自然言語処理 137－6
音声言語情報処理 31－1
（2000．6．2）

統計的日本語形態素解析に対する拡張 HMM モデル

浅原 正幸　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{masayu-a,matsu}@is.aist-nara.ac.jp

我々は日本語形態素解析器『茶筌』のための学習ツールを開発している。現在『茶筌』では階層構造を持った品詞体系を採用し、タグの種類は約 500 にもなっている。このため、通常の品詞 tri-gram モデルの作成は困難で、品詞 bi-gram モデルでも適切な量のタグづけコーパスを得ることは難しい。通常、このような細かいタグを取り扱うために、複数のタグを同値類へとグループ化することによってタグの数を減らすことが行われる。我々はこれを拡張し、マルコフモデルの条件付き確率計算について各件でタグの同値類を変更するようにした。さらに、例外的な現象によるデータスパースネスに対処するため、単語レベルまで品詞として見るモデルと、選択的 tri-gram モデルを導入した。また、単語レベルまで品詞として見る場合には、単語－品詞間スムージング を導入した。『茶筌』にこれらのモデルを適用し、各拡張の評価を行った。

キーワード: 統計的形態素解析、機械学習、隠れマルコフモデル、階層的品詞体系、選択的 tri-gram、単語－品詞間 スムージング

Extended Hidden Markov Model for Japanese Morphological Analyzer

Masayuki Asahara　　Yuji Matsumoto
Graduate School of Infomation Science, Nara Institute Science and Technology
{masayu-a,matsu}@is.aist-nara.ac.jp

We are developing a learning tool for Japanese morphological analyzer, ChaSen. Currently we use a fine-grained POS tag set in which the number of the tag is about 500. We cannot apply a normal tri-gram model, and even a bi-gram model with a moderate size of annotated corpus. A usual technique to cope with such fine-grained tags is to reduce the size of tag set by grouping the set of tags into equivalent classes. We pursue it further and introduce the concept of *position-wise grouping* where the tag set is partitioned into different equivalent classes at each position in the conditional probabilities in Markov Model. Besides, to cope with data sparseness problem caused by exceptional phenomena, several other techniques such as word-level statistics and smoothing of word-level and POS-level statistics, and *selective tri-gram model.* We then give results of experiments to see the effect of the tools applied to ChaSen.

Keywords : Statistic Morphological Analysis, Machine Learning, Hidden Markov Model, Hiearchical Part of Speech, Selective tri-gram, Smoothing between words and part of speech

1　はじめに

近年大規模なタグ付きコーパスが利用できるようになり、多くの統計的形態素解析器が開発され高い精度と頑強性を達成できるようになった。一方、各ユーザや各言語のなかで十分な量のコーパスが得られておらず、学習モデルの改善需要は依然としてある。本論文では、このような需要に応える、拡張した統計モデルについて述べる。この拡張モデルは、可変長マルコフモデル [6] に基づいた統計的日本語形態素解析器『茶筌』 [9] に使用されている。以下では、拡張モデルの

概要について述べる。

『茶筌』では品詞体系として、IPA 品詞体系に少し手を加えたものを採用している。そのタグの数は 500 にもなる。助詞などのいくつかの単語については 1 単語を 1 品詞としてみなすため、実際のタグの数はさらに多い。タグの数が大きいために、tri-gram 連接規則を構築することが困難である。全てのタグについて異なるものとすると bi-gram 連接規則を構築することも難しい。

このような詳細なタグに対処する手法として、複数のタグを同値類へとグループ化し、タグの数を減らす方法 [4] がある。本稿では、新しく各件で異なるグループ化を導入した。タグ集合を、マルコフモデルの条件付き確率が各件で異なる同値類へと分割する。この手法は、多くの活用形を持つ日本語の形態素解析に非常に有効である。日本語は前の単語の活用形は後の単語に対して非常に重要であるが、逆に後の単語の活用形は前の単語に対しあまり重要ではないという特徴を持つ。また、話し言葉では、2 つ以上の形態素が 1 つの単語へと縮約するという多くの縮約表現が見られる。このような単語は、右から見るのと左から見るのとで異なる品詞に属するような振舞いを行う。前後で異なるグループ化を利用することにより、このような単語を各件によって異なる品詞と同一視することが可能となる。

