社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

単語リストと生コーパスによる確率的言語モデルの分野適応

森　信介†

† 日本アイ・ビー・エム株式会社東京基礎研究所

あらまし　本論文では、単語リストと生コーパスが利用可能な状況における確率的言語モデルの分野適応について述べる。このような状況の下での一般的な対処は、単語リストを語彙に加えた自動単語分割システムによる生コーパスの自動単語分割の出力文を可能な限り人手で修正し、パラメータ推定に利用することである。しかしながら、文単位での修正では、正確な単語分割が容易でない箇所が含まれることになり、作業効率の著しい低下を招く。加えて、文単位で順に修正していくことが、限られた作業量を割り当てる最良の方法であるかということも疑問である。本論文では、コーパスの修正を単語単位とし、修正箇所を単語リストで与えられる適応分野に特有の単語に集中することを提案する。これにより、上述の困難を回避し、適応分野に特有の単語の統計的な振る舞いを捕捉するという、適応分野のコーパスを利用する本来の目的にのみコーパス修正の作業を集中することが可能となる。実験では、自動単語分割の結果の人手による修正の程度や方法を複数用意し、その結果得られるコーパスから推定された確率的言語モデルの予測力やそれに基づく仮名漢字変換の精度を計算した。この結果、適応分野に特有の語彙の出現箇所に修正のコストを集中することにより、少ない作業量で効率良く確率的言語モデルを分野適応できることが分かった。
キーワード　仮名漢字変換, 音声認識, 言語モデル, コーパス

Language Model Adaptation with a Word List and a Raw Corpus

Shinsuke MORI†

† IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.

Abstract　In this paper, we discuss stochastic language model adaptation methods given a word list and a raw corpus. In this situation, a general method is to segment the raw corpus by a word segmenter equipped with a word list, correct the output sentences annotated with word boundary information by hand, and build a model from the segmented corpus. In this sentence-by-sentence error correction method, however, the annotator encounters difficult points and this results in a decrease of the productivity. In addition, it is not sure that sentence-by-sentence error correction from the beginning is the best way to dispense a limited work force. In this paper, we propose to take a word as a correction unit and concentrically correct the positions in which words in the list appear. This method allows us to avoid the above difficulty and go straight to capture the statistical behavior of specific words in the application field. In the experiments, we used a variety of methods to prepare a segmented corpus and compared the language models from the corpora in predictive power and Kana-kanji conversion accuracy. The results showed that concentrating on the error correction around the words in the list, we can build a better language model with less effort.
Key words　Kana-Kanji Convertor, Speech Recognition, Language Model, Corpus

1. まえがき

確率的言語モデルは、文字列を出力とする言語処理において幅広く用いられている。音声認識システム [1] の多くが、解選択において、音響モデルとともに確率的言語モデルを参照する。文字誤り訂正や仮名漢字変換においても、確率的言語モデルを用いる方法が提案されている [2] [3]。

多くの確率的言語モデルは単語や単語列の頻度に基づいており、これは正しく単語に分割された例文 (単語分割済みコーパス) に対して計数される。この単語分割済みコーパスは、一般的と考えられる分野においては既に利用可能となっているが、新たに確率的言語モデルを用いる分野 (医療現場やコールセンターでの音声認識など) の言語資源としては、単語に分割されていない例文 (生コーパス) やその分野の単語リストのみが利

用可能であることが多い。このような状況の下での一般的な対処は、単語リストを語彙に加えた自動単語分割システム [4] により生コーパスの各文を単語に分割し、可能な限り多くの文の分割結果を人手で修正し、自動解析の結果と合わせて単語分割済みコーパスとすることである。

単語分割の修正量は、多ければ多いほど統計結果の信頼性が増し、確率的言語モデルの能力は高くなる。しかしながら、単語分割の修正作業は非常にコストや時間がかかるので、コーパスの一部分を修正の対象とし、残りの部分に関しては自動分割の結果をそのまま用いるということがしばしば行なわれる。文単位で修正する場合には、文法の専門家でさえも正確な単語分割が容易でない機能語列などの箇所が必然的に含まれることになるが、このようなの箇所での分割方針を作業者に徹底することは非常に困難であり、作業効率の著しい低下を招く。加えて、文単位で順に修正していくことが、限られた作業量を割り当てる最良の方法であるかということも疑問である [5] 。

