自然言語処理 116－3
（1996. 11. 18）

単語頻度の期待値に基づく未知語の自動収集

永田昌明
NTT 情報通信研究所 知的通信処理研究部

本稿では、単語の出現頻度の期待値に基づいて日本語テキストから未知語を収集する方法を報告する。まず頑健なN-best単語分割プログラムを用いて日本語テキスト中の単語候補の頻度の期待値を求め、次に頻度の期待値が予め決めた閾値以下である単語候補を取り除くことにより未知語候補の集合を得る。本手法における単語頻度の期待値の計算法は、一般化前向き後向きアルゴリズムの近似計算法に相当する。人手により単語分割された470万語のEDRコーパスで単語分割プログラムを訓練し、未知語率2.1%の試験文(1000文)でテストしたところ、未知語収集の精度は再現率43.7%，適合率52.3%であった。

Automatic Extraction of New Words from Japanese Texts based on Expected Word Frequency

Masaaki NAGATA
NTT Information and Communication Systems Laboratories
nagata@nttnly.isl.ntt.co.jp

We present a novel new word extraction method from Japanese texts based on expected word frequencies. First, we compute expected word frequencies from Japanese texts using a robust stochastic N-best word segmenter. We then extract new words by filtering out erroneous word hypotheses whose expected word frequencies are lower than the predefined threshold. The method is derived from an approximation of the generalized version of the Forward-Backward algorithm. When the Japanese word segmenter is trained on a 4.7 million word segmented corpus and tested on 1000 sentences whose out-of-vocabulary rate is 2.1%, the accuracy of the new word extraction method is 43.7% recall and 52.3% precision.

1 はじめに

コンピュータによる日本語の形態素解析は、英語のヒアリングに似たところがある。「make や get など基本単語 3000 語の用法をマスターすれば日常英会話は OK!」なんて真っ赤な嘘であることを多くの人は経験的に知っているに違いない。確かに出現頻度順上位 3000 語を覚えていれば、日常会話に出現する異なり単語の大部分はカバーできるだろう。しかし、たった一つの単語を知らないために文の意味が皆目分からなくなることは多い。知らない単語が一つあると文のセグメンテーションに失敗し、前後の単語も聞きとれなくなるからである。

語学の基本は単語力である。同様に、日本語の形態素解析プログラムの能力を決めるのは、対象テキストに出現する語彙が辞書に登録されている割合である。ここでは、対象テキスト中の辞書に登録されていない単語の割合である未知語率 (OOV rate: Out-Of-Vocabulary rate) を尺度に用いる。

本報告では、未知語率が小さい辞書を作成するために、日本語のテキストから未知語を自動的に収集する方法について述べる。この研究の目標は、新しい領域やアプリケーションのための辞書の作成、および、対象領域における新語の収集や死語の削除のような辞書の保守を自動化することである。この形態素解析を用いたアプリケーションの一例については [Nagata, 1996b] を参照して頂きたい。

2 日本語形態素解析における未知語問題

最初に日本語の形態素解析における未知語処理の問題点を概観する。仮に "ペンシルバニア大学は ENIAC の 50 周年を祝う。" という入力文において "ペンシルバニア" と "ENIAC" を辞書未登録語とする。この文の 3 通りの形態素解析候補を図 1に示す。ここで <UNK> は未知語を表す。図 1において、文字列 "ENIAC" は正しく未知語と同定されているが、文字列 "ペンシルバニ大学" の単語分割には曖昧性がある。

第 1 候補では "大学" が辞書に登録されているのでシステムは "ペンシルバニア" を未知語と推定している。第 2 候補では "ペンシル" (鉛筆) が辞書に登録されているので "バニア大学" を未知語と推定している。"スタンフォード大学" や "ケンブリッジ大学" は辞書に登録されているのであり得ない解釈ではない。第 3 候補では "ペンシル" と "大学" が辞書に登録されているので "バニア" を未知語と推定している。

