2)	確率分割	2-gram	448,172,469	0.57%	4.921	152.1

モデル B〜G はモデル A と補間されている。自動分割と手動分割の語彙は、モデル A の語彙と分割済みコーパスに出現するすべての単語との和集合である。また、確率分割における語彙数は、モデル A の語彙と確率的単語分割コーパスの読点を含まないすべての部分文字列との和集合である。ただし、G の新聞 (2) に関しては 61 文字以上の文字列を語彙としていない。

すなわち、自動分割の結果により 1 または 0 に決定される単語境界確率を自動単語分割システムの精度の分だけ割り引く。具体的には、自動単語分割システムにより単語境界であると判定された点では $P_i = 0.987$ とし、単語境界でないと判定された点では $P_i = 1-0.987$ とした。

4.5 言語モデル

基本となる言語モデルは以下の通りである。

モデル A: 単語分割済みの一般分野のコーパスから未知語モデルを含む単語 2-gram モデルを構築した (第 2.1 項参照)☆。これは単語 1-gram モデルと補間されており、補間係数は学習コーパスを 9 つに分割し削除補間法で求めた。このモデルの語彙 (既知語) は、9 つの部分学習コーパスの 2 つ以上に出現する 20,788 語とした。既知語に含まれない単語は未知語とみなされ、未知語記号の出現確率は既知語以外を未知語記号 `UW` に置き換えて単語 n-gram 頻度を計数することにより推定される。

様々な比較のために以下のモデルを構築した。言語モデルの分野適応が目的であるから、これらのモデル (適応分野のモデル) は上述のモデル A(一般分野のモデル) と線形補間されている[7] ☆☆。

モデル B (従来手法): 生コーパスの既知の利用方法は、これを自動で単語に分割することである。したがって、対象分野の学習コーパスを生コーパスとみなし、自動単語分割システムにより改めて自動で単語に分割することで得られる自動単語分割済みコーパスから単語 2-gram モデルを構築した。

モデル C (提案手法): 対象分野の学習コーパスを確率的単語分割コーパスとみなして提案手法により単語 2-gram モデルを推定した。

モデル D (上限): 対象分野の学習コーパスを単語分割済みコーパスとして利用し、単語 2-gram モデルを推定した。これは、生コーパスが正しく単語に分割されている場合を想定した理想的な状況であり、このモデルの構築には多大なコストがかかる。

以下の 3 つのモデルはモデル C の条件の一部を変更した結果得られるモデルである。

モデル E: モデル C において、単語 2-gram モデルではなく、単語 1-gram モデルを構築する。

モデル F: 対象分野の生コーパスと同程度の文字数からなる異なる分野の生コーパスを利用する。

モデル G: モデル C と同程度の予測力となる量の異なる分野の生コーパスを利用する。

4.6 評 価

各モデルの予測力を表 3 に示す。モデル A とモデル B の結果の比較から、生コーパスとしてであっても、適応対象の分野のコーパスが利用できれば、これを自動分割した結果に対して頻度の統計をとることで対象分野における予測力の向上が図れることが分かる。これは、従来の知見を追認する結果である。モデル B とモデル C の結果の比較から、生コーパスを決定的に自動分割するのではなく、確率的に分割されたコーパスとみなすことが、より有効な利用方法であることが分かる。つまり、自動分割の結果以外の分割の可能性を考慮に入れて単語の統計をとることで、適応対象の分野特有の単語や表現の近辺での致命的な分割誤りの影響が軽減されることが分かる。

生コーパスが正しく単語に分割されているという理想的な状況を前提とするモデル D のパープレキシティ

☆ 一般に、単語 2-gram モデルよりも単語 3-gram モデルの方が予測精度が高いが、モデル構築の方法の比較を目的とした場合には単語 2-gram モデルで十分である。

☆☆ 補間係数の推定は、モデル毎に、一般分野の学習コーパスの尤度が最大になるように削除補間法により推定した。本来、適応分野の学習コーパスの尤度を最大化すべきであるが、これには適応分野の学習コーパスが単語に分割されている必要がある。

は、従来手法のモデル B のパープレキシティーよりも約 10%低いが、現実的な条件のみを仮定するモデル C はこの半分の約 5%の低下を実現している。

生コーパスから単語 1-gram モデルを推定するモデル E の予測力は、単語 2-gram を用いるモデル C よりもかなり低い。生コーパスであっても 2-gram が有効であることが分かる。