本論文では、コーパスの修正を一文単位ではなく単語単位とし、修正箇所を単語リストなどで与えられる適応分野に特有の単語の周辺に集中することを提案する。これにより、上述のような困難を回避することが可能となり、さらに、適応分野に特有の単語の統計的な振る舞いを捕捉するという、適応分野のコーパスを利用する本来の目的にのみコーパス修正の作業を集中することが可能となる。このようにして得られるコーパスは一部分の単語境界情報のみが正確な文を含む。このようなコーパスから有限の語彙に対して確率的言語モデルを推定するために、本論文では、生コーパスから無限の語彙に対して確率的言語モデルを推定する方法 [6] を語彙が有限の場合に応用する方法について述べる。

実験では、生コーパスの単語境界の人手による修正の程度や方法を複数用意し、その結果得られるコーパスから推定される確率的言語モデルの予測力やそれに基づく仮名漢字変換の精度を計算した。実験の結果、単語リストの各単語に対して 2 箇所の出現のみを人手でマークする方法では、単語数の割合にして生コーパス全体の 5.22%のみの修正により、単語数の割合にして生コーパス全体の 45.00%の文を文単位で修正した場合と同程度の仮名漢字変換の精度を達成することができた。また、単語リストの各単語に対して全ての出現箇所を人手でチェックすることで、コーパス全体に対して自動分割の結果を人手で修正するのと同程度の予測力と変換精度を達成できた。この結果から、適応分野に特有の語彙の出現箇所に修正のコストを集中することにより、少ない作業量で効率良く確率的言語モデルを分野適応できるといえる。

2. 確率的言語モデル

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。自動単語分割は解析系の一例であり、文字列が与えられると尤度が最大になる単語の列を計算する。認識系の代表例の音声認識では、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

日本語の確率的言語モデルは、日本語のアルファベット列 $\mathcal{X}^*$ 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、確率値をすべてのアルファベット列に渡って合計すると 1 以下になる必要がある。

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

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 $(i \leqq 0)$ 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 UT 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 UT 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。

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

この式の中の x_i $(i \leqq 0)$ 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。

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

2.2 応　　用

確率的言語モデルの応用は、自然言語認識と自然言語解析に大別できる。

認識系の代表例は、音声認識である。確率的言語モデルを用いる音声認識では、音響特徴量の列 $\boldsymbol{s}$ を入力とし、以下の式のように、確率最大となる単語列 $\boldsymbol{w}$ を出力する。

$$\hat{\boldsymbol{w}} = \mathop{\mathbf{argmax}}_{\boldsymbol{w} \in \mathcal{X}^*} P(\boldsymbol{s}|\boldsymbol{w}) P(\boldsymbol{w})$$

この式における $P(\boldsymbol{w})$ が確率的言語モデルである。確率的言語モデルの予測力と認識系の精度との関係は、解析的に導出できるような確固とした関係ではない。音声認識に対して実験的に得られた関係として、西村ら [7] は相関係数 0.6 を報告している。

解析系の代表例は、単語分割 (と品詞付与) である。確率的言語モデルによる単語分割 [4] は、以下の式が示すように、ある言語の文字列 $\boldsymbol{x}$ を入力とし、生成確率が最大となる単語列 $\boldsymbol{w}$ を出力する。

$$\hat{\boldsymbol{w}} = \mathop{\mathbf{argmax}}_{\boldsymbol{w} = \boldsymbol{x}} P(\boldsymbol{w})$$

ここで $\boldsymbol{w} = \boldsymbol{x}$ は、単語列 $\boldsymbol{w}$ を文字列とみなした場合、入力 $\boldsymbol{x}$ と等しいことを表す。

3. 単語リストと生コーパスによる分野適応

この節では、適応対象の分野の単語リストと、それらが出現する生コーパスが利用可能である場合に、それらから確率的言語モデルを推定する方法を述べる。

3.1 確率的単語分割コーパスからの単語 n-gram 確率の推定

単語分割済みコーパスは、各文字間に単語境界が存在するか否かの情報が人手により付与されている。生コーパスはこの情報を持たないが、各文字間に単語境界が存在する確率を付与し、それによって生コーパスを確率的に単語に分割されたコーパス(確率的単語分割コーパス)とみなすことにより、無限の語彙に対する単語 n-gram 頻度や単語 n-gram 確率を計算する方法が提案されている [6]。以下では、この方法を説明する。