入力文が未知語を含む場合、図 1の "ペンシルバニア", "バニア大学", "バニア" のように互いに重なり合う単語候補が得られる。未知語処理の大きな課題の一つは、互いに重なり合う単語候補の中で最も尤もらしい候補を選択する基準を確立することである。しかしそれ以前に、図 1に示すようなかなり尤もらしい形態素解析候補の集合を得ること自体が難しい課題である。入力文が未知語を含む場合、一つの未知語の同定誤りが周辺の単語のセグメンテーション誤りを引き起こす。そこで以下では、まず未知語があっても頑健に動作する形態素解析法について述べ、次に未知語の収集法について述べる。

3 統計的言語モデル

3.1 単語分割モデル

日本語文字列 $C = c_1 c_2 \ldots c_m$ が単語列 $W = w_1 w_2 \ldots w_n$ に分割され、その品詞列が $T = t_1 t_2 \ldots t_n$ とする。日本語の形態素解析は、与えられた文字列に対する単語列と品詞列の同時確率 $P(W,T|C)$ を最大化する単語列と品詞列の組を求める問題である。文字列 C は共通なので $P(W,T)$ を最大化すればよい。

$$(\hat{W},\hat{T}) = \arg\max_{W,T} P(W,T|C) = \arg\max_{W,T} P(W,T) \quad (1)$$

一般に英語では $P(W,T)$ を品詞付けモデル (tagging model) と呼ぶが、日本語では単語分割の方が品詞付けより重要な問題なので、以下では $P(W,T)$ を単語分割モデル (word segmentation model) と呼ぶ。

我々は頑健な形態素解析プログラムを構成するために、三つの単語分割モデル (品詞 trigram, 単語 bigram, 単語 unigram) を比較した。品詞 trigram モデルでは、品詞 trigram 確率 $P(t_i|t_{i-2},t_{i-1})$ と品詞別単語出現確率 $P(w_i|t_i)$ の積で $P(W,T)$ を近似する。

$$P(W,T) = \prod_{i=1}^{n} P(t_i|t_{i-2},t_{i-1})P(w_i|t_i) \quad (2)$$

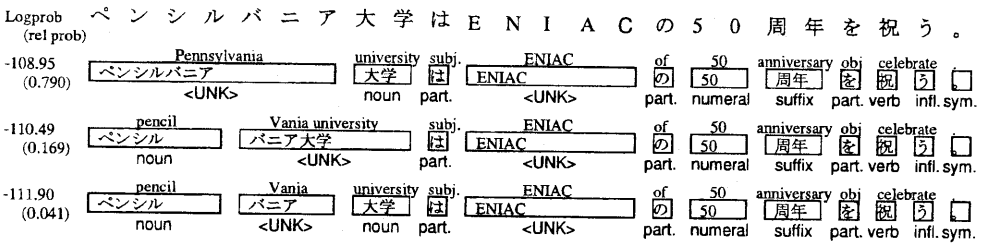

図 1: 未知語を含む日本語形態素解析の例

単語 unigram モデルおよび単語 bigram モデルでは、単語出現確率 $P(w_i, t_i)$ および単語 bigram 確率 $P(w_i, t_i|w_{i-1}, t_{i-1})$ で $P(w_i, t_i)$ を近似する[1]。

$$P(W,T) = \prod_{i=1}^{n} P(w_i, t_i) \tag{3}$$

$$P(W,T) = \prod_{i=1}^{n} P(w_i, t_i|w_{i-1}, t_{i-1}) \tag{4}$$

3.2 単語モデル

任意の単語候補に対して適当な確率を割り当てる統計的単語モデルを導入する。正確には、すべての未知語は <UNK> という特別な品詞に属すると考え、ある単語 w_i が未知語であるとき、その表記が長さ k の文字列 $c_1 \ldots c_k$ である確率 $P(c_1 \ldots c_k|\text{<UNK>})$ を単語モデルと定義する。この確率は以下のような単語長確率と単語表記確率の積に分割できる。

$$\begin{aligned} P(w_i|\text{<UNK>}) &= P(c_1 \ldots c_k|\text{<UNK>}) \\ &= P(k|\text{<UNK>})P(c_1 \ldots c_k|k, \text{<UNK>}) \end{aligned} \tag{5}$$

