自然言語処理 124－6
（1998．3．13）

誤り駆動型の確率モデル学習による日本語形態素解析

北内 啓　宇津呂 武仁　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{akira-k,utsuro,matsu}@is.aist-nara.ac.jp

本研究では、日本語形態素解析の確率モデル学習におけるパラメータ推定の精度を上げるための有効な品詞分類を自動的に学習した。解析誤りをもとに詳細化する品詞分類を素性として取り出し、品詞分類を段階的に細かくしていく。学習によって得られた品詞分類を用いて bi-gram のマルコフモデルに基づくパラメータ推定を行うことにより、形態素解析の精度を向上させた。
実験により、人手で調整して決めた品詞分類に比べ、より少ないパラメータ数でより高い精度を得ることができた。また、品詞分類によってパラメータ数や精度がどのように変化するかといった、品詞分類全体の性質をとらえることができた。

[キーワード] 形態素解析, 確率モデル, 誤り駆動, コーパス

Error-driven Model Learning of Japanese Morphological Analysis

KITAUCHI Akira　UTSURO Takehito　MATSUMOTO Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

This paper proposes a method of an learning optimal set of part-of-speech tags which gives the highest performance in morphological analysis. In our method, considering patterns of errors in the morphological analysis, first, candidates of more specific part-of-speech tags to be included in the model of morphological analyzer are generated. Then, the most effective candidate which gives the greatest decrease in errors is employed. In the experimental evaluation of the proposed method, we achieve a morphological analyzer of higher performance compared with a model with a hand-tuned set of part-of-speech tags, and with much smaller number of parameters.

[keyword] Morphological Analysis, Stochastic Model, Error Driven Model Learning, Corpus

1　はじめに

形態素解析は自然言語処理の中で最も基本的な技術であり、応用範囲も広く、従来から多くの形態素解析システムが開発され、実際に使用されてきた。しかし、その多くは文法や辞書の整備を人手で行うもので、メンテナンスに手間がかかる。特に、単語のコスト値や品詞どうしの連接のしやすさなどのパラメータを人手で調整するのは非常に困難である。

そのような手間を軽減するため、品詞タグ付コーパスを用いて形態素解析のための言語的特徴を統計的に学習し、解析精度を向上させるという手法が行われるようになってきている。しかし、これらの手法の多くは、文法体系など、ある決められた枠組みの中でのパラメータ値を自動的に求めるというもので、品詞分類など文法体系そのものを動的に決定するという手法はあまり行われていない。

例えば、日本語の品詞分類は活用をもつため、活用型や活用形のすべての組合わせを考えると品詞の種類は数百に及ぶ。さらに、助詞や助動詞などは語彙レベルでもコーパス中の分布が異なり、膨大な種類の品詞分類を考慮する必要がある。これほど種類の多い品詞について高い精度が得られるような品詞分類を人手で求めるのは、非常に手間がかかる。

そこで本研究では、品詞タグ付コーパスを用いた日本語形態素解析の統計的学習において、パラメータ推定の精度を向上させるのに有効な品詞分類を自動的に求め、その品詞分類を用いてパラメータ推定を行った。その結果、人手で決めた適当な品詞分類に比べ、少ないパラメータ数でより高い精度を得ることができた。

本件研究の手法では、ひとつひとつの品詞分類を素性とみなし、解析誤りをもとに素性集合を求めていくことで品詞分類を決定する。具体的には、まず初期状態としてかなり粗い品詞分類から学習を開始し、解析誤りの多い品詞や単語に注目して素性を抽

出する。抽出した素性を一時的に追加した素性集合を用いてパラメータ推定を行い、解析精度が十分上がる場合はその素性を正式に素性集合に追加する。このように素性の抽出と追加の手順を繰り返すことにより、段階的に品詞分類を細かくしていき、最終的に精度がどれくらい向上するかを測定する。

本研究のパラメータ推定の方法は、基本的には bi-gram のマルコフモデルに基づいている。しかし、コーパス中で特定の品詞連接が特徴的に出現する場合など、ある連接部分の連接確率とその連接以外の部分の連接確率を別々に求めたいときがある。本研究ではこの部分の連接確率も求められるようにすることで、解析精度を向上させた。

2 誤り駆動型の確率モデル学習による日本語形態素解析

2.1 日本語形態素解析の確率モデル

本研究の確率モデルは、基本的には N-gram のマルコフモデルに基づいている。N-gram のマルコフモデルにおける日本語形態素解析は

> 与えられた入力文 S に対する単語列 $W = w_1 \cdots w_n$ と品詞列 $T = t_1 \cdots t_n$ の同時確率 $P(W,T|S)$ を最大にするような単語列と品詞列の組 $(\hat{W},\hat{T})$ を求める。