大きなタグ集合を扱うときデータスパースネスの問題は常に重要な問題である。『茶筌』で採用しているタグ集合では、通常の tri-gram モデルを導入するのは非現実的である。そのため、我々は bi-gram モデルをベースとした、選択的 tri-gram モデルを導入した。選択的 tri-gram モデルとは、特別な連接だけを tri-gram 連接で記述し、通常の bi-gram モデルと統合するモデルである。この tri-gram モデルでも、データスパースネス問題を解決するために、bi-gram 連接とのスムージングを利用している。

これらの手法の併用により、適切なサイズのタグ付きコーパスから確率パラメータを学習し、統計的形態素解析器の性能を上げることができた。

第 2 節では、統計的形態素解析の基本概念とその問題点について述べる。第 3 節では、拡張モデルについて詳述する。第 4 節では、様々な条件でツールの評価を行う。第 5 節で関連研究を提示し、最後に第 6 節でまとめと今後の課題について述べる。

2 背景

2.1 統計的形態素解析の確率モデル

統計的形態素解析の一般的なモデルとして、隠れマルコフモデルが知られている。

形態素解析は入力文 S 単語列 $W = w_1, \ldots, w_n$ に対する品詞タグ列 $T = t_1, \ldots, t_n$ を決定することと定義できる。目標は次の確率値を最大にするような T を発見することである:

$$arg\max_T P(T|W)$$

ベイズの定理を利用して、$P(W,T)$ は品詞タグ生起確率列と単語の生起確率列として展開される:

$$\begin{aligned} arg\max_T P(T|W) &= arg\max_T \frac{P(T,W)}{P(W)} \\ &= arg\max_T P(T,W) \\ &= arg\max_T P(W|T)P(T) \end{aligned}$$

単語生起確率はその品詞タグからのみに、品詞タグ生起確率は bi-gram モデル (もしくは tri-gram モデル) のみに制限して近似をする:

$$\begin{aligned} P(W|T) &= \prod_{i=1}^{n} P(w_i|t_i) \\ P(T) &= \prod_{i=1}^{n} P(t_i|t_{i-1}) \end{aligned}$$

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

これらの値は、タグ付きコーパスの頻度から最尤推定を用いて推定される:

$$\begin{aligned} P(w_i|t_i) &= \frac{F(w_i,t_i)}{F(t_i)} \\ P(t_i|t_{i-1}) &= \frac{F(t_{i-1},t_i)}{F(t_{i-1})} \\ P(t_i|t_{i-2,i-1}) &= \frac{F(t_{i-2},t_{i-1},t_i)}{F(t_{i-2},t_{i-1})} \end{aligned}$$

このようにしてタグ付きコーパスから学習されたパラメータを利用して、単語列 W に最尤な品詞タグ列 T を決定する。品詞タグ列の決定は、動的計画法の一種である Viterbi algorithm による。

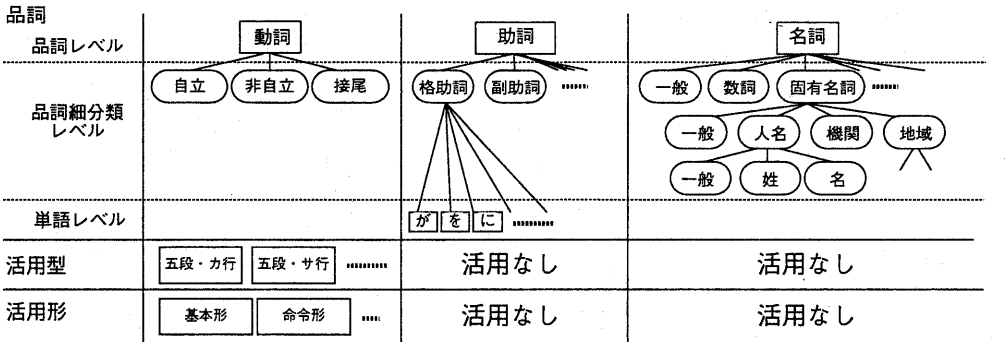

図 1: 階層的品詞体系

2.2 階層的品詞体系

『茶筌』では、IPA 品詞体系を少し変更した品詞タグ集合を利用している。この品詞タグ集合は、「品詞情報」、「活用型」、「活用形」の 3 つ の要素からなる。「活用型」、「活用形」の 2 つの情報は、活用語にしか現れない。