以下では $P(k|\text{<UNK>})$ を単語長モデル、$P(c_1 \ldots c_k|k)$ を単語表記モデルと呼ぶ。

単語長確率は平均単語長 λ をパラメタとするポワソン分布に従うと仮定する。すなわち、長さ 0 の単語区切り記号を考え、この区切り記号の平均間隔が平均単語長に等しくなるようにランダムに配置されたモデルで単語分割を近似する。

$$P(k|\text{<UNK>}) = \frac{(\lambda-1)^{k-1}}{(k-1)!} e^{-(\lambda-1)} \tag{6}$$

[1] 式 (3) と式 (4) では、一つの単語は表記 w_i と品詞 t_i の組で定義されると考える。

単語表記確率は単語内文字 bigram 確率の積で近似する。なお語頭と語末は特別な記号 "#" で表す。単語内文字 bigram は、単語の先頭 (接頭辞)・中間・末尾 (接尾辞) に現れる文字 bigram に対して大きな確率を割り当て、単語境界をはさむ文字 bigram には小さな確率を割り当てる性質を持つ。

$$P(c_1 \ldots c_k|k, \text{<UNK>}) = P(c_1|\#) \prod_{i=2}^{k} P(c_i|c_{i-2})P(\#|c_k) \tag{7}$$

単語モデル $P(w_i|\text{<UNK>})$ を使って、未知語に対する単語分割モデルを定義する。まず品詞 trigram モデルでは、未知語 w_i に対する品詞別単語出現確率 $P(w_i|t_i)$ は、定義により単語モデルに等しい。

$$P(t_i|t_{i-2}, t_{i-1}) = P(\text{<UNK>}|t_{i-2}, t_{i-1}) \tag{8}$$

$$P(w_i|t_i) = P(w_i|\text{<UNK>}) \tag{9}$$

単語 unigram モデルでは、未知語 w_i に対する単語出現確率 $P(w_i, t_i)$ は未知語出現確率と単語モデルの積である。また単語 bigram モデルでは、未知語 w_i に対する単語 bigram 確率 $P(w_i, t_i|w_{i-1}, t_{i-1})$ は未知語を含む単語 bigram 確率と単語モデルの積である。

$$P(w_i, t_i) = P(\text{<UNK>})P(w_i|\text{<UNK>}) \tag{10}$$

$$P(w_i, t_i|w_{i-1}, t_{i-1}) = P(\text{<UNK>}|w_{i-1}, t_{i-1})P(w_i|\text{<UNK>}) \tag{11}$$

未知語の出現確率 $P(\text{<UNK>}|t_{i-2}, t_{i-1})$, $P(\text{<UNK>})$, $P(\text{<UNK>}|w_{i-1}, t_{i-1})$ は、学習コーパスにおいて 1 回しか出現しなかった単語をすべて <UNK> に置き換えたコーパスから求める。この方法は、低頻度単語 (系列) に割り当てられた確率を未知語 (未知系列) に再配分するという点で、バックオフ法 [Katz, 1987] におけるチューリング推定と基本的に同じである。

4 一般化前向き後向き再推定

4.1 一般化前向き / ビタビアルゴリズム

英語の品詞付けでは、式(1)を最大化する品詞系列はビタビアルゴリズムで求め[Church, 1988]、式(2)のパラメタの最尤推定量は前向き後向きアルゴリズムで求める[Cutting et al., 1992]。しかし、日本語では単語候補が互いに重なり合うので、これらのアルゴリズムは適用できない。

そこで日本語の単語分割のために一般化前向きアルゴリズムと一般化ビタビアルゴリズムを次のように定義する。長さ n の日本語文字列を $C = c_1 c_2 \ldots c_n$ とし、部分文字列 $c_{p+1} \ldots c_q$ を c_p^q で表す。辞書に相当するものとして、文字列 c_p^q から単語候補のリスト $\{w_i\}$ への写像 D を考える。以下の説明では w_i が表記と品詞の組を表すことにし、単語分割モデルとして単語 bigram を用いる場合だけを扱う。