対象分野とは異なる生コーパスを用いたモデル F とモデル C の結果の比較から、適応対象と異なる分野のコーパスは、大きさが同程度の場合には対象分野のコーパスほど効果がないことがわかる。この傾向は、従来手法である自動分割による場合と同様である。また、対象分野とは異なるより大きい生コーパスを用いたモデル G とモデル C の結果の比較から、新聞記事を用いた場合、同程度の効果を出すために約 1,187 倍の大きさが必要であったことが分かる。

生コーパスのサイズの影響を調べるために、生コーパスのサイズを変えて (1/1, 1/3, 1/9) モデル B(自動分割) とモデル C(確率分割) を構築し、それぞれの予測力を計算した。図 2 がこの結果である。このグラフから、生コーパスのサイズに関わりなく生コーパスを確率的単語分割コーパスとして利用する提案手法のほうが自動分割結果をそのまま用いる従来手法よりも優れていることが分かる。換言すれば、ある予測力を実現するために必要となる適応分野のコーパスの収集コストが提案手法により削減できることがわかる。また、モデル B(自動分割) のパープレキシティーは最も右の点でも強い減少傾向があることがわかる。したがって、対象分野の例文を収集し、より大きな生コーパスを作ることでさらなるモデルの改善が期待できるといえる。

多くの分野において、無視できる程度のコストでかなりの量の例文を収集可能である。それゆえ、対象分野の多くの例文を可能な限り収集し、提案手法を用いて単語 n-gram モデルを構築することが、音声認識などに必要な確率的言語モデルを新たな分野に適応させるよい方法である。

5. おわりに

本論文では、生コーパスから単語単位の言語モデルを構築する方法を提案した。これにより自動単語分割システムが一意に自動分割した結果に対する単語の統計よりもよい言語モデルが構築でき、自動分割の結果を人手で修正する必要性が軽減される。

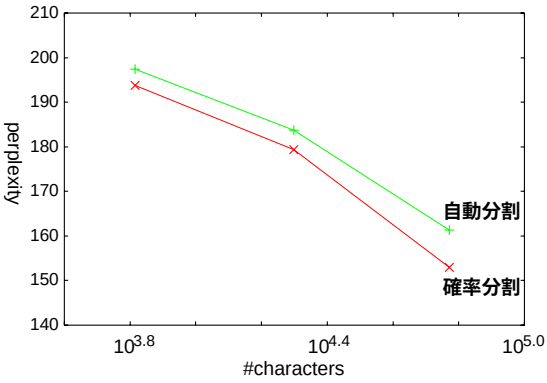


図 2 生コーパスのサイズとパープレキシティーの関係
Fig. 2 Relation between raw corpus size and perplexity.

付　　録

A.1 単語 0-gram 頻度

生コーパスの単語 0-gram 頻度、すなわち期待単語数に関して以下の命題が成り立つ。

命題 1 n_r 文字からなる生コーパス $C_r = \boldsymbol{x}_1^{n_r}$ の各文字境界 (x_i と x_{i+1} の間) が単語境界である確率 P_i $(1 \leq i \leq n_r - 1)$ と生コーパスの中の期待単語数 $f(\cdot)$ の間には以下の関係式が成り立つ。

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

証明 1 X_i を以下で定義される確率変数とする。

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

この定義から明らかなように、X_i の期待値 $E(X_i)$ は P_i に等しい。ゆえに、期待値の加法性☆から、期待単語数は以下のようになる。

$$1+E(\sum_{i=1}^{n_r-1} X_i) = 1+\sum_{i=1}^{n_r-1} E(X_i) = 1+\sum_{i=1}^{n_r-1} P_i \quad \blacksquare$$

A.2 単語 1-gram 確率の妥当性

以下の命題から式 (4) により定義される単語 1-gram 確率が、確率として正しく定義されていることが示される。

命題 2 生コーパス中における単語 w の表記の出現位置の集合を $O_1 = \{(i,k) \,|\, \boldsymbol{x}_{i+1}^k = w\}$ とし、単語 w の期待頻度が