この品詞情報部分は階層構造を持っている。品詞情報は一番上のレベルで、15 のカテゴリーに分類されている (名詞、動詞、...)。「品詞情報」のそれ以下のレベルを「品詞細分類」と呼ぶ。例えば、名詞はさらに普通名詞 (一般)、固有名詞、数詞などに分類される。固有名詞はさらに、一般、人名、組織、地域に分割される。このように任意の深さの品詞分類レベルが記述することができる。最も下のレベルに単語を定義し、1 単語 を 1 品詞とみなすことが可能である。

動詞、形容詞、助動詞は活用語に属し、語尾変化する。これらの活用語は、特定の「活用型」の集合へと分類され、それぞれの「活用型」は「活用形」の集合を持っている。

日本語の「活用形」は後続する単語によって変化することが知られている。「活用形」は、 bi-gram もしくは tri-gram 連接の t_{i-1} の位置 ($P(t_i|t_{i-1})$ や $P(t_i|t_{i-2}, t_{i-1})$ の中の t_{i-1}) に出現する場合には重要な役割を果すが、他の位置の場合には区別する必要はない。図 1 に、品詞の階層構造を示す。

2.3 統計モデルの問題点

統計的自然言語処理におけるほとんどの問題は、学習データのスパースネスから起こる。上に示した品詞体系の場合、単語まで見ないモデルでも、タグの数は約 500 にもなる。bi-gram モデルでさえ、データスパースネスの問題は起きる。tri-gram モデルの場合はこの問題はさらに大きくなる。これらの問題はグループ化によってタグ集合を小さくすることにより改善することが期待される。

一方、さまざま例外的な言語現象も問題となる。いくつかの単語は、他の同じタグを振られた単語と異なる連接特性を持つ。このような例外は、単語単位、もしくは単語をグループ化した単位で、他の同じタグに属する単語の統計値と別に統計を取る必要がある。また、bi-gram 文脈では品詞決定ができない現象もある。本蔽では、これらの例外的現象をグループ化、単語レベルの統計、単語－品詞間 smooothing と 選択的 tri-gram によって取り扱う。それぞれの特徴については、次節で詳しく説明する。

3 統計モデルの拡張

本節では、上で述べた問題対処するために導入した統計モデルの拡張について概観する。

我々は確率モデルに対し、以下の 5 つの拡張を行った。

- 各件で独立したグループ化
- 単語レベルの統計値
- 単語－品詞間スムージング

- 選択的 tri-gram モデル
- コーパス中に現れない単語の推定

本稿では tri-gram 連接 t_{i-2}, t_{i-1}, t_i や bi-gram 連接 t_{i-1}, t_i に対し、t_{i-2} を前々件、t_{i-1} を前件、t_i を後件と呼ぶ。

3.1 各件で独立したグループ化

『茶筌』では、非常に細かいタグ集合を使っている。このために、確率パラメータの量を減らすために、タグ集合をいくつかの同値類へと分類することが重要になってくる。さらに、前節で述べた通り、いくつかの品詞 (もしくは単語) は、現れる位置によって、連接規則の上で異なる振舞いをする。例えば、「活用形」は後続の単語の曖昧性の解消に対して重要な役割を果たす。「活用形」は bi-gram もしくは tri-gram 連接の前件 t_{i-1} の位置に表われるもののみ考慮に入れれば良い。これは、動詞の統計値を取る際、その場所によって異なるグループ化をすべきであることを意味する。