一般化前向きアルゴリズムでは、前向き確率 $\alpha_p^q(w_i)$ を、文字列 c_0^q が出現し、かつ、その最後の単語 w_i の表記が文字列 c_p^q である確率と定義する。前向き確率は式(12)により再帰的に計算できる。

$$\alpha_q^r(w_{i+1}) = \sum_{0 \le p < q} \sum_{w_i \in D(c_p^q)} \alpha_p^q(w_i) P(w_{i+1}|w_i)$$
$$w_{i+1} \in D(c_q^r), 0 \le q < n, q < r \le n \quad (12)$$

一般化ビタビアルゴリズムは式(12)の総和を最大化に置き換えたものである。$\phi_p^q(w_i)$ は、最後の単語 w_i の表記 c_p^q であるような、文字列 c_0^q に対する最尤単語列(単語分割候補)の確率と定義する。

$$\phi_q^r(w_{i+1}) = \max_{0 \le p < q} \max_{w_i \in D(c_p^q)} \phi_p^q(w_i) P(w_{i+1}|w_i)$$
$$w_{i+1} \in D(c_q^r), 0 \le q < n, q < r \le n \quad (13)$$

なお元の前向きアルゴリズムとビタビアルゴリズムは、式(12)および式(13)において、p と r を $p = q-1$ および $r = q+1$ に限定した場合に相当する。

未知語を扱うために、写像 D は、辞書に登録されていない文字列 c_p^q に対して、品詞が<UNK>の単語候補を返す。遷移確率 $P(w_{i+1}|w_i)$ は単語モデルに基づいて割り当てる。従って、一般化前向きアルゴリズムおよび一般化ビタビアルゴリズムでは、入力文中のすべての部分文字列を単語と仮定し、それらのすべての組合せを調べる。さらに、一般化後向きアルゴリズムも同様に定義できるので、一般化前向き後向きアルゴリズムは容易に導出できる。

4.2 単語の頻度の期待値

一般化前向きアルゴリズムまたは一般化ビタビアルゴリズムを用いれば、入力文に対するすべての単語分割候補を求めることができる。コーパスの第 i 番目の文の第 j 番目の単語分割候補を O_j^i とする。$P(O_j^i)$ は単語分割モデルから求められるので、第 i 文における単語 w_α の頻度 $C^i(w_\alpha)$ は以下のようになる。

$$C^i(w_\alpha) = \sum_j \left(\frac{P(O_j^i)}{\sum_k P(O_k^i)} \times n_j^i(w_\alpha) \right) \quad (14)$$

ここで $n_j^i(w_\alpha)$ は単語 w_α が第 i 文の j 候補に出現した回数を表す。コーパス中の単語出現頻度の期待値 $C(w_\alpha)$ は、すべての文に関する総和から求める。

$$C(w_\alpha) = \sum_i C^i(w_\alpha) \quad (15)$$

4.3 テキスト中の未知語の収集

テキスト中の単語頻度の期待値(式(15))は、テキスト中の任意の部分文字列の「単語らしさ」の尺度と考えることができる。そこで単語出現頻度の期待値の閾値 θ とし、単語候補 w_α の出現頻度の期待値が θ より大きく、かつ、w_α が辞書に登録されていなければ w_α を未知語として収集する。

$$C(w_\alpha) > \theta \quad (16)$$

ここでは単語の出現頻度の期待値を式(14)から求める方法を示したが、一般化前向き後向きアルゴリズムを用いれば、すべての単語分割候補を列挙せずに単語の頻度の期待値を求められる。しかし、一般化前向き後向きアルゴリズムは、計算量とスケーリングに少し難点があるので、本手法では、式(14)を N-best 単語分割候補の重み付き総和で近似した。N-best 単語分割候補は、前向き DP 後向き A^* アルゴリズム[Nagata, 1994]により求めた[2]。

[2] HMM のパラメタ再推定において、最尤候補のみを用いる方法をビタビ再推定と呼ぶ。これに対して本方法は N-best 再推定とでも呼ぶべきもので、ビタビ再推定よりは精度が高く、一般化前向き後向き再推定よりは計算量が少ない。