という問題に帰着され、$P(W,T|S)$ は以下の式によって与えられる。

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

$P(w_i|t_i)$ が単語生起確率、$P(t_i|t_{1,i-1})$ が品詞連接確率である。本研究では bi-gram のマルコフモデルを用いており、品詞連接確率 $P(t_i|t_{1,i-1})$ は以下のように近似される。

$$P(t_i|t_{1,i-1}) \approx P(t_i|t_{i-1})$$

本研究では、パラメータ推定を行う際の品詞分類に基づく品詞の連接を素性と呼ぶ。すなわち、前件と後件の品詞分類を素性の集合ととらえることができる。品詞分類を決定することは、式 (1) において前件の品詞 t_{i-1} や後件の品詞 t_i の分類を決定することに相当する。解析誤りをもとに最適な素性集合を求め、その素性集合が表す品詞分類のもとで単語生起確率と品詞連接確率のパラメータ推定を行う。

パラメータ推定は、まず bi-gram のマルコフモデルと同じように、以下のように最尤推定法によって確率値を求める。

$$P(w_i|t_i) = \frac{C(w_i)}{C(t_i)} \qquad (2)$$

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})} \qquad (3)$$

ここで、前件の品詞を s_i とし、前件の品詞分類を後件の品詞分類と関係なく自由に決定できるようにする。すると、連接確率は式 (4) のようになる。

$$P(t_i|s_{i-1}) = \frac{C(s_{i-1},t_i)}{C(s_{i-1})} \qquad (4)$$

式 (4) の連接確率は、式 (3) の場合と同様に最尤推定法によって導くことができる。

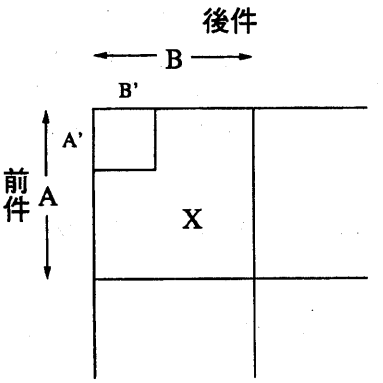

図 1: 特定の連接の連接確率を別に求める場合

次に、例外的な連接に対するパラメータ推定について述べる。例えば、図 1のように粗い品詞分類の連接 (A,B) に対し、細かい品詞分類の連接 (A',B') だけを特別視して、(A',B') の部分の連接確率と、(A,B) から (A',B') を取りのぞいた部分の連接確率を別々に求めることを考える。C_c を連接の総出現数、C_w を単語の総出現数とおくと、$C_c = C_w$ であるから、(A',B') の同時確率 $P(A',B')$ と連接確率 $P(B'|A')$ は以下のように求められる。

$$\begin{aligned} P(A',B') &= \frac{C(A',B')}{C_c} \\ P(B'|A') &= \frac{P(A',B')}{P(A')} = \frac{C(A',B')}{C_c} \Big/ \frac{C(A')}{C_w} \\ &= \frac{C(A',B')}{C(A')} \end{aligned}$$

ここで、$(A-A',B')$,$(A',B-B')$,$(A-A',B-B')$ における連接確率がすべて同じ確率値 k であるとする。すなわち、

$$\begin{aligned} P(B'|A-A') &= P(B-B'|A') \\ &= P(B-B'|A-A') = k \end{aligned}$$

$$\frac{P(A-A',B')}{P(A-A')} = \frac{P(A',B-B')}{P(A')}$$
$$= \frac{P(A-A',B-B')}{P(A-A')} = k$$

同時確率の総和を考えると、

$$P(A',B') + P(A-A',B') + P(A',B-B')$$
$$+ P(A-A',B-B') = P(A,B)$$

であるから、$C_c = C_w$ より

$$k \times (P(A-A') + P(A') + P(A-A'))$$
$$= P(A,B) - P(A',B')$$
$$k \times \frac{(C(A-A') + C(A') + C(A-A'))}{C_w}$$
$$= \frac{C(A,B) - C(A',B')}{C_c}$$
$$\text{ゆえに } k = \frac{C(A,B) - C(A',B')}{C(A-A') + C(A') + C(A-A')}$$

一般的な場合も同様に考え、互いに共通部分をもたない n 個の連接 $(A_1,B_1),\cdots,(A_n,B_n)$ に対し $X = \bigcup_{i=1}^{n}(A_i \times B_i)$ をひとつの領域と見て連接確率を求めると以下のようになる。

$$P(B_1|A_1) = \ldots = P(B_n|A_n)$$
$$= \sum_{i=1}^{n} C(A_i,B_i) / \sum_{i=1}^{n} C(A_i)$$

