

単語頻度の再推定による自己組織化単語分割

永田昌明

NTT 情報通信研究所

nagata@nttnly.isl.ntt.co.jp

本稿では、小さな単語リストと大量のプレーンテキストから日本語の単語分割プログラムを作成する方法を提案する。本手法は、単語単位の統計的言語モデル、初期値推定手続き、再推定手続きから構成される。まず文字種に関するヒューリスティクスを用いて訓練テキストから抽出した単語候補を単語リストに加え、単語リスト中の単語と最長一致する訓練テキスト中の文字列の頻度から単語頻度の初期値を推定する。次に単語頻度に基づく統計的言語モデルを用いて訓練テキストを単語分割し、単語リストと単語頻度を再推定する。1719 個の単語と 390 万文字のテキストに対して本手法を用いて単語分割プログラムを訓練したところ、単語分割精度は再現率 86.3%、適合率 82.5% であった。

A Self-organizing Japanese Word Segmenter using Heuristic Word Identification and Re-estimation

Masaaki NAGATA

NTT Information and Communication Systems Laboratories

We present a self-organized method to build a stochastic Japanese word segmenter from a small number of basic words and a large amount of unsegmented training text. It consists of a word-based statistical language model, an initial estimation procedure, and a re-estimation procedure. Initial word frequencies are estimated by counting all possible longest match strings between the training text and the word list. The initial word list is augmented by identifying words in the training text using a heuristic rule based on character type. The word-based language model is then re-estimated to filter out inappropriate word hypotheses. When the word segmenter is trained on 3.9M character texts and 1719 initial words, its word segmentation accuracy is 86.3% recall and 82.5% precision.

1 はじめに

日本語は単語を分かち書きする習慣がない。しかし、一般に個々の文字の発音や意味から単語の発音や意味を構成的にできないので、日本語を扱う自然言語処理アプリケーションでは、単語分割が必要になる。中国語やタイ語も同様の問題を持っている。

ある意味では日本語の単語分割問題は解決済みである。人手により単語分割された訓練テキストが大量にあれば、統計的言語モデルと動的計画法を用いて95%以上の単語分割精度を得られる[Nagata, 1994, Yamamoto, 1996, 竹内・松本, 1997]。しかし人手による単語分割は非常に高価であり、対象領域ごとに単語分割済みコーパスを大量に用意するのは難しい。

我々の研究目標は、日本語単語分割の教師なし学習、すなわち、単語リストとプレーンテキストから自己組織的な方法で単語分割プログラムを作成することである。現在では、1万語～10万語程度の単語リストおよび10MB～100MB程度のテキストを用意することは難しくない。教師なし学習が実現できれば、単語分割プログラムの開発コストが大幅に低下し、自然言語処理の適用範囲が大幅に広がる¹。

日本語の教師なし単語分割に関する研究は少ない。[Yamamoto, 1996] および [竹内・松本, 1997] では、規則に基づく単語分割プログラムを用いて訓練テキストを単語分割することにより統計的言語モデルの初期値を求め、再推定手続きによりモデルを洗練する。このアプローチは、JUMAN[松本ほか, 1994]のような規則に基づく単語分割プログラムが存在することを前提としているので、人手の介在なしに新たな対象領域に適用することは難しい。また語彙収集基準や文法体系に関する自由度も小さい。

中国語の単語分割ではもう少し自己組織的な方法が研究されている。[Sproat et al., 1996] は、辞書中の単語のコーパスにおける文字列頻度を初期値としてビタビ再推定により単語 unigram モデルを構築した。[Chang et al., 1995] は、小さな単語分割済みコーパス (seed corpus) と大きな (単語分割されていない)

コーパスから Viterbi 再推定により単語 unigram モデルを求めた。[Luo and Roukos, 1996] は、訓練テキストを2つに分け、一方で単語分割を行い、他方で単語頻度を再推定するというステップを交互に繰り返す訓練法を提案した。