図 2: 単語頻度の期待値計算の例

4.4 単語の出現頻度の期待値の計算例

ここでは単語頻度の期待値計算の例を示す。コーパスの第 i 文が文字列 "言語学入門" であり、その上位 3 個の単語分割候補を図 2とする。図 2の左端は単語分割候補の相対確率であり、式 (14) の $\frac{P(O_j^i)}{\sum_k P(O_k^i)}$ に相当する。この文における各単語候補の出現頻度の期待値は以下のようになる。

$$\begin{aligned} C^i(\text{入門}) &= 0.7 + 0.2 + 0.1 = 1.0 \\ C^i(\text{言語学}) &= 0.7 \\ C^i(\text{言語}) = C^i(\text{学}) &= 0.2 \\ C^i(\text{言}) = C^i(\text{語学}) &= 0.1 \end{aligned}$$

この文の単語数の期待値 $\sum_\alpha C^i(w_\alpha)$ は 2.3 である。もし、どの単語候補も辞書に登録されておらず、閾値 θ が 0.15 ならば、"入門","言語学", "言語", "学" を単語と認定し、" 言" と "語学" は無視する。

次に単語頻度の期待値のコーパスにおける総和を求める例を示す。まず "ペンシルバニア大学は ENIAC の 50 周年を祝う。" という文がコーパスに含まれ、その上位 3 個の単語分割候補を図 1とすれば、"ペンシルバニア", "バニア大学", "バニア" の単語頻度の期待値はそれぞれ 0.790, 0.169, 0.041 となる。さらに "ホワイトハウスはペンシルバニア通りにある。" という文もコーパスに含まれ、"ペンシルバニア", "バニア通り", "バニア" の単語頻度の期待値がそれぞれ 0.825, 0.127, 0.048 とする。コーパス全体における単語頻度は以下のようになる。

$$\begin{aligned} C(\text{ペンシルバニア}) &= 0.790 + 0.825 = 1.615 \\ C(\text{バニア大学}) &= 0.169 \\ C(\text{バニア通り}) &= 0.127 \\ C(\text{バニア}) &= 0.041 + 0.048 = 0.089 \end{aligned}$$

"ペンシルバニア" は頻度の期待値が増加し、未知語と同定される可能性が高くなる。一般に、ある単語候補がコーパス中により多く出現するほど、たとえ文の単語分割に曖昧性があったとしても、その単語候補は未知語として収集される可能性が高くなる。

5 実験

5.1 言語データ

我々は単語分割プログラムの訓練と試験のために「EDR 日本語コーパス Version 1.0」[EDR, 1995] を用いた。EDR コーパスは、新聞・雑誌・辞書・百科辞典・教科書などから収集され、形態論・統語論・意味論レベルの様々な注釈が付与された約 500 万単語 (20 万文) のコーパスである。この実験では単語区切り・読み・品詞の情報を用いた。

まずコーパス全体の約 90% に相当する文を無作為に抽出して訓練集合とし、残りの 10% の中からテスト集合 1(100 文) とテスト集合 2 (1000 文) を無作為に抽出した。表 1に訓練集合とテスト集合の文・単語・文字の数を示す。

表 1: 訓練データと試験データの量

	訓練集合	テスト集合 1	テスト集合 2
文	192802	100	1000
単語	4746461	2463	25177
文字	7521293	3912	39875