☆ 確率変数 X と Y に対して $E(X) + E(Y) = E(X + Y)$

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

で定義されるとき

$$\sum_{w\in\mathcal{S}_r} f_r(w) = f_r(\cdot)$$

である。ここで $\mathcal{S}_r$ は C_r の全ての部分文字列の集合で、文字列の位置の違いを無視したものである☆。

証明 2

$$\begin{aligned}
&\sum_{w\in\mathcal{S}_r} f_r(w)\\
&= \sum_{w\in\mathcal{S}_r}\sum_{\boldsymbol{x}_{i+1}^{k}=w} P_i\left[\prod_{j=i+1}^{k-1}(1-P_j)\right]P_k\\
&= \sum_{0\le i<k\le n_r} P_i\left[\prod_{j=i+1}^{k-1}(1-P_j)\right]P_k\\
&= \sum_{k=1}^{n_r}\left(\sum_{i=0}^{k-1} P_i\left[\prod_{j=i+1}^{k-1}(1-P_j)\right]\right)P_k\\
&= \sum_{k=1}^{n_r} P_k \qquad \text{(図 3 参照)}\\
&= 1+\sum_{k=1}^{n_r-1} P_k \quad (\because P_{n_r}=1)\\
&= f_r(\cdot)\ \blacksquare
\end{aligned}$$

A.3 単語 n-gram 確率の妥当性

単語 n-gram の場合についても、式 (5) で定義される単語 n-gram 確率が確率として適正に定義されていることが示される。

命題 3 生コーパス中における n 単語の列 $\boldsymbol{w}_1^n$ の表記の出現位置の集合を $O_n = \{(i, e_1^n) \,|\, \boldsymbol{x}_{b_m}^{e_m} = w_m, 1 \le m \le n\}$ とし、

$$\begin{aligned}
&e_1^n = (e_1, e_2, \cdots, e_n)\\
&b_1 = i+1,\ b_{m+1} = e_m + 1 \qquad (6)
\end{aligned}$$

と表すとする。単語 n-gram $\boldsymbol{w}_1^n$ の期待頻度が

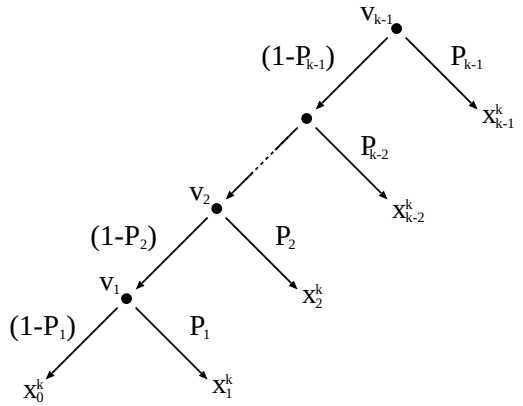
$$f_r(\boldsymbol{w}_1^n) = \sum_{(i,e_1^n)\in O_n} P_i\Big(\prod_{m=1}^{n}\Big[\prod_{j=b_m}^{e_m-1}(1-P_j)\Big]P_{e_m}\Big)$$

で定義されるとき、

$$\sum_{\boldsymbol{w}_1^n\in\mathcal{S}_r} f_r(\boldsymbol{w}_1^n) = f_r(\cdot) \qquad (7)$$

ある。ただし、$n \ge 2$では、文頭に限り $w_1, w_2, \cdots, w_{n-1}$ は ϵ を取ることができるとする。

☆ 確率論[8] において $\mathcal{S}_r$ を標本空間とし、その全ての部分集合を事象とすると、$P(w) = f_r(w)/f_r(\cdot)$ によって定義される P は確率測度となる。



位置 k に単語境界があるとの仮定のもと、葉に対応する文字列が単語である確率は根から葉までのエッジの確率の積となる。図を左下から調べていくと、ノード v_1 から葉までのエッジの確率の積の和は $(1-P_1)+P_1 = 1$ であり、ノード v_2 から葉までのエッジの確率の積の和は $(1-P_2)(1-P_1)+(1-P_2)P_1+P_2 = (1-P_2)+P_2 = 1$ である。同様の計算を繰り返すことで、根 v_{k-1} からの葉までの確率の積の和 $\sum_{i=0}^{k-1} P_i\left[\prod_{j=i+1}^{k-1}(1-P_j)\right]$ も 1 である。厳密な証明は、k に対する数学的帰納法により可能である。

図 3 証明 2 の概念図
Fig. 3 Product in Proof 2

証明 3 $f_r(\boldsymbol{w}_1^n)$ の定義式の和の計算の対象 ($\sum$ の中身) を形式的に、

$$T(i;\boldsymbol{e}_1^n) = P_i\Big(\prod_{m=1}^{n}\Big[\prod_{j=b_m}^{e_m-1}(1-P_j)\Big]P_{e_m}\Big)$$