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。

単語 0-gram 頻度　確率的単語分割コーパスにおける単語 0-gram 頻度 $f_r(\cdot)$ は、そのコーパス中の期待単語数であり、以下のように定義される。

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

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。
（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、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。

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

単語 n-gram 頻度　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 $(\boldsymbol{x}_{b_m}^{e_m}=w_m,\ 1\leqq\forall m\leqq n;\ e_m+1=b_{m+1},\ 1\leqq\forall m\leqq n-1;\ b_1=i+1;\ e_n=k)$ 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。
（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。

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

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

とした。

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、確率的単語分割コーパスにおける単語 1-gram 確率を以下のように定義する。

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

単語 n-gram 確率　決定的に単語に分割されたコーパスからの単語 n-gram 確率の最尤推定の場合と同様に、確率的単語分割コーパスにおける単語 n-gram 確率を以下のように定義する。

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

3.2 有限の語彙に対する確率的単語分割コーパスからの単語 n-gram 確率の推定

確率的言語モデルを用いる音声認識においては、認識される語彙には発音が付与されている必要がある。また、確率的言語モデルを用いる仮名漢字変換においてもキー入力列が付与されている表記 (単語) のみが変換結果として出現し得る。このように、現実的な応用では有限の語彙に対する確率的言語モデルを構築する必要がある。分野適応において単語リストが与えられている場合には、一般コーパスから得られる語彙と対象分野の単語リストを語彙として、対象分野の生コーパスから確率的言語モデルを構築する。この際に、未知語モデルを含めて確率的言語モデルの条件を満たすためには、未知語記号を含む単語 n-gram 確率を正しく定義する必要がある。

単語分割済みコーパスにおいては、まず語彙 $\mathcal{W}_k$ に属さない単語をコーパスの全ての出現場所において未知語記号 UT に置き換え、その上で未知語記号を語彙に含まれる単語と同様に扱って頻度計算を行なう。決定的に単語に分割されていない確率的単語分割コーパスに対しては、この方法を採ることができない。また、語彙以外の任意の文字列に対する単語 n-gram 頻度を計数しその和を計算する方法も考えられる。語彙以外の任意の文字列は、実際には無限集合ではなく、コーパスの部分文字列のみを対象とすれば十分であるが、これは非常に大きな数となるので、この計算方法も現実的ではない。しかしながら、単語 n-gram 頻度の以下の性質を用いることにより、確率的単語分割コーパスに対しても未知語記号を含む単語 n-gram 頻度を容易に計算することができる(注1)。

$$\begin{cases} f_r(\boldsymbol{w}_u \mathtt{UT} \boldsymbol{w}_v) = \displaystyle\sum_{w \in \mathcal{X}^+ - \mathcal{W}_k} f_r(\boldsymbol{w}_u w \boldsymbol{w}_v) \\ \displaystyle\sum_{w \in \mathcal{X}^+} f_r(\boldsymbol{w}_u w \boldsymbol{w}_v) = \sum_{w \in \mathcal{X}^+ - \mathcal{W}_k} f_r(\boldsymbol{w}_u w \boldsymbol{w}_v) + \sum_{w \in \mathcal{W}_k} f_r(\boldsymbol{w}_u w \boldsymbol{w}_v) \end{cases}$$

$$\Rightarrow f_r(\boldsymbol{w}_u \mathtt{UT} \boldsymbol{w}_v) = \sum_{w \in \mathcal{X}^+} f_r(\boldsymbol{w}_u w \boldsymbol{w}_v) - \sum_{w \in \mathcal{W}_k} f_r(\boldsymbol{w}_u w \boldsymbol{w}_v) \quad (3)$$

ここで $\boldsymbol{w}_u, \boldsymbol{w}_v \in \{\mathcal{W}_k \cup \{\mathtt{UT}\}\}^*$ は語彙と未知語記号からなる長さ 0 以上の任意の列である。

未知語記号の 1-gram 頻度　確率的単語分割コーパスにおける未知語記号の 1-gram 頻度 $f_r(\mathtt{UT})$ は、コーパスに対して計数した単語 1-gram 頻度と単語 0-gram 頻度に対して成り立つ関係

$$f_r(\cdot) = \sum_{w \in \mathcal{X}^+} f_r(w)$$

と式 (3) において $\boldsymbol{w}_u = \boldsymbol{w}_v = \varepsilon$ (ε は空列を表す) とすることで得られる等式

$$f_r(\mathtt{UT}) = \sum_{w \in \mathcal{X}^+} f_r(w) - \sum_{w \in \mathcal{W}_k} f_r(w)$$