教師なし単語分割の大きな課題の一つは未知語の扱いである。[Sproat et al., 1996] では、名詞の複数形、中国人の名前、外来語の発音表記のような規則性のある語形成過程について語彙規則を用意した。[Chang et al., 1995] では、ある文字列が語を構成するかを文字 n-gram 統計に基づいて判定する2クラス分類器 (two-class classifier) を用いた。

本稿で提案する手法の特徴は、(1) 再推定可能な統計的単語モデル、(2) 単語頻度の初期値の推定法、(3) ヒューリスティックな単語同定法による初期単語リストの拡張と再推定により不適切な単語候補の排除の組み合わせ、の3つである。図1に本稿で提案する日本語単語分割プログラムのブロック図を示す。以下では、まず統計的単語モデルと単語分割アルゴリズムについて述べ、次に単語頻度の初期値の推定法と初期単語同定法について述べる。最後に様々な条件下で教師なし単語分割の実験結果を報告する。

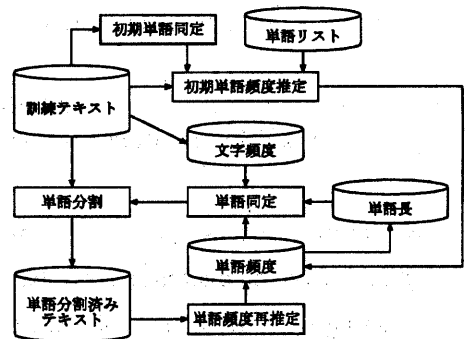


図1: 教師なし日本語単語分割のブロック図

2 言語モデルと単語分割アルゴリズム

2.1 単語分割モデル

文字列 $C = c_1 c_2 \dots c_m$ が単語列 $W = w_1 w_2 \dots w_n$ に分割されるとする。単語分割は、与えられた文字

¹一般にテキスト中の単語出現率の情報があれば単語分割プログラムを作成できる。逆に単語分割プログラムがあればテキスト中の単語出現率を求められる。この「卵と鶏の関係」から抜け出すことが教師なし単語分割の中心的課題である。

列に対する単語列の同時確率 $P(W|C)$ を最大化する単語列 $\hat{W}$ を求める問題である。文字列 C は共通なので $P(W)$ を最大化すればよい。

$$\hat{W} = \arg \max_W P(W|C) = \arg \max_W P(W) \quad (1)$$

ここでは計算量を考慮して同時確率 $P(W)$ を単語 unigram モデル (単語出現確率 $P(w_i)$ の積) で近似する。

$$P(W) = \prod_{i=1}^n P(w_i) \quad (2)$$

2.2 単語モデル

入力文中の任意の部分文字列に対して適当な単語確率を割り当てるために以下のような統計的単語モデルを定義する。厳密には、単語モデルはある単語 w_i が未知語であるときに、その表記が文字列 $c_1 \dots c_k$ である確率である。これを以下のような単語長確率 $P(k)$ と単語表記確率 $P(c_1 \dots c_k)$ の積に分解する。

$$P(w_i | \langle \text{UNK} \rangle) = P(c_1 \dots c_k | \langle \text{UNK} \rangle) = P(k)P(c_1 \dots c_k) \quad (3)$$

ここで k は文字列の長さ、 $\langle \text{UNK} \rangle$ は未知語を表す。

単語長確率 $P(k)$ は訓練テキストの平均単語長 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ると考える。また単語表記確率 $P(c_1 \dots c_k)$ は文字 unigram 確率の積で近似する。

$$P(k) = \frac{(\lambda - 1)^{k-1}}{(k-1)!} e^{-(\lambda-1)} \quad (4)$$

$$P(c_1 \dots c_k) = \prod_{i=1}^k P(c_i) \quad (5)$$

文字 unigram 確率は単語分割されていないテキストから推定できる。平均単語長 λ は単語頻度から次式により求められる。

$$\lambda = \frac{\sum |w_i| C(w_i)}{\sum C(w_i)} \quad (6)$$

ここで $|w_i|$ and $C(w_i)$ は単語 w_i の長さと頻度を表す。従ってこの言語モデルにおいて再推定が必要なパラメタは単語頻度だけである。

図 2 に EDR コーパスの実際の単語長分布およびポワソン分布からの推定値を示す。ここでは、すべての単語 ($\lambda = 1.6$) と出現頻度 1 の単語 ($\lambda = 4.8$) の 2 つの分布を示す。後者は未知語の分布に近い。