活用形をもつ品詞については、単語生成確率に対して近似を行う。基本的には、品詞 $T=(p,t,f)$, 形態素 $W=(p,t,f,w)$ に対し、

$$P(W|T) = \frac{C(W)}{C(T)} = \frac{C(p,t,f,w)}{C(p,t,f)}$$

のように確率値を計算する。しかし、本研究で利用している日本語形態素解析システム茶筌 [3] は活用形ごとに単語生成確率をもつことができず、すべての活用形に共通な単語生成確率をひとつだけ持つ。そこで、$P(W|T)$ を以下のように近似する。

$$P(W|T) = \frac{C(p,t,f,w)}{C(p,t,f)} \approx \frac{C(p,t,w)}{C(p,t)} \qquad (5)$$

2.2 誤り駆動型の素性選択

本研究では、品詞分類を粗くした状態を初期状態とし、解析誤りをもとに詳細な品詞分類を素性として抽出、追加していくことにより、品詞分類を細かくしていき最適な素性集合を学習した。ここではその学習方法について説明する。この方法以外にも、細かい品詞分類を初期状態として、品詞分類を段階的に粗くしていくことにより素性集合を学習していく方法などが考えられる。

2.2.1 概要

学習には以下の 4 種類のコーパスを用いる。

1. 素性選択時パラメータ推定用訓練コーパス A
   学習時、パラメータ推定を行うために用いる。
2. 素性選択用訓練コーパス B
   学習時、解析誤りから素性を抽出、追加するために用いる。
3. 評価時パラメータ推定用訓練コーパス C
   学習の結果得られた素性集合を用いて評価を行う際、訓練コーパスとしてパラメータ推定を行うために用いる。
4. 評価用テストコーパス D
   学習の結果得られた素性集合を用いて評価を行う際、テストコーパスとして解析を行うために用いる。

学習は以下の手順で行う。

1. 初期化
   まず、初期の品詞分類を用いてコーパス A でパラメータ学習を行い、コーパス B を解析する。解析時のパラメータ値には、コーパス A で学習した単語生成確率と品詞連接確率のほかに、コーパス A,B,C,D に含まれるすべての単語にごく小さな単語生成確率を与えたものを使う。
2. 素性候補の選択と追加
   以下の手順を、適当な回数だけ、あるいは抽出する素性がひとつもなくなるまで繰り返す。
   (a) 素性候補の選択
   コーパス B を解析した結果とコーパス B に付与されている品詞とを比較して得られた解析誤りをもとに、素性 (品詞連接) の候補を抽出する。
   (b) 素性の追加
   素性候補の中から適当に取り出した素性を一時的に素性集合に追加し、コーパス A でパラメータ推定を行い、コーパス B を解析する。解析時のパラメータ値には、コーパス A で学習した単語生成確率と品詞連接確率のほかに、コーパス A,B,C,D に含まれるすべての単語にごく小さな単語生成確率を与えたものを使う。
   解析の結果、十分に精度が上がっていれば追加した素性を正式に採用して素性集合に追加する。
3. 評価

コーパスと解析結果の例

解析文	コーパス				解析結果
さっぱり	副詞 - 助詞類接続	*	*	さっぱり	←(同左)
と	助詞 - 副詞化	*	*	と	助詞 - 格助詞 - 引用 * * と
おいしい	形容詞 - 自立	形容詞・イ段	基本形	おいしい	←(同左)
炒め	動詞 - 自立	一段	連用形	炒める	名詞 - 一般 * * 炒め物
物	名詞 - 接尾 - 一般	*	*	物	
です	助動詞	特殊・デス	基本形	です	←(同左)
。	記号 - 句点	*	*	。	←(同左)

解析誤りから取り出された連接の例

前件				後件			
副詞 - 助詞類接続	*	*	さっぱり	助詞 - 格助詞 - 引用	*	*	と
副詞 - 助詞類接続	*	*	さっぱり	助詞 - 副詞化	*	*	と
助詞 - 格助詞 - 引用	*	*	と	形容詞 - 自立	形容詞・イ段	基本形	おいしい
助詞 - 副詞化	*	*	と	形容詞 - 自立	形容詞・イ段	基本形	おいしい
形容詞 - 自立	形容詞・イ段	基本形	おいしい	動詞 - 自立	一段	連用形	炒め
形容詞 - 自立	形容詞・イ段	基本形	おいしい	名詞 - 一般	*	*	炒め物
名詞 - 一般	*	*	炒め物	助動詞	特殊・デス	基本形	です