訓練テキスト中の異なり単語数は 133281 個であり、頻度 2 以上の 65152 単語を辞書に登録した。単語分割モデルは、頻度 1 の単語を<UNK>で置き換えた後に、品詞 trigram, 単語 unigram, 単語 bigram の 3 種類を作成した。単語 bigram は 758172 個のうち頻度 2 以上の 294668 個を使用した。また、訓練テキスト中の異なり文字数は 3534 個であり、頻度 2 以上の 3167 個を既知記号とした。頻度 1 の文字を未知記号タグに置き換えて作成した単語内文字 bigram のうち頻度 2 以上の 91198 個を単語表記モデルとして使用した。すべての N-gram 確率は削除補完法により平滑化した [Jelinek, 1985]。また、単語長モデルのパラメタは、すべての単語の平均単語長 1.58 と低頻度単語の平均単語長 4.49 の 2 種類を試した。

```
JC00092627
ロックフェラー研究所はアメリカの大富豪ロックフェラーが設立
した学術研究所です。

コーパス                          システム出力
ロックフェラー研究所 / ロック     | ロックフェラー / ロックフェ
                                  > 研究所 / ケンキュウジョ / 名詞
は / ハ / 助詞                    は / ハ / 助詞
アメリカ / アメリカ / 名詞        アメリカ / アメリカ / 名詞
の / ノ / 助詞                    の / ノ / 助詞
大 / ダイ / 接頭語                大 / ダイ / 接頭語
富豪 / フゴウ / 名詞              富豪 / フゴウ / 名詞
ロックフェラー / ロックフェラ     ロックフェラー / ロックフェ
が / ガ / 助詞                    が / ガ / 助詞
設立 / セツリツ / 動詞            設立 / セツリツ / 動詞
し / シ / 語尾                    し / シ / 語尾
た / タ / 助動詞                  た / タ / 助動詞
学術研究所 / ガクジュツケンキ     | 学術研究所 /NIL/<UNK>
です / デス / 助動詞              です / デス / 助動詞
。 /。 / 記号                     。 /。 / 記号

sys=15, std=14, matched=13
precision=87.7 (13/15), recall=92.9 (13/14)
```

図 3: コーパスとシステム単語分割の比較

5.2 評価尺度

単語分割精度は再現率と適合率で評価する。まず正解コーパス中の単語数 (Std)、システム出力の単語数 (Sys)、および、両者の照合数 (M) を求め、再現率 (M/Std)、適合率 (M/Sys) を計算する。

“ロックフェラー研究所はアメリカの大富豪ロックフェラーが設立した学術研究所です。” という文の単語分割に対する再現率と適合率の計算例を図 3に示す。辞書未登録語である’学術研究所’が正しく単語分割されていることに注目して欲しい。

未知語抽出精度は再現率・適合率・ F- 尺度で評価する。まず正解コーパス中の未知語数 (Std)、システムが未知語と同定した単語数 (Sys)、および両者の照合数 (M) を求め、再現率 (M/Std) と適合率 (M/sys) を計算する。未知語抽出の再現率と適合率は単語頻度の閾値に大きく依存するので、総合的な評価のために F- 尺度を用いる。 F- 尺度は情報検索で用いられるもので次式で定義される。

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R} \qquad (17)$$

ここで P と R は再現率と適合率を表し、β は適合率に対する再現率の相対的重要度である。本実験では $\beta = 1.0$ とし、再現率と適合率の重みを等しくした。

5.3 単語分割精度

予備実験として、最も頑健な単語分割法を調べるために、品詞 trigram, 単語 unigram, 単語 bigram の 3 つの単語分割モデルを比較し、さらに単語長モデル (平均単語長) に関しても全単語と低頻度単語の 2 種類を比較した。表 2に、テスト集合 1(100 文) に対する単語分割精度を示す。

単語分割モデルとしては、品詞 trigram や単語 unigram に比べて単語 bigram の方が明らかに精度が高い。単語長モデルについては、低頻度語の平均単語長を用いる方が精度は高いが、その差は僅かである。結局、以下の実験では、単語 bigram(単語分割モデル)、文字 bigram(単語表記モデル)、低頻度語の平均単語長 (単語長モデル) を用いることにした[3]。

5.4 未知語抽出精度

本報告で提案した未知語抽出法をテスト集合 2(1000 文) に適用した結果を表 3に示す。テスト集合 2 には 538 個の未知語があり、未知語率は 2.1% である。その中で頻度 2 の未知語が 8 個、残りの 530 個はすべて頻度 1 である。単語頻度の期待値は上位 10 個の単語分割候補から計算し、未知語の長さは最大 8 文字までに制限した。

表 3に示すように、単語頻度の期待値の閾値 θ が高いほど、未知語抽出の適合率が高くなり、再現率は低くなる。再現率と適合率の重みを同等としたときの最適の閾値は 0.10 付近であり、その場合、再現率 43.7%, 適合率 52.3% である。