また、口語表現には縮約表現が多く出現する。例えば、助動詞「ちゃう」は「て (助詞) +しまう (助動詞)」の 2 つの単語から構成される縮約形であり、他の単語とは別の振舞いをする。これらの振舞いを統計的に学習する方法として、その単語のさまざまな使用例を集め、正確にタグづけしたあと学習データに追加する方法がある。これに対して、各件で独立したグループ化を利用することにより、この問題に対して別の解を与えることができる。この単語について 条件付き確率 $P(t_i|t_{i-1})$ を計算する際、後件 t_i については「て」と同じ同値類にグループ化し、前件 t_{i-1} については「しまう」と同じ同値類にグループ化して、これらのクラスからその統計的振舞いを学習することが可能になる。

以下、各件で独立したグループ化について詳説する。簡単のため、bi-gram モデルについて説明する。$\mathcal{T} = \{A, B, \cdots\}$ を元の品詞タグ集合とする。このタグ集合に対し、2 つ のタグ集合を導入する。一つは 後件についてのタグ集合 $\mathcal{T}^c = \{A^c, B^c, \cdots\}$、もう一つは、前件についてのタグ集合 $\mathcal{T}^p = \{A^p, B^p, \cdots\}$ である。これらの集合間に、同値類を作成する写像を定義する。後件に対する写像 : $I^c(\mathcal{T} \to \mathcal{T}^c)$ および前件に対する写像 : $I^p(\mathcal{T} \to \mathcal{T}^p)$ を定義する。

図 2 では、これらの写像による、分割の例をしめす。この例の写像は次の通り:

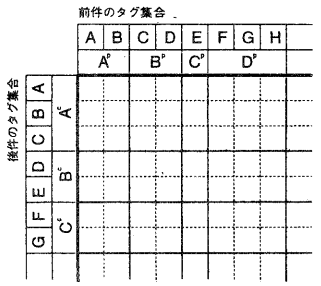

図 2: 各件で独立したグループ化

(前件)「形容詞-自立 * 体言接続」

形容詞-自立	形容詞・アウオ段	体言接続
形容詞-自立	形容詞・イ段	体言接続
形容詞-自立	形容詞・文語	体言接続

図 3: グループ化の例

$$
\begin{aligned}
I^c &= \{A \to A^c, B \to A^c, C \to A^c, \\
&\quad D \to B^c, E \to B^c, \ldots\} \\
I^p &= \{A \to A^p, B \to A^p, C \to B^p, \\
&\quad D \to B^p, E \to C^p, \ldots\}
\end{aligned}
$$

同値類のクラスを表現するために、タグ t が、後件で $[t]^c$ に属し、前件で $[t]^p$ に属しているとすると、単語生起確率、品詞生起確率は次式になる:

$$
\begin{aligned}
P(w_i|t_i) &= \frac{F(w_i, [t_i]^c)}{F([t_i]^c)} = \frac{F(w_i, t_i)}{F([t_i]^c)} \\
P(t_i|t_{i-1}) &= \frac{F([t_{i-1}]^p, [t_i]^c)}{F([t_{i-1}]^p)}
\end{aligned}
$$

図 3 に、グループ化の例を示す。この例では、「活用形」が共通する 3 つの品詞を同一視している。

3.2 単語レベルの統計値

いくつかの単語は、他の同じ品詞分類にある単語と異なった振舞いをする。特に、日本語の助詞、助動詞、接頭詞, 接尾詞などは、異なる文脈的振舞いをすることが知られている。これらに対し、単語を異なる品詞タグとして定義し、個別に統計値を取るように拡張した。