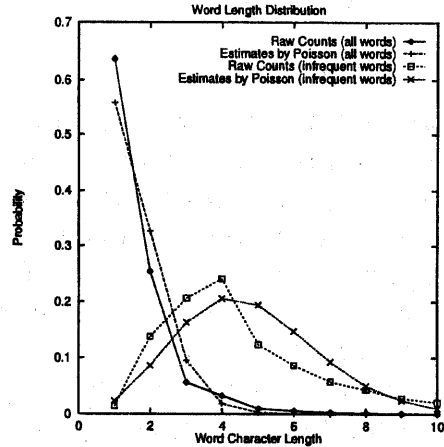


図 2: EDR コーパスにおける単語長の分布

2.3 ビタビ再推定

入力文に対する最尤な単語分割はビタビアルゴリズムを拡張した動的計画法を用いて求める [Nagata, 1994]。

このアルゴリズムは文頭から一文字ずつ進み、各文字位置においてその位置で終了する単語列とその位置から開始する単語候補の全ての組み合わせを調べる。

計算量を考慮して単語 unigram の再推定にはビタビ再推定を用いた。まず単語頻度の初期推定値に基づいてビタビアルゴリズムを訓練コーパスに適用する。次に最尤単語列を正解と見なし単語頻度を再推定する。この手続きを収束するまで繰り返す。

3 単語頻度の初期値

3.1 最長一致

一般に単語辞書と何らかのヒューリスティックスを組み合わせて訓練コーパスを単語分割すれば、単語頻度の初期値を推定することができる。日本語の単語分割において (中国語でも) 最も一般的なヒューリスティックスは最長一致である [Wu and Tseng, 1993]。ここでは [Sproat et al., 1996] の食欲なアルゴリズム (左最長一致) を用いた。文頭から始め、ある位置で辞書と一致する最長の単語を探し、その終わりの次の文字を開始点として同様の手続きを文末まで繰り返す。このアルゴリズムは実装が簡単で単語分割が

一意に決定できるという利点がある。

3.2 文字列頻度

[Sproat et al., 1996] では、単語を構成する文字列の訓練コーパスにおける頻度 (文字列が本当にその単語を表すかに関係なく) から単語頻度の初期値を推定する。コーパス中の単語延べ総数は辞書中の各単語の文字列頻度の和とする。テキスト T 中の文字列 W の出現頻度は接尾辞配列 (suffix array) と呼ばれるデータ構造を作成することにより効率的に求められる。接尾辞配列はテキスト T の全ての接尾辞をソートしたリストである [Manber and Myers, 1993]。

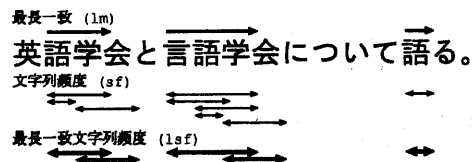
3.3 最長一致文字列頻度

文字列頻度から求めた単語頻度の推定値は、短い単語の頻度が実際よりかなり大きくなる傾向がある。この問題を解決するために我々は「最長一致文字列頻度」を考案した。この方法では、テキスト T における文字列 (単語) W_1 の出現箇所において、この出現箇所が辞書 D 中の他の文字列 (単語) W_2 の出現箇所の部分文字列でない場合の数を求める。