から以下のように、単語 0-gram 頻度と語彙に対する単語 1-gram 頻度の和から計算される。

$$f_r(\mathtt{UT}) = f_r(\cdot) - \sum_{w \in \mathcal{W}_k} f_r(w)$$

未知語記号を含む 2-gram 頻度　任意の単語 $w_1 \in \mathcal{W}_k$ と未知語記号からなる列の確率的単語分割コーパスにおける頻度 $f_r(w_1\mathtt{UT})$ はコーパスに対して計数した単語 2-gram 頻度と単語 1-gram 頻度に対して成り立つ関係

$$f_r(w_1) = \sum_{w \in \mathcal{X}^+} f_r(w_1 w), \quad \forall w_1 \in \mathcal{W}_k \cup \{\mathtt{UT}\}$$

と式 (3) において $\boldsymbol{w}_u = w_1,\ \boldsymbol{w}_v = \varepsilon$ とすることで得られる等式

$$f_r(w_1\mathtt{UT}) = \sum_{w \in \mathcal{X}^+} f_r(w_1 w) - \sum_{w \in \mathcal{W}_k} f_r(w_1 w)$$

から以下のように単語 1-gram 頻度と既知語に対する単語 2-gram 頻度の和から計算される。

$$f_r(w_1\mathtt{UT}) = f_r(w_1) - \sum_{w \in \mathcal{W}_k} f_r(w_1 w)$$

同様に未知語記号と任意の単語 $w_2 \in \mathcal{W}_k$ からなる列の確率的単語分割コーパスにおける頻度 $f_r(\mathtt{UT}w_2)$ は、以下のように計算される。

$$f_r(\mathtt{UT}w_2) = f_r(w_2) - \sum_{w \in \mathcal{W}_k} f_r(w w_2)$$

さらに未知語記号の 2-gram 頻度 $f_r(\mathtt{UT}\,\mathtt{UT})$ は

$$f_r(\cdot) = \sum_{w_1 \in \mathcal{X}^+} \sum_{w_2 \in \mathcal{X}^+} f_r(w_1 w_2)$$

を用いることで以下のように計算される。

$$f_r(\mathtt{UT}\,\mathtt{UT}) = f_r(\cdot) - \sum_{w_1 \in \mathcal{W}_k} f_r(w_1\mathtt{UT}) - \sum_{w_2 \in \mathcal{W}_k} f_r(\mathtt{UT}w_2) - \sum_{(w_1 w_2) \in \mathcal{W}_k \times \mathcal{W}_k} f_r(w_1 w_2)$$

未知語記号を含む n-gram 頻度 ($n \geqq 3$)　未知語記号を含む一般の n-gram 頻度も 2-gram 頻度の場合と同様に計算することが可能である。

未知語記号を含む n-gram 確率 ($n \geqq 1$)　未知語記号を含まない場合と同様に、確率的単語 n-gram 頻度を確率的単語 $(n-1)$-gram 頻度で割ることで未知語記号を含む単語 n-gram 確率が得られる。

以上から、語彙を有限とし未知語記号を仮定する場合でも、確率的単語分割コーパスに対する単語 n-gram 確率を推定できることが示された。

4. 生コーパスの利用方法

適応対象の分野のコーパスは、その分野の言語的な特徴を的確に捉えるために重要である。この利用方法としては、以下の 3 つが代表的である。

- 未知語の取り出し

  生コーパスに対して文字 n-gram の統計などを取り、ある程度の頻度があり、かつ前後の文字の分布にばらつきがある文字列などを単語候補として抽出する [8] [9] この結果得られた単語候補は、人手でチェックされる。さらに確率的言語モデルの応用に応じて読みの付与などを行なう。

(注1)：正確には複数の未知語記号を含む n-gram に対するの記述も必要であるが、式が繁雑になるためここでは 1 つの未知語記号のみを含む場合のみ記述した。

	女性エコノミスト、キャ	サリン	・カミリさんなどは「今
	ローム・デーヴィッド・	サリン	ジャーは 20 世紀アメリカ
○	に次ぐおぞましい地下鉄	サリン	事件、長い不況に追い打
○	理が始まった中川被告は	サリン	生成を認めながら「目的
	っているのを知りながら	サリン	流出を阻止する義務を怠

図 1　単語リストの KWIC による単語境界情報付与の例
Fig. 1　An example of corpus annotation by KWIC.