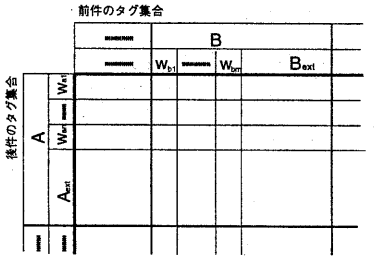

図 4: 単語レベルへの拡張

(後件)「形容詞-非自立よい」

形容詞-非自立	形容詞・アウオ段	ガル接続	よい
形容詞-非自立	形容詞・アウオ段	仮定形	よい
形容詞-非自立	形容詞・アウオ段	仮定縮約 1	よい
形容詞-非自立	形容詞・アウオ段	仮定縮約 2	よい
	⋮		
形容詞-非自立	形容詞・アウオ段	ガル接続	良い
形容詞-非自立	形容詞・アウオ段	仮定形	良い
形容詞-非自立	形容詞・アウオ段	仮定縮約 1	良い
形容詞-非自立	形容詞・アウオ段	仮定縮約 2	良い
	⋮		

図 5: 単語レベルでのグループ化の例

タグ集合 T を、いくつかの単語を個別の品詞タグ (単語レベル) として定義した新しいタグ集合 T^{ext} に拡張する。

注意すべき点は、品詞レベルの統計値は、いくつかの同じグループの単語を独立にした際に修正される点である。タグ集合 T 中に、品詞タグ A, B があると仮定する。品詞 A 中の単語 $w_{a_1}, \ldots, w_{a_n}$ と品詞 B 中の単語 $w_{b_1}, \ldots, w_{b_m}$ をタグ集合 T^{ext} では独立したとする。この時、タグ集合 T^{ext} 中の品詞タグ A_{ext}, B_{ext} は次のようになる:

$$
\begin{aligned}
A_{ext} &= A \setminus \{w_{a_1}, \ldots, w_{a_n}\} \\
B_{ext} &= B \setminus \{w_{b_1}, \ldots, w_{b_m}\}
\end{aligned}
$$

$A-B$ 連接確率を推定するために、頻度 $F(A,B)$ ではなく頻度 $F(A_{ext}, B_{ext})$ を利用する。図 4 に、上の条件での単語レベルの拡張を例示する。

これらのタグ集合の拡張は、各件で独立したグループ化の特別な場合である。全ての単語レベルのタグから タグ集合 T^{ext} への同値類写像が定義できる。写像 I^c は A_{ext} 中の全ての単語 を A_{ext} へと写像し、単語 $\{w_{a_1}, \ldots, w_{a_n}\}$ はそれぞれの単語自身へと写像する。同様に I^p も B_{ext} 中の全ての単語 を B_{ext} へと写像し、単語 $\{w_{b_1}, \ldots, w_{b_m}\}$ をそれぞれの単語自身へと写像する。

図 5 に、単語レベルでのグループ化の例を示す。後件では「活用形」は重要でないことから、全ての「活用形」を同一視している。また、表記の揺れ (漢字/かな) も、グループ化により吸収している。

3.3 単語ー品詞間スムージング

単語を独立したタグとして見るときに、その生起頻度が低い場合、十分な統計量を得るために事例を蓄積なければならない。別の解として、品詞レベルの統計値とのスムージングを考えることができる。単語の統計値のスパースネスを緩和するために、その単語の属する品詞の統計値を利用する。

ここで、2 つのスムージング係数を定義する。

λ_c を後件におけるスムージング率、λ_p を前件におけるスムージング率とする。これらの値は各単語毎に決定することができる。

単語 w_i を独立に統計を取るものとし、その品詞を t_i とする。もし後件にスムージングが適用されるとき、タグ生起確率は次のように定義される (w_i 自身が独立したタグである):

$$
\begin{aligned}
& P(w_i|t_{i-1}) \\
&= ((1-\lambda_c)P(t_i|t_{i-1}) + \lambda_c P(w_i|t_{i-1})) \\
&= \frac{(1-\lambda_c)F(t_{i-1},t_i) + \lambda_c F(t_{i-1},w_i,t_i)}{F(t_{i-1})}
\end{aligned}
$$

前件の単語についてスムージングを適用した場合は次のようになる (t_{i-1} を w_{i-1} の品詞であるとする):

$$
\begin{aligned}
& P(t_i|w_{i-1}) \\
&= (1-\lambda_p)P(t_i|t_{i-1}) + \lambda_p P(t_i|w_{i-1}) \\
&= (1-\lambda_p)\frac{F(t_{i-1},t_i)}{F(t_{i-1})} + \lambda_p \frac{F(w_{i-1},t_{i-1},t_i)}{F(w_{i-1},t_{i-1})}
\end{aligned}
$$