最長一致文字列頻度を求めるために、テキスト T と辞書 D に対して接尾辞配列 S_T と S_D を作成する。まず S_T を用いて文字列 W のテキスト T における全ての出現位置のリスト L_W を求める。次に S_D を用いて W を部分文字列として含む辞書中の全ての単語 $\bar{W}$ を求め、 S_T を用いて $\bar{W}$ のテキストにおける出現位置のリスト $L_{\bar{W}}$ を求める。辞書 D に関するテキスト T における単語 W の最長一致文字列頻度は、差集合 $L_W - L_{\bar{W}}$ の要素数に等しい。

例として、入力文が「英語学会と言語学会について語る。」であり、辞書に「言語学、言語、語学、学会、語」があるとき、3つの頻度推定法の違いを図3に示す。最長一致文字列頻度 (lsf) 法はテキストにおける全ての最長一致の可能性を考慮するのに対し、最長一致 (lm) 法は一つの可能性しか考慮しない。また文字列頻度 (sf) 法において短い単語の頻度が大きくなり過ぎるという問題が最長一致文字列頻度法では解決されていることは明らかである。

最長一致文字列頻度の問題点は、もし単語 W_1 が他の単語 W_2 の部分文字列であり、訓練テキストにお



	最長一致	文字列頻度	最長一致文字列頻度
言語学	1	1	1
言語	0	1	0
語学	1	2	1
学会	0	2	2
語	1	3	1
合計	3	9	5

図 3: 単語頻度の初期値の推定法の比較

いて W_1 が常に W_2 の部分文字列として出現する場合 (図3の「言語」と「言語学」)、 W_1 の頻度の推定値が0になってしまうことである。しかし、大きな訓練テキストではこのような例は少ない。また、この場合、 W_1 を辞書に残しておく必要があるかには議論の余地がある。現在は出現頻度を平滑化することで対処している。

4 語彙獲得ツールにより単語リストの拡張

日本語では、文字種の変化点が単語境界である可能性が高い。日本語には句読点以外に少なくとも5つの字種 (漢字、ひらがな、カタカナ、アルファベット、アラビア数字) がある。このヒューリスティクスだけでも単語分割プログラムを作成できるが、その精度は低い²。しかし、初期単語リストを拡張するための語彙獲得ツールとして文字種による単語分割法が非常に有効である。

初期単語リスト作成手順は以下の通り。まず訓練コーパスを文字種の変化点により分割し、頻度付き単語候補リストを作成する。ただし、ひらがな文字列は助詞・助動詞などの機能語列である可能性が高いので単語候補リストから取り除く。こうして得られた単語候補リストを元の単語辞書とマージする。

²他の単語頻度推定法による単語分割精度が70-80%程度なのに対して、文字種変化による単語分割の精度60%以下である。

拡張された初期単語リストには単語ではない文字列が多く含まれているが、その大半は再推定の過程で取り除かれる³。

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に EDR コーパス Version 1.0 [EDR, 1995] を用いた。EDR コーパスの大きさは 510 万語 (20 万 8 千文) で、新聞・雑誌・辞書・百科事典・教科書などから収集されており、単語分割・読み・品詞を始めとする豊富な言語情報タグが人手により付与されている。

この実験では、10 万文ずつ無作為に選択して 2 つの訓練集合を作成した。一方の訓練集合 (training-0) は初期単語リストを作成するのに使用し、他方の訓練集合 (training-1) は単語分割プログラムを訓練するのに使用した。また残りの 8 千文から無作為に 100 文を選択し、単語分割精度の評価に用いた。表 1 に訓練集合と試験集合における文・単語・文字の数を示す⁴。

表 1: 訓練データと試験データの量

	training-0	training-1	test
文	100000	100000	100
単語 (延べ)	2460188	2465441	2538
単語 (異なり)	85966	85967	919
文字	3897718	3906260	3984

人手により単語分割された training-0 の頻度に基づき、頻度の閾値を 1, 2, 5, 10, 50, 100, 200 とする 7 種類の初期単語リスト (D1-D200) を作成した。それぞれの単語リストの語彙数と試験文に対する未知語率 (out-of-vocabulary rate) を表 2 の第 2 列と第 3 列に示す。例えば、D200 は training-0 に 200 回以上出現した単語から構成される。D200 は 826 語し