図 2: 解析結果、および解析誤りから取り出された連接の例

学習の結果得られた素性集合を用いて、コーパスCでパラメータ学習を行い、コーパスDを解析し、最終的にどのくらい精度が向上したか測定する。

学習時と同様、解析時のパラメータ値には、コーパスCで学習した単語生成確率、品詞連接確率のほかに、コーパスA,B,C,Dに含まれるすべての単語にごく小さな単語生成確率を与えたものを使う。

2.2.2 初期の素性集合

学習を始めるにあたり、まず最初の素性集合、すなわち前件と後件の品詞分類を決めておく必要がある。素性集合は品詞細分類、活用型、活用形の3個の要素を組み合わせることで決定される。

これらの3個の要素を組み合わせることで様々な素性集合を作り出し、初期の素性集合を決定する。同様に、実験で抽出する素性も様々なものが考えられる。特定の品詞連接をひとつの素性として区別することもできる。

2.2.3 素性候補の抽出

コーパスAを用いてパラメータ推定した後、コーパスBを解析し、その解析誤りをもとに素性候補を抽出する。まず形態素の誤りを前件・後件の組として取りだし、現在の品詞分類よりも細かい分類になるように素性の候補を抽出する。

形態素の誤りを連接として取り出すには、誤りのある形態素の前後のすべての連接を抜き出せばよい。例えば、コーパスと解析結果の形態素がそれぞれ

コーパス $m_1\ m_2\ m_3\ m_4$
解析結果 $m_1\ m_2'\ m_3'\ m_4$

のようになっており、m_2 と m_2'、m_3 と m_3 はそれぞれ単語が同じで品詞が異なっていた場合、

- (m_1, m_2), (m_1, m_2')
- (m_2, m_3), (m_2, m_3'), (m_2', m_3), (m_2', m_3')
- (m_3, m_4), (m_3', m_4)

の8個の連接が取り出される。図2に例をあげる。

前件	(助詞 - 格助詞 - 引用)
後件	(形容詞 - 自立 形容詞・イ段 基本形 おいしい)

図 3: 素性候補のもととなる連接の例

このように取り出されたすべての連接について、現在の品詞分類よりも細かい分類の素性の候補を抽出する。例えば、現在の前件・後件の品詞分類がそれぞれ (助詞 - 格助詞 - 引用) と (形容詞 - 自立 形容詞・イ段 *) だった場合、図3のような連接に対して抽出する素性を考える。素性候補を抽出する方法は実験によって異なり、大きく分けて以下の3通りの方法がある。

1. 連接の片方のみを素性候補として抽出
   前件の品詞分類と後件の品詞分類を組み合わせた連接を用いてパラメータ学習を行う際、連接のうち前件あるいは後件の片方を素性候補として抽出する。素性集合は前件と後件を組み合わせたものになっているので、例えば、前件の品詞 a を素性候補として抽出した場合は、実際には現在の後件の品詞分類 $B = \{b_1, \cdots, b_n\}$ に対し、$\{(a, b_1), \cdots, (a, b_n)\}$ の n 個の素性を抽出したことになる。
   現在の品詞分類より細かく、誤り形態素連接と同じかそれよりも粗い品詞分類を素性として抽

出する。図3の連接に対しては、図4の素性が抽出される。

前件
助詞 - 格助詞 - 引用　＊　＊　と

後件			
形容詞 - 自立	形容詞・イ段	基本型	
形容詞 - 自立	形容詞・イ段	＊	おいしい
形容詞 - 自立	形容詞・イ段	基本型	おいしい

図 4: 抽出された素性 (連接の片方のみ) の例

2. 連接の片方と同レベルの素性すべてを抽出
   1 と同じように、連接のうち前件あるいは後件の片方を素性候補として抽出し、同時に同レベルの品詞分類を素性として抽出する。品詞細分類、活用型、活用形、語彙の 4 通りの要素を組み合わせることにより、様々な品詞分類を考えることができる。図3の連接に対しては、図5の素性が抽出される。

前件
助詞 - 格助詞 - 引用　＊　＊　と

後件			
形容詞 - 自立	形容詞・イ段	語幹	＊
形容詞 - 自立	形容詞・イ段	基本形	＊
	...		
形容詞 - 自立	形容詞・イ段	＊	大きい
形容詞 - 自立	形容詞・イ段	＊	美しい
	...		
形容詞 - 自立	形容詞・イ段	語幹	大きい
形容詞 - 自立	形容詞・イ段	基本形	大きい
	...		
形容詞 - 自立	形容詞・イ段	語幹	美しい
形容詞 - 自立	形容詞・イ段	基本形	美しい
	...		