図 4に、単語頻度の閾値が 0.5 の場合に、抽出に成功した未知語 (matched)、誤って抽出した単語候補 (sys-matched)、抽出に失敗した未知語 (std-matched) の例を示す。再現率や適合率はあまり高くないが、

[3] 表 2において、単語 unigram の単語分割精度が約 90% であることは注目に値する。単語分割では単語の出現頻度の情報が最も重要であり、文脈情報 (品詞の連接制約など) は 2 次的な要素であることが分かる。

表 2: 言語モデルと単語分割精度 (100 個のテスト文)

単語長モデル	品詞 trigram		単語 unigram		単語 bigram	
	再現率	適合率	再現率	適合率	再現率	適合率
全単語	91.6	88.8	88.7	87.3	94.6	89.4
低頻度単語	91.5	89.3	88.8	87.6	94.7	89.9

```
matched=196 (抽出に成功した未知語)
３万１４８７ しんかい２０００ キリシタン ジャップ トム・ニース トリップ フリードリッヒ レトリック
暗語 印紙 開明 楽天主義 凶悪犯 作業帽 賜杯 褒まし 日伯援護協会 百松 傍聴者 与信 ...
sys-matched=103 (誤って抽出した単語候補)
９０万７０００余 ＳＴＫ製 エクソン社 エストリッジ ジャカール機械 ファクチュア化 フローティング マニュ
乾式構法 汗牛充 順々 清掃局 占い師 村山大臣 東大宇宙航空 灘中、灘高 二浪 年功給 破壊読出し 陸上幕僚長 ...
std-matched=342 (抽出に失敗した未知語)
４０４ ＢＢＮアドバンスト・コンピューター社 X線天文学 あっと言う間 ギャラップ調査 レジャー産業
ロックフェラー研究所 引きも切らず 教員住宅 勤労意欲 国際情報化社会 仕立て物 実験班 真珠採取 吹き付け
清掃車 先客 先鞭をつけ 短音 倒れ込 ...
threshold=0.5, std=538, sys=299, matched=196
recall=36.4 (196/538), precision=65.6 (196/299)
```

図 4: 未知語収集の例

表 3: 未知語抽出精度 (1000 個のテスト文)

閾値 θ	再現率	適合率	F- 尺度
>0.00	56.1	34.2	42.5
>0.10	43.7	52.3	47.6
>0.50	36.4	65.6	46.8
>0.90	25.3	76.8	35.8
>0.95	23.2	78.1	35.8
>0.99	17.3	81.6	28.5

我々は本手法の未知語収集能力は満足できるレベルにあると考えている。その理由は次節で議論する。

6 考察

日本語単語分割の性能評価の問題点は、多くの人が合意できる唯一の正解が存在しないことである。一般に、本報告のように唯一の正解 (EDR コーパスの単語分割) との完全一致に基づく評価は、(ある被験者の) 許容可能性に基づく評価よりも過小評価になる傾向がある。(少なくとも筆者にとっては) システムが収集した未知語の多くは許容可能であるし、システムが収集に失敗した未知語の多くは既に辞書に登録されている単語の組合せに分割できる。

例えば、“データ・コミュニケーション” は EDR コーパスでは 1 つの単語として扱われているが、システム出力では “データ” と “コミュニケーション” (両方とも辞書に登録されている) に分割されたため、未知語抽出誤りとなる。また、システム出力では “ハノーヴァ公” を未知語として抽出しているが、EDR コーパスでは “ハノーヴァ” と “公” に分割されているため誤りとなる。ちなみにシステム辞書には、接尾辞’公’は登録されているが’ハノーヴァ’は登録されていない。しかし、EDR コーパスの別の箇所では “ハノーヴァー国立図書館” が 1 語になっている。大部分の未知語抽出誤りは、このパターンである。

もちろん明らかな誤りもあり、それらは大きく 3 つに分類できる。第 1 のタイプは「長い単語の切断」である。例えば、“イラストレーション” という単語の長さは 9 文字なので、予め設定した未知語の最大