³ここでは最も簡便な初期単語リスト作成法として文字種変化を利用したが、正規化頻度など「単語らしき文字列」を同定する手法ならば何を用いても良い。

⁴Training-1 は training-0 と同じ情報源から得られたブレンデドテキストとして使用しており、元のコーパスに付与されていた単語分割の情報は利用していない。

かないが、試験文中の単語の 76.6% (未知語率 23.4%) をカバーしている。これも Zipf の法則の一例である。

5.2 評価尺度

一般に単語分割精度は再現率と適合率で表現する [Nagata, 1994, Sproat et al., 1996]。人手により単語分割されたコーパス中の単語数を Std, 単語分割プログラムが出力した単語数を Sys, 両者の一致数を M とするとき、再現率は M/Std , 適合率は M/Sys である。常に再現率と適合率の 2 つの値を扱うのは煩わしいので、総合的な性能指標として次式の F- 尺度を用いる。

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R} \quad (7)$$

ここで P と R は適合率と再現率を表し、 β は適合率に対する再現率の相対的重要度を表す。以下では $\beta = 1.0$ とし、再現率と適合率を同等の重みで扱う。

5.3 単語頻度推定法の比較

前節で述べた 3 つの単語頻度推定法、最長一致 (lm)、文字列頻度 (sf)、最長一致文字列頻度 (lsf) を比較した。表 2 の第 6, 7, 8 列に様々な初期単語リストの大きさ (D1-D200) に対する単語分割精度 (F- 尺度) を示す。比較のために、訓練集合 training-1 (training-0 ではない) の人手による単語分割から求めた単語頻度 (wf) を用いた際の単語分割精度を表 2 の第 5 列に示す。表 2 をグラフにしたものを図 4 に示す。

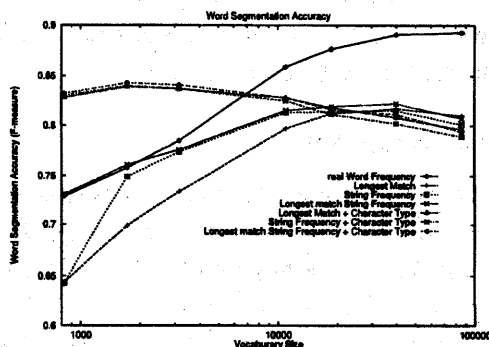


図 4: 初期単語リストの大きさと単語分割の精度

表 2: 単語分割精度

	頻度	語彙数	未知語率	wf	lm	sf	lsf	lm+ct	sf+ct	lsf+ct
D1	≥ 1	85966	0.010	0.893	0.810	0.801	0.807	0.796	0.789	0.794
D2	≥ 2	39994	0.017	0.891	0.817	0.815	0.822	0.808	0.802	0.811
D5	≥ 5	18689	0.037	0.877	0.812	0.814	0.819	0.818	0.811	0.816
D10	≥ 10	10941	0.060	0.859	0.797	0.813	0.815	0.828	0.825	0.828
D50	≥ 50	3159	0.134	0.785	0.734	0.774	0.776	0.837	0.837	0.841
D100	≥ 100	1719	0.181	0.758	0.699	0.749	0.761	0.839	0.840	0.843
D200	≥ 200	826	0.234	0.729	0.644	0.643	0.731	0.828	0.830	0.832

本当の単語頻度 (wf) を用いた単語分割の精度は、どの単語頻度推定法よりも 5-10% 程度優れている。3つの単語頻度推定法の中では、最長一致文字列頻度法 (lsf) が最も優れている。どの単語頻度推定法も D1 の単語分割精度が D2 よりも低い。これに対して本当の単語頻度を用いた場合は D1 の方が D2 より高い。