- 自動分割による単語分割済みコーパス

  自動単語分割システム [4] により自動的に単語に分割し、これを単語分割済みコーパスとして利用する。単語分割システムは、人手により正しく単語に分割された一般的なコーパスから構築されるので、適応対象の分野の文に対する解析精度は必ずしも高くない。特に、適応分野に特有の単語や表現の周辺で分割を誤る傾向がある。しかしながら、適応対象の分野の単語分割済みコーパスは、多少の誤りが含まれていても、確率的言語モデルの構築に有用であることが知られている。

- 人手による単語分割済みコーパス

  理想的には、適応対象の分野のコーパスの全ての文が正しく (単語分割の指針に沿って) 単語に分割されていることが望ましい。このときに確率的言語モデルの能力は最高になる。

単語分割の修正作業は、作業量を増やせば増やすほど確率的言語モデルの能力は高くなる。現実には、単語分割の修正作業は非常にコストや時間がかかるので、コーパスの一部分を修正の対象とし、残りの部分に関しては自動分割の結果をそのまま用いるということが行なわれる。しかし、この方法が有限の作業量を割り当てる最良の方法であるか疑問が残る。

単語分割の修正作業は、コーパスに単語境界の情報を付与することである。単語境界の情報の最小単位は各文字の間に単語境界があるか否かである。一般的に行なわれる文単位の修正作業は、文頭から順に各文字の間の単語境界情報が正しいかを確認し、必要に応じて修正することである。我々は修正作業を、単語リストなどで与えられる適応分野に特有の単語の周辺に集中することを提案する。具体的には、図 1 に示されるように、単語リストに含まれる語 (例では「サリン」) の対象分野のコーパスでの出現位置を KWIC(Key Word In Context) 形式で提示し、注目している文字列が各文脈において単語として用いられているかのチェックをする。単語として用いられている箇所にマーク (図中では「○」) を付け、それ以外の箇所では何もしないという作業を行なう。各単語についてマークする箇所の数を制限するというこも有効であろう。そうすれば、判断の難しい箇所で時間を浪費することを避けることもできる。

5. 評　　価

適応分野の生コーパスの利用方法について比較検討するために、生コーパスに対する人手による単語境界情報の付与の程度や方法を複数用意し、その結果得られるコーパスから推定される確率的言語モデルの予測力やそれに基づく仮名漢字変換の精度を計算した。

5.1 実験条件

実験には、一般的な分野のコーパスとして会話辞典の例文と、適応対象として新聞記事を用いた (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、適応分野のコーパスは、主として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、比較対象としての理想的な状況を実現するためである。適応分野の単語リストは、適応分野のコーパスにのみ出現する 21,855 単語からなる。

表 1　コーパス
Table 1　Corpora.

用途	分野	文数	単語数	文字数
学習	会話	14,754	187,658	254,436
学習	新聞	20,700	625,761	917,830
テスト	会話	1,639	21,105	28,655
テスト	新聞	2,300	68,566	100,091

基本となる確率的言語モデルは以下の通りである。

Base　単語分割済みの一般分野のコーパスから単語 2-gram モデルを構築した。既知語の数は 5,112 語である。適応分野の単語リストは、未知語モデルにおいて出現確率を嵩上げされ、未知語に対して出現しやすくなる外部辞書語 [10] として利用する。適応分野のコーパスは利用しない。

この確率的言語モデルの同一分野のテストコーパスに対するクロスエントロピーは 4.509 であり、テストセットパープレキシティーは 64.28 であった(注2)。

実験に利用した自動単語分割システムは、この言語モデルに基づいており、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)。同一分野のテストコーパスに対する単語境界の推定精度は 98.26%であった(注3)。

後述する実験において、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。単語境界の推定結果の信頼度には、自動単語分割システムの精度 $\alpha = 98.26\%$ を利用した。すなわち、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 $P_i = \alpha$ とし、単語境界でないと判定された点では $P_i = 1 - \alpha$ とした。

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、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 [11]。

(注2)：テストセットパープレキシティーは平均単語長に影響されるので、異なるテストコーパスに対する結果との比較には適さない。
(注3)：対象分野のテストコーパス (表 1 参照) に対する単語境界の推定精度は 89.25%であった。