前件と後件両方の単語に適用した場合次のようになる:

$$
\begin{aligned}
& P(w_i|w_{i-1}) \\
= \; & (1-\lambda_p)((1-\lambda_c)P(t_i|t_{i-1}) + \lambda_c P(w_i|t_{i-1})) + \\
& \lambda_p((1-\lambda_c)P(t_i|w_{i-1}) + \lambda_c P(w_i|w_{i-1}))
\end{aligned}
$$

$$= (1-\lambda_p)(\frac{(1-\lambda_c)F(t_{i-1},t_i)+\lambda_c F(t_{i-1},w_i,t_i)}{F(t_{i-1})})+$$
$$\lambda_p(\frac{(1-\lambda_c)F(w_{i-1},t_{i-1},t_i)+\lambda_c F(w_{i-1},t_{i-1},w_i,t_i)}{F(w_{i-1},t_{i-1})})$$

3.4 選択的 tri-gram モデル

大きなタグ集合に対して、単純な tri-gram モデルを定義することは不可能である。しかし、品詞決定に tri-gram の文脈を必要とする場合がある。そこで、限定した tri-gram 連接のみを導入する。これを選択的 tri-gram モデルと呼ぶ。本モデルでは、これらの tri-gram 統計と bi-gram 統計とを混合して利用する。

異なる長さの文脈を混合する考え方は今までなかったわけではない。可変長マルコフモデルは Ron [6] により提案されている。Ron の手法は、n を変化させた n-gram の混合モデル で、学習アルゴリズムとともに提示されている。このようなモデルでは、文脈の集合 (有限状態集合) を、有限状態が決定的に明確にするために相互に分割すべきである。

ここで、我々は tri-gram 統計に対し少し異なった改良を行った。tri-gram 連接を例外的な文脈として考え、bi-gram 文脈と tri-gram 文脈とが交わりを持つ場合に、tri-gram 文脈を bi-gram 文脈に含まれる例外として考える。この考えでは、モデルの中で全ての文脈は相互に独立したものとして考える。そして、我々のモデルを Ron の定式へと変換することができる。長い文脈を短い文脈の例外として解釈する場合に、この定式化はより簡潔である。

後件のグループ化 (T^c) については、bi-gram における同じグループ化を共有することを仮定している。しかし、前件、前々件については、タグ集合に対して bi-gram と異なるグループ化が定義できる。先行する位置に対して、新しい 2 つ のタグ集合を導入する:

Tri-gram 連接に対する前件についてのタグ集合:

$$T^{p'} = \{A^{p'}, B^{p'}, \cdots\}$$

Tri-gram 連接に対する前々件についてのタグ集合:

$$T^{pp'} = \{A^{pp'}, B^{pp'}, \cdots\}$$

Tri-gram 連接について、前件に対しての同値類写像を $I^{p'}(T \to T^{p'})$ とし、前々件に対しての同値類写像を $I^{pp'}(T \to T^{pp'})$ とする。

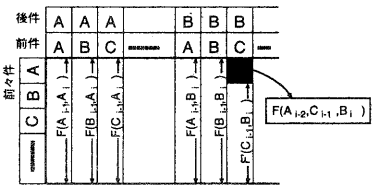

図 6: 選択的 tri-gram

(形容詞-＊＊連用テ接続)(助詞-係助詞 ＊＊は)(助動詞-ナイ)
(助動詞 特殊・ダ 連用形)(助詞-係助詞 ＊＊は)(助動詞-ナイ)
(助動詞 特殊・ダ 連用形)(助詞-係助詞 ＊＊も)(助動詞-ナイ)
(助動詞 特殊・ダ 連用形)(助詞-係助詞 ＊＊しか)(助動詞-ナイ)