5.4 初期単語リスト拡張の効果

文字種に基づく語彙獲得法 (ct) と単語頻度推定法 (lm, sf, lsf) を組み合わせた場合の単語分割精度を表 2 の第 9,10,11 列に示す。ここでは訓練集合 training-1 から 108975 個の同一文字種列 (ひらがな列を除く) を単語候補として初期単語リストに追加した。

初期単語リストの拡張により単語分割精度は大きく改善される。単語頻度推定法による差は小さいが、3つの中では最長一致文字列頻度法が最も優れている。驚くべきことに、最も単語分割精度が高いのは、非常に小さな初期単語リスト (D100: 1719 語) と語彙獲得ツールを組み合わせた場合で、再現率 86.3%, 適合率 82.5% (F- 尺度 0.843) であった。

5.5 再推定の効果

再推定の効果を調べるために、3つの初期単語リスト (D1, D2, D100) および 2つの単語頻度推定法 (文字列頻度 (sf), 最長一致文字列頻度 + 同一文字種列 (lsf+ct)) の組み合わせをテストした。

ここではビタビ再推定を 3 回行った。それ以上繰り返しても変化はなさそうだった。再推定の各段階における、試験集合に対する単語分割頻度、訓練テ

キストにおける延べ単語数、辞書中の異なり単語数を図 5 に示す。

一般に単語分割精度は再推定による変化が少ない。初期単語リストが大きい (D1 と D2) ときは精度が少し良くなり、初期単語リストが小さい (D100) ときは精度が少し悪くなる。これは英語の品詞付けプログラムの教師なし学習の実験結果と一致するのかもしれない。[Kupiec, 1992] は非常に精緻な教師なし学習法を提案しているが、[Elworthy, 1994] は再推定は常に有効とは限らないと報告している。しかし、本実験の結果は単語 unigram を使用したことに起因しており、再推定が有効ではないと判断するのはまだ早いと報告者は考えている。

再推定の明らかな効能は、単語頻度の補正と不適切な単語候補の削除である。1 回目の再推定における延べ単語数の急激な減少は、実際より大きく推定されていた初期単語頻度がより適切な値に補正されたことを表す。また 1 回目の再推定における異なり単語数の減少は、語彙獲得ツールにより初期単語リストに付加された不適切な単語候補が削除されたことを表す。

6 議論

6.1 単語 unigram モデルの性質

まず最初に単語 unigram モデルの性質を明らかにしておく。もし 2つの単語分割候補の単語数が異なれば、単語数の少ない方が優先される。これは最長一致を優先されることとほぼ同じである。例えば、入力文字列が c_1c_2 で、辞書に c_1c_2 , c_1 , c_2 という 3