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

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。

さらに確率的言語モデルの応用として仮名漢字変換 [3] を採用し、文単位で一括変換した場合の第一候補の変換精度を計算した(注4)。これは、音声認識において、音響モデルの誤りの影響を排した場合と考えることもできる。

5.3 適応分野の生コーパスの利用方法

適応分野の生コーパスの利用方法について比較検討するために、生コーパスの自動分割結果に対する単語境界情報の人手による修正の程度や方法として、以下の 6 つを準備した。

Auto 適応分野の生コーパスを自動的に単語分割し、その結果をそのまま用いる。これは、自動分割システムにより単語境界と判定された点では $P_i = 1$ とし、単語境界でないと判定された点では $P_i = 0$ とする確率的単語分割コーパスと等価である。

Raw 適応分野の生コーパスを確率的単語分割コーパスとして用いる。すなわち、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 $P_i = \alpha$ とし、単語境界でないと判定された点では $P_i = 1-\alpha$ とした。

Well-done 適応分野の生コーパスの全文を人手により正しく単語に分割し、これを **Auto** と同様に決定的に単語に分割されたコーパスとして利用する。

45%-done 適応分野の生コーパスの最初から 281,398 単語目まで (45.00%) を人手により正しく単語に分割し、その残りを自動的に単語分割した。これを **Auto** と同様に決定的に単語に分割されたコーパスとして利用する。

Medium まず **Raw** と同様に単語境界確率を設定する。さらに、単語リストに含まれる文字列が生コーパス中に単語として出現している全ての箇所において、その文字列内の単語境界確率を 0 とし、その文字列の直前と直後の単語境界確率を 1 とする。これは、生コーパスに対する単語リストに含まれる文字列の KWIC を見て、その文字列が単語として出現している場合にマークをつける作業をした結果に相当する。チェック箇所は単語数でのべ 138,483 箇所 (22.13%) である。

Rare まず **Raw** と同様に単語境界確率を設定する。さらに、単語リストに含まれる文字列が生コーパス中に単語として出現している最初の 2 箇所において、その文字列内の単語境界確率を 0 とし、その前後の単語境界確率を 1 とする。これは、生コーパスに対する単語リストに含まれる文字列の KWIC を見て、その文字列が単語として出現している場合にマークをつける作業を各文字列に対して 2 つのマークがつくまで行なった結果に相当する。チェック箇所は単語数でのべ 32,643 箇所 (5.22%) である。

(注4)：評価基準は文献 [3] と同一である。

表 2 各モデルの予測精度と仮名漢字変換の精度

Table 2 Predictive powers and conversion accuracies of each model.

モデル	生コーパスの利用方法	H	PP	再現率	適合率
Base	–	7.558	1938	62.74%	72.34%
Auto	自動分割	6.618	755.7	80.52%	85.24%
Raw	確率分割	6.276	536.5	84.70%	87.85%
Rare	部分修正	6.133	465.2	86.57%	89.24%
Medium	部分修正	5.889	364.2	88.34%	90.50%
45%-done	部分修正	6.049	427.4	86.56%	89.32%
Well-done	完全修正	5.858	353.1	88.90%	90.90%

以上のようにして得られる適応分野のコーパスから、**Base** モデルの既知語と単語リストに含まれる単語を語彙として単語 1-gram 確率と単語 2-gram 確率を計算し、**Base** モデルと補間して適応分野のための確率的言語モデルを構築した。

5.4 評　　価

各モデルの予測力と仮名漢字変換の精度を表 2 に示す。**Base** とコーパス修正のコストがない **Auto** の **Raw** の結果から、適応分野のコーパスは可能な限り収集し、言語モデルの推定に利用するのがよいといえる。利用方法においては、**Auto** と **Raw** の結果の比較から、誤りを含む自動分割結果を 100%信頼してそのまま用いるのではなく、単語境界か否かの判定結果を割り引いて確率的単語分割コーパスとして用いるほうがよいといえる。