図 7: 選択的 tri-gram の例

t に対して $I^{pp'}$ により定義された同値類を $[t]^{pp'}$ とする。tri-gram 連接確率を次のように定義する:

$$P(t_i|t_{i-2},t_{i-1}) = P([t_i]^c|[t_{i-2}]^{pp'},[t_{i-1}]^{p'})$$
$$= \frac{F([t_{i-2}]^{pp'},[t_{i-1}]^{p'},[t_i]^c)}{F([t_{i-2}]^{pp'},[t_{i-1}]^{p'})}$$

図 6 に tri-gram モデルの頻度計算を図示する。

Tri-gram 文脈を含む bi-gram 文脈 については、bi-gram の統計値はその tri-gram の統計値を外して計算される。

例えば、モデルに tri-gram 文脈 $A-C-B$ が含まれているとき、bi-gram 文脈 $C-B$ の統計値は次のようになる (F はコーパス中の実際の頻度を示し、F' は確率計算のために使用する、推定頻度とする):

$$F'(C,B) = F(C,B) - F(A,C,B)$$

Tri-gram 文脈の選択は簡単ではない。各 tri-gram 文脈をエラー率もしくはエラー数の多い順に出力し、この出力を見て人手で tri-gram 文脈を追加している。

図 7 に、選択的 tri-gram の例を示す。この例は、品詞決定に tri-gram 文脈を要する助動詞「ない」のためのものである。「＊」 は任意の細分類、任意の活用型、任意の活用形を表す。後件の「(助動詞-ナイ)」は、助動詞「ない」の全ての活用形、全ての表記を同一視したものである。この tri-gram 文脈により、係助詞に後置する「ない」の曖昧性を解消することができる。

3.5 コーパス中に現れない単語の推定

辞書の全ての単語が学習コーパスに出現しないため、未知語の生起確率を何らかの方法で割り当てる必要がある。未知の現象を推定するいくつかの方法が存在する。本モデルでは、リドストーン推定[1]を適用して、全ての観察される頻度を以下のようにして追加修正する。

$$P(w|t) = \frac{F(w,t) + \alpha}{\sum_{v \in t} F(v,t) + \alpha|t|}$$

現在、外部から足し込む単語のデフォルトの頻度 α を 0.5 としている。

4 評価

4.1 評価実験

提示してきた拡張が、通常の bi-gram を どのように改善してきたかを評価するために、いくつかの実験を行った。

前件については活用形に基づいて動詞をグループ化し、全ての助詞、助動詞、記号を単語まで見るものとし、各単語に属する品詞とのスムージングを行った。単語-品詞間のスムージング率は、全ての単語について 0.9 に固定した。選択的 tri-gram については、曖昧な助詞「の」と助動詞「ない」と「ある」についての区別をするために定義している。これは非常に単純な拡張ではあるが、学習ツールの効果を評価するのに十分であると考える。

評価では 5-fold cross evaluation を行った。タグ付きコーパスを学習データ (80%) と評価データ (20%) に分割する。実験は 5 回繰り返し、結果を平均した。全データサイズは 37490 文 922932 単語。

評価は次の 3 レベルで行った。

- level1: 単語境界のみ一致
- level2: 単語境界と品詞のトップレベルが一致
- level3: 品詞の全情報が一致

評価に際し以下の 5 つのモデルを作成した:

- D: 通常の bi-gram
- D_w: D + 助詞などの単語レベルでの統計
- D_{wg}: D_w + グループ化
- D_{ws}: D_w + 単語-品詞間スムージング
- D_{wgt}: D_{wg} + 選択的 tri-gram

モデルを評価するために、F 値を以下の式で定義する:

$$\begin{aligned} \text{再現率 } (Recall) &= \frac{\text{正解単語数}}{\text{コーパス中の単語数}} \\ \text{適合率 } (Precision) &= \frac{\text{正解単語数}}{\text{システム出力単語数}} \\ F_\beta &= \frac{(\beta^2+1) \cdot \text{再現率} \cdot \text{適合率}}{\beta^2 \cdot (\text{再現率} + \text{適合率})} \end{aligned}$$