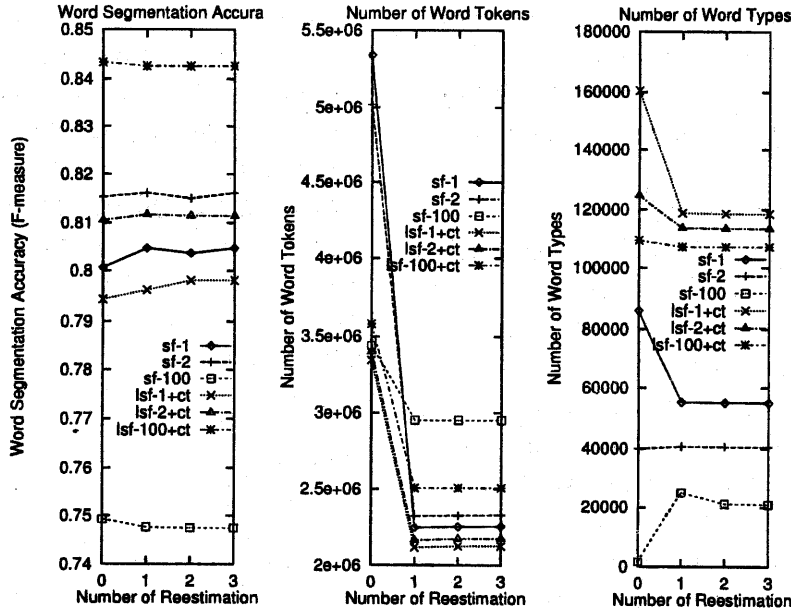


図 5: 単語分割精度、延べ単語数、異なり単語数

つの単語があるとする。 c_1c_2 より $c_1|c_2$ が優先されるためには、次の関係が成り立つ必要がある。

$$\frac{C(c_1c_2)}{N} < \frac{C(c_1)}{N} \frac{C(c_2)}{N} \quad (8)$$

ここで $C(\cdot)$ は単語頻度を表し、 N は延べ単語数を表す。仮に、 N を 100 万とすると、 $C(c_1c_2) = 1$ の場合、 c_1 と c_2 の頻度が非常に大きくない限り (例えば $C(c_1) \approx C(c_2) > 1000$)、 c_1c_2 が選ばれる。

もし 2 つの単語分割の単語数が同じならば、単語頻度の積が大きい方が優先される。例えば、入力文字列が $c_1c_2c_3$ で、辞書に c_1c_2 , c_3 , c_1 , c_2c_3 という 4 つの単語があるとする。 $c_1c_2|c_3$ より $c_1|c_2c_3$ が優先されるためには、次の関係が成り立つ必要がある。

$$\frac{C(c_1c_2)}{N} \frac{C(c_3)}{N} < \frac{C(c_1)}{N} \frac{C(c_2c_3)}{N} \quad (9)$$

分母の N は共通なので、単語頻度の積の大きい方が優先されるのは明らかである。

6.2 単語分割誤りの分類

単語分割誤りは大きく 3 つに分類される。第 1 のタイプは誤りではなく人手による単語分割の判断の

揺れ、あるいは日本語の単語分割が持つ本質的な曖昧性が原因である。例えば、人手により単語分割されたコーパスの中では、文字列「外国人労働者」が一つの単語とされている場合もあれば、「外国人」と「労働者」に分割されている場合もある。しかし、単語 unigram モデルを用いた単語分割では常に一つの単語と同定する。誤りの 3-5% はこのタイプである。

第 2 のタイプは未知語の分解である。例えば、「妙 (形容動詞の語幹)」が辞書にあるために、「珍妙」という文字列 (未知語) が「珍」と「妙」の 2 つの単語と同定されてしまう。未知語中の部分文字列が辞書中の単語と偶然一致した場合、辞書中の単語と残りの部分文字列に分解される場合が非常に多い。これは、長い単語に割り当てる確率が小さくなり過ぎるという文字 unigram モデルの欠点である。

第 3 のタイプは最長一致の誤りである。これは文末のひらがな表記された機能語列で頻発する。例えば、「集ま | つ | て | き | た」と分割されるべきなのに、より単語数が少ない「集ま | っ | て (提題詞) | きた (北)」が選ばれてしまう。ひらがなは異なり文字数が少ないので (< 100)、初期単語リストが大き

いほど、一つのひらがな単語が他のひらがな単語列と偶然に一致する可能性が高くなる。これが初期単語リストを大きくしても、ある時点で単語分割精度の向上しなくなる(または低下する)原因である。

6.3 再推定の効果の分類

再推定により生じる単語分割の変化は、単語境界の補正と細分割の2つに大きく分類される。前者は単語数は不変で単語境界だけが移動し、後者は一つの単語を2つ以上に分割する。

文末述語のひらがな列に関しては、最初の単語分割と正しい単語分割の単語数が同じ場合、再推定により単語分割が改善されることが多い。例えば、「連れ去 | られ | たま (玉) | まだ (副詞)」は「連れ去 | られ | た | まま | だ」に修正される。

最長一致の誤りに関しても、より短い個々の単語の頻度が非常に大きければ、再推定により改善される。例えば、「控え | たい (鯛)」は「控え | た | い」に修正される。