長 (8 文字) を越えていて、かつ、“イラスト” という単語が辞書に登録されているので、“レーション” が未知語として抽出されてしまう。

第 2 のタイプは「数詞の断片化」である。これは本システムが字句解析 (tokenization) を全く行なっていないことに原因がある。例えば、“1676” という数詞は、“16” と “76”, “1” と “676”, “16” と “7” と “6” というように任意の部分列に分割されてしまう。

第 3 のタイプは「名詞と助詞の連結」である。本システムは、“A の B”, “A と B”, “A,B” などの名詞句を一つの名詞とみなしてしまう場合がある。例えば、システムは “可制御かつ可観測” を未知語として抽出したが、正解は “可 / 制御 / かつ / 可 / 観測” という単語分割である。これは未知語や低頻度語を含む複数の短い単語の確率の積よりも一つの長い未知語の確率の方が大きくなるためだと思われる。

7 関連研究

近年、日本語や中国語を対象とした単語分割および辞書作成の研究が盛んである。[Chang et al., 1995] は、単語分割されていない大きなコーパス (約 31 万文) と単語分割された小さなコーパス (1000 文) を用いて中国語の辞書を自動的に作成する方法を提案した。彼らの手法は、単語 unigram を用いたビタビ再推定、および、文字 N-gram を特徴量として文字列が単語かどうかを判定する 2 クラス分類器 (Two-Class Classifier) を用いる。システム出力を二つのオンライン中国語辞書 (約 2 万語) と比較し、2 文字単語では再現率 56.88%, 適合率 77.37%、3 文字単語では再現率 6.12%, 適合率 85.97% と報告している。

日本語では、[Nagao and Mori, 1994] では、suffix array 法を用いて任意の長さの文字 N-gram を求め、文字 N-gram の情報が単語抽出に重要であることを指摘しているが、具体的な評価はない。

本手法は、単語分割に単語 bigram を使用し、未知語収集には単語頻度の期待値を用いる。単語分割されたコーパスを正解とみなし、未知語率 2.1% の 1000 文に対して再現率 43.7%, 適合率 52.3% を得ている。[Chang et al., 1995] と本手法を比較することは、言語 (中国と日本語)、訓練に用いた (単語分割された / されていない) コーパスの大きさ、初期単語リストの大きさ、試験に用いたコーパスの大きさと未知語率、正解データの種類 (オンライン辞書と単語分割されたコーパス) などが異なるので不可能である。しかし、確率モデルに基づく単語頻度の期待値を用いる本手法の方がより簡単かつ有用であると主張したい。

8 おわりに

本報告では、日本語テキストから未知語を収集する方法を述べた。今後は、未知語処理技術を応用して、プレーンテキストから単語 bigram を学習する方法を確立したい。

参考文献

[Chang et al., 1995] Jing-Shin Chang, Yi-Chung Lin, and Keh-Yih Su. 1995. Automatic Construction of a Chinese Electronic Dictionary, VLC-95, pp.107-120.

[Church, 1988] Kenneth W. Church. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, ANLP-88, pp.136-143.

[Cutting et al., 1992] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A Practical Part-of-Speech Tagger, ANLP-92, pp.133-140.

[EDR, 1995] Japan Electronic Dictionary Research Institute. 1995. *EDR Electronic Dictionary Version 1 Technical Guide*, EDR TR2-003. Also available as *The Structure of the EDR Electronic Dictionary*, `http:///www.iijnet.or.jp/edr/`.

[Jelinek, 1985] Frederick Jelinek. 1985. Self-organized Language Modeling for Speech Recognition. IBM Report.

[Katz, 1987] Slava M. Katz. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Trans. ASSP-35, No.3, pp.400-401.

[Nagao and Mori, 1994] Makoto Nagao and Shinsuke Mori. 1994. A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, COLING-94, pp.611-615.

[Nagata, 1994] Masaaki Nagata. 1994. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm. COLING-94, pp.201-207.

[Nagata, 1996a] Masaaki Nagata. 1996. Automatic Extraction of New Words from Japanese Texts using Generalized Forward-Backward Search. EMNLP, pp.48-59.

[Nagata, 1996b] Masaaki Nagata. 1996. Context-Based Spelling Correction for Japanese OCR. COLING-96, pp.806-811.