コーパス修正のコストがない **Raw** の予測力や変換精度は、自動分割の結果を人手で完全に修正した場合の予測力と変換精度 **Well-done** に対してかなり低く、修正のコストを払うことで改善する余地があることが分かる。自動分割結果の修正は文単位で行なうのが一般的であるが、単語リストに含まれる単語が出現する箇所に限定して、文の一部分のみをチェックする場合の結果が **Rare** と **Medium** である。単語リストの各単語に対して 2 箇所の出現のみを人手でマークする **Rare** では、単語数の割合にして 5.22%のみがマークの対象になるが、仮名漢字変換の精度はコーパスの最初から順に 45.00%の単語をチェックする **45%-done** の精度にほぼ等しい。単語リストの各単語に対して全ての出現箇所を人手でチェックする **Medium** と自動分割の結果を人手で完全に修正する **Well-done** の予測力と変換精度は同程度である。この結果から、適応分野に特有の語彙の出現箇所に修正のコストを集中すれば、コーパス全体の約 22.13%の単語のみのチェックで、予測力においても、仮名漢字変換の精度においても、コーパス全体の分割結果を人手で修正したコーパスを利用する場合にかなり近い性能を達成することが可能であるといえる。

文単位で分割結果を修正する方法と、特定の文字列の KWIC を見てそれが各文脈で単語として用いられているかをマークするするのは、1 単語あたりのチェックに要するコストが等しいとは限らない。しかしながら、**Rare** と **45%-done** のチェック対象の単語数には 9 倍の差がある。特定の単語の KWIC における 1 箇所のチェックが、注目単語の前後 4 単語の修正を含めた分割修正 (合計 9 単語) に相当する時間を要するとは思えず、**Rare** と **45%-done** の総修正コストの順序関係は変わらないであろう。加えて、文全体に対して分割結果の修正を行なう場合には、主に活用語尾や助詞や助動詞からなる、文法の専門家でさえも正確な単語分割が容易でない箇所が含まれることになるが、このような正確な単語分割が困難な機能語などの列の分割方針を作業者に徹底することは非常に困難である。単語リストに含まれる単語のみをチェック対象にすれば、このような困難を回避することが可能となり、さらに、適応分野に特有の単語の統計的な振る舞いを捕捉するという、適応分野のコーパスを利用する本来の目的のみにコーパス修正のコストを集中することが可能となる。以上のことから、適応分野に特有の語彙の出現箇所に修正のコストを集中し、この結果得られる部分的に修正されたコーパスを確率的単語分割コーパスとみなして確率的言語モデルを構築することにより、音声認識や仮名漢字変換などの適応対象の分野における精度をより低いコストでより短時間で向上させることが可能となる。

6. おわりに

本論文では、単語リストと生コーパスが利用可能であることを前提として、確率的言語モデルを分野適応する際に、コーパスの修正の程度や方法について比較検討を行なった。予測力や仮名漢字変換の精度を評価基準とする実験の結果、生コーパスの自動単語分割の結果の人手による修正を単語リストに含まれる単語が出現する箇所に限ることで、確率的言語モデルの適応分野における性能をより効率よく向上させることが可能となることが分かった。

文　献

[1] F. Jelinek: “Self-organized language modeling for speech recognition”, Technical report, IBM T. J. Watson Research Center (1985).

[2] M. Nagata: “Context-based spelling correction for Japanese OCR”, Proceedings of the 16th International Conference on Computational Linguistics (1996).

[3] 森, 土屋, 山地, 長尾：“確率的モデルによる仮名漢字変換”, 情報処理学会論文誌, **40**, 7, pp. 2946–2953 (1999).

[4] M. Nagata: “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] D. Hakkani-Tür, G. Tur, M. Rahim and G. Riccardi: “Unsupervised and active learning in automatic speech recognition”, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (2004).

[6] S. Mori and D. TAKUMA: “Word n-gram probability estimation from a Japanese raw corpus”, International Conference on Speech and Language Processing (2004).

[7] 西村, 伊東, 山崎, 荻野：“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”, 情報処理学会研究報告, 第 SLP15 巻, pp. 27–34 (1997).

[8] 森, 長尾：“n グラム統計によるコーパスからの未知語抽出”, 情報処理学会研究報告 (1995).

[9] 中渡瀬：“統計的手法による単語の切り出しについて”, 電子情報通信学会技術研究会報告, pp. 69–74 (1995).

[10] 森, 山地：“日本語の情報量の上限の推定”, 情報処理学会論文誌, **38**, 11, pp. 2191–2199 (1997).

[11] P. F. Brown, S. A. D. Pietra and R. L. Mercer: “An estimate of an upper bound for the entropy of English”, Computational Linguistics, **18**, 1, pp. 31–40 (1992).