再推定による好ましくない変化の代表例は、低頻度単語を高頻度単語列(あるいは高頻度語と未知語)に分解してしまうことである。例えば、「使節」という単語は頻度が小さいので、共に頻度が大きい単語である「使」と「節」に分割されてしまう。

前節でも述べたように、再推定の効能は、語彙獲得ツールにより初期単語リストに付加された不適切な単語を削除することである。例えば、文字列「ソ連製戦車」は文字種により「ソ」と「連製戦車」に分割されて初期単語リストに加えられる。しかし、「ソ連」と「製」が単語リストにあるので、再推定後は「ソ」と「連製戦車」が単語リストから取り除かれ「戦車」が加えられる。

7 結論と課題

本稿では、小さな単語リストと大量のテキストから自己組織的な方法で確率的日本語単語分割プログラムを作成する方法を述べた。この単語分割プログラムの欠点は、未知語の一部が辞書中の単語と偶然一致した場合に未知語が分割されること、および、ひらがな表記された文末述語で誤った最長一致が起

りやすいことである。前者は短い単語に大きな確率を割り当てる傾向を持つ文字 unigram モデルを用いた単語モデルに起因する。後者は分割数最小の解を優先するという単語 unigram モデルに起因する。

本実験で単語 bigram や文字 bigram を使わなかった理由は、適当な初期値の推定法が存在しなかったからである。そこで本研究の次のステップとしては、単語 unigram モデルを用いた単語分割プログラムを用いて単語 bigram 頻度と文字 bigram 頻度の初期値を求め、これを再推定により洗練するという手続きを検討している。

参考文献

- [Chang et al., 1995] J. Chang, Y. Lin, and K. Su: Automatic Construction of a Chinese Electronic Dictionary, WVLC-95, pp.107-120, 1995.
- [EDR, 1995] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (第1版), 1995.
- [Elworthy, 1994] D. Elworthy: Does Baum-Welch Re-estimation Help Taggers? ANLP-94, pp.53-58, 1994.
- [Kupiec, 1992] J. Kupiec: Robust Part-of-Speech Tagging using a Hidden Markov Model. *Computer Speech and Language*, 6, pp.225-242, 1992.
- [Luo and Roukos, 1996] X. Luo and S. Roukos: An Iterative Algorithm to Build Chinese Language Models, ACL-96, pp.139-143, 1996.
- [松本ほか, 1994] 松本・黒橋・宇津呂・長尾: 日本語形態素解析システム JUMAN 使用説明書 version 2.0, 1994.
- [Manber and Myers, 1993] U. Manber and G. Myers: Suffix Arrays: A New Method for On-Line String Searches, *SIAM J. Comput.*, Vol.22, No.5, pp.935-948, 1993.
- [Nagata, 1994] M. Nagata: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, COLING-94, pp.201-207, 1994.
- [Nagata, 1997] M. Nagata: A Self-Organizing Japanese Word Segmenter using Heuristic Word Identification and Re-estimation, To appear in WVLC-97, 1997.
- [中渡瀬, 1996] 中渡瀬: 正規化頻度による形態素境界の推定, NL 研究 96-113-3, pp.13-18, 1996.
- [Sproat et al., 1996] R. Sproat, C. Shih, W. Gale, and N. Chang: A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, Vol.22, No.3, pp.377-404, 1996.
- [竹内・松本, 1997] 竹内・松本: 隠れマルコフモデルによる日本語形態素解析のパラメータ推定, 情報論, Vol.38 No.3, pp.500-509, 1997.
- [Wu and Tseng, 1993] Z. Wu and G. Tseng: Chinese Text Segmentation for Text Retrieval: Achievements and Problems, *Journal of ASIS*, Vol.44, No.9, pp.532-544, 1993.
- [Yamamoto, 1996] M. Yamamoto: A Re-estimation Method for Stochastic Language Modeling from Ambiguous Observations, WVLC-96, pp.155-167, 1996.