それぞれのモデルに対し、F 値 ($\beta = 1$ とした) を学習データと評価データの各レベルについて評価した。結果を表 1, 2 に示す。

4.2 考察

結果から以下のようなことが言える。グループ化は評価データに対して良い結果を達成している。グループ化が学習データに対して不得手である。これは、グループ化が、学習データに対する過学習を緩衝することによると考える。

選択的 tri-gram は、level2, level3 において顕著な改善が見られた。通常の bi-gram モデルと比較すると、level3 で 0.4-0.5% 、level2 で 0.2-0.3% の改善がみられている。

単語－品詞間スムージングは学習データのグループ化を施したものを改善させることができた。しかし、他の環境では精度が改善されなかった。この実験ではスムージング率は全ての単語について固定されている。各単語について異なるスムージング率を設定することが必要であると考える。

5 関連研究

Cutting ら [2] は、パラメータの数を減らすために可能なタグ集合の上の同値類へのグループ化を提案している。そして、Schmid [7] は、この同値類上にスムージングを導入している。これらのクラスは品詞タグの分割に利用されており品詞タグの混合については考慮されていない。

Brill [1] は 変形規則による方法を提案している。Tri-gram 連接の選択については、我々の方法に似ている。

[1] ラプラス推定を拡張したもの。

表 1: 学習データ (F 値 %)

dataset	level1	level2	level3
D	98.84	98.36	97.36
D_w	98.96	98.58	97.81
D_{wg}	98.92	98.46	97.61
D_{ws}	98.96	98.58	97.80
D_{wgt}	98.92	98.55	97.70

表 2: 評価データ (F 値 %)

dataset	level1	level2	level3
D	98.69	98.12	96.91
D_w	98.75	98.24	97.22
D_{wg}	98.80	98.26	97.20
D_{ws}	98.76	98.27	97.23
D_{wgt}	98.78	98.35	97.27

Haruno ら [3] は、誤り駆動の手法に基づき複数の可変長規則を構築し、これらのモデルを混合する手法を提案している。この手法では、グループ化とスムージング技術は導入されていない。

北内ら [8] は、誤り駆動の方法でタグ集合の改善を決定する方法を示している。この方法は、タグ集合の階層構造レベルによってタグ集合を決定している。この手法では単語レベルの区別と階層構造を越えたグループ化は考慮に入れられていない。

6 まとめと今後の課題

本論文では、形態素解析のための統計モデルについてのいくつかの拡張を提案した。また、簡単な実験を行い、各拡張の効果を評価した。

いくつかの単語について個別に頻度を数え、品詞レベルの統計値とのスムージングを導入することにより、データスパースネスの問題を緩和することができた。各件毎のグループ化により、効果的な確率パラメータ環境の改善を達成することができた。選択的 tri-gram により、簡単に例外的な言語現象を記述することができるようになった。

今後の課題として、これらのモデルを自動的もしくは半自動的に改善できる方法を開発する予定である。例えば、誤り駆動による手法などが個別に頻度を計算する単語や、tri-gram 文脈の選択に応用できるだろう。

形態素解析器『茶筌』と学習ツールキットは以下の URI から入手できる。

`http://cl.aist-nara.ac.jp/lab/nlt/chasen/`

参考文献

[1] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, Vol. 21, No. 4, pp. 543–565, 1995.

[2] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[3] M. Haruno and Y. Matsumoto. Mistake-driven mixture of hierarchical tag context trees. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 230–237, July 1997.

[4] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[5] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[6] D. Ron, Y. Singer, and N. Tishby. Learning Probabilistic Automata with Variable Memory Length. In *COLT-94*, pp. 35–46, 1994.

[7] H. Schmid. Inprovements in part-of-speech tagging with an application to german. In *EACL SIGDAT workshop*, pp. 47–50, 1995.

[8] 北内啓, 宇津呂武仁, 松本裕治. 誤り駆動型の素性選択による日本語形態素解析の確率モデ ル学習. 情報処理学会論文誌, Vol. 40, No. 5, pp. 2325–2337, 5 1999.

[9] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム 「茶筌」 version 2.0 使用説明書 第二版, 12 1999.