とおく。ここで各 b_m は、式 (6) により、$i,\ \boldsymbol{e}_1^n$ から定まる。また、文頭にのみ ϵ を許すことを反映するため、n-gram の単語境界に相当するインデックス集合 E_n を次のように定義しておく。

$$\begin{aligned}
E_n = \{&(i_1, i_2, \cdots, i_n) \,|\, 0 = i_1 = i_2 \cdots = i_l,\\
&i_l < i_{l+1} < i_{l+2} < \cdots < i_n \le n_r,\\
&0 \le l \le n-1\}
\end{aligned}$$

以下、n に関する数学的帰納法で示す。$n = 1$ の時は、命題 2 より成り立つ。$n-1$ で式 (7) が成り立つと仮定すると、以下が成り立つ。

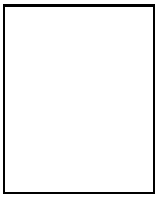
$$
\begin{aligned}
&\sum_{\boldsymbol{w}_1^n \in \mathcal{S}_r} f_r(\boldsymbol{w}_1^n) \\
&= \sum_{\boldsymbol{w}_1^n \in \mathcal{S}_r} \sum_{\boldsymbol{x}_{b_1}^{e_1}=w_1} \sum_{\boldsymbol{x}_{b_2}^{e_2}=w_2} \cdots \sum_{\boldsymbol{x}_{b_n}^{e_n}=w_n} T(i;\boldsymbol{e}_1^n) \\
&= \sum_{(i,e_1,e_2,\cdots,e_n)\in E_{n+1}} T(i;\boldsymbol{e}_1^n) \\
&= \sum_{(e_1,e_2,\cdots,e_n)\in E_n} \left(\sum_{i=0}^{e_1} P_i \left[\prod_{j=i+1}^{e_1-1} (1-P_j) \right] \right) T(e_1;\boldsymbol{e}_2^n) \\
&= \sum_{(e_1,e_2,\cdots,e_n)\in E_n} T(e_1;\boldsymbol{e}_2^n) \quad (\because \text{証明 2 と同様}) \\
&= \sum_{\boldsymbol{w}_2^n \in \mathcal{S}_r} f_r(\boldsymbol{w}_2^n) \quad (\because \text{帰納法の仮定}) \\
&= f_r(\cdot)
\end{aligned}
$$

よって数学的帰納法により $n = 1, 2, \cdots$ について式 (7) が示された。 ■

参 考 文 献

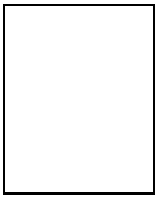
1) Jelinek, F.: Self-Organized Language Modeling for Speech Recognition, Technical report, IBM T. J. Watson Research Center (1985).
2) Nagata, M.: Context-Based Spelling Correction for Japanese OCR, *Proceedings of the 16th International Conference on Computational Linguistics* (1996).
3) 永田昌明: 統計的言語モデルと N-best 探索を用いた日本語形態素解析法, 情報処理学会論文誌, Vol. 40, No. 9, pp. 3420–3431 (1999).
4) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 201–207 (1994).
5) Manber, U. and Myers, G.: Suffix Arrays: A New Method for On-Line String Searches, *SIAM J. Comput*, Vol. 22, No. 5, pp. 935–948 (1993).
6) Brown, P. F., Pietra, S. A. D. and Mercer, R. L.: An Estimate of an Upper Bound for the Entropy of English, *Computational Linguistics*, Vol. 18, No. 1, pp. 31–40 (1992).
7) Clarkson, P.R. and Robinson, A.J.: Language Model Adaptation Using Mixtures And An Exponentially Decaying Cache, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 799–802 (1997).
8) Williams, D.: *Probability With Martingales*, Cambridge Mathematical Textbooks (1991).

(平成？年？月？日受付)
(平成？年？月？日採録)



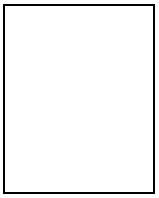
森　信介（正会員）

1998 年京都大学大学院博士後期課程修了．同年日本アイ・ビー・エム (株) 入社．東京基礎研究所において計算言語学の研究に従事．工学博士．1997 年本学会山下記念研究賞受賞．言語処理学会会員．



宅間　大介（正会員）

2003 年東京大学大学院数理科学研究科修了．同年日本アイ・ビー・エム (株) 入社．東京基礎研究所において自然言語処理の研究に従事．



倉田　岳人（正会員）

2004 年東京大学大学院情報理工学系研究科修了．同年日本アイ・ビー・エム (株) 入社．東京基礎研究所において音声言語処理の研究に従事．音響学会会員．