図 5: 抽出された素性 (連接の片方と同レベル) の例

3. 連接そのもの、つまり前件と後件のペアを素性として抽出
   特殊な連接を前件と後件のペアの形で抽出し、その部分の連接確率だけをパラメータ学習する。品詞細分類、活用型、活用形、語彙を組み合わせた分類を素性として追加する。
   現在の品詞分類が (A, B) のとき、誤り形態素連接 (a, b) に対し、「A と同じかそれよりも細かく、a と同じかそれよりも粗いすべての品詞からなる集合」を X、「B と同じかそれよりも細かく、b と同じかそれよりも粗いすべての品詞からなる集合」を Y とすると、抽出される素性は、$X \times Y - (a, b)$ となる。

図3の連接に対しては、図6の素性候補が抽出される。

2.2.4 最適素性の選択方法

解析誤りを元に素性の候補を取り出した後、以下の手順によって最適素性を選択し、素性集合に追加していく。学習全体の手順とともに説明する。

1. 初期化
   以下のように初期値を設定する。

   試験済み素性集合 $F_t = \phi$

   追加済み素性集合 $F_a = \phi$

   パラメータ推定用素性集合 $F_p = F_{pinit}$

   まず、品詞分類 F_p を用いてコーパス A でパラメータ学習を行い、コーパス B を解析する。
2. 素性候補の選択
   コーパス B を解析した結果とコーパス B に付与されている品詞とを比較して得られた解析誤りをもとに、素性候補の集合 F_c を抽出する。
3. 素性候補の追加 (繰り返し)
   以下の処理を停止条件 1 または停止条件 2 を満たすまで繰り返す。

   (a) F_c に含まれ F_t に含まれていない素性のうち、最も頻度の多い素性候補をひとつ取り出し、f とする。

   (b) 素性 f を素性集合 F_p に一時的に加え、コーパス A を用いてパラメータ推定を行い、コーパス B を解析する。

   (c) 素性 f を F_t に加える。

   停止条件 1 $F_c \subset F_t$ になったとき。
   停止条件 2 今回の解析結果の誤り形態素数と、前回素性集合に素性を正式に追加したときの誤り形態素数との差を求める。減少した誤り形態素数が 3 個以上であり、かつ今回減少した誤り形態素数が前回減少した誤り形態素数の 1/10 よりも多かった場合、すなわち

   (今回減少した誤り形態素数) ≥ 3 かつ
   (今回減少した誤り形態素数)
   $>$ ((前回減少した誤り形態素数)/10

   を満たしていた場合は精度が十分に上がったとみなし、素性 f を正式に素性集合 F_p に追加、さらに f を F_a にも追加し 2 へ戻る。
4. $F_a \neq \phi$ の場合は $F_a = F_t = \phi$ として 2 へ戻り、$F_a = \phi$ の場合はすべての学習を終了する。

前件				後件			
助詞 - 格助詞 - 引用	*	*		形容詞 - 自立	形容詞・イ段	基本型	*
助詞 - 格助詞 - 引用	*	*		形容詞 - 自立	形容詞・イ段	*	おいしい
助詞 - 格助詞 - 引用	*	*		形容詞 - 自立	形容詞・イ段	基本型	おいしい
助詞 - 格助詞 - 引用	*	*	と	形容詞 - 自立	形容詞・イ段		
助詞 - 格助詞 - 引用	*	*	と	形容詞 - 自立	形容詞・イ段	基本型	*
助詞 - 格助詞 - 引用	*	*	と	形容詞 - 自立	形容詞・イ段	*	おいしい
助詞 - 格助詞 - 引用	*	*	と	形容詞 - 自立	形容詞・イ段	基本型	おいしい

図 6: 抽出された素性 (連接ペア) の例

3 実験と考察

3.1 人手によって調整した素性集合

学習によって獲得した素性集合との比較のために、人手で素性集合をいくつか作成した。品詞分類の細かさを適当に調節し、4個の素性集合を用意した。

Fm_1 品詞大分類のみを区別した素性集合。品詞細分類、活用型、活用形は区別していない。

Fm_2 品詞細分類までを区別した素性集合。活用型、活用形は区別していない。

Fm_3 品詞細分類、活用型、活用形のすべてを区別した素性集合。

Fm_4 品詞細分類、活用型、活用形のすべてを区別し、さらに助詞と助動詞については語彙まで区別した素性集合。

再現率、適合率、連接規則のパラメータ数について、学習して得られた素性集合と比較する。

3.2 実験

実験には、4種類のコーパス (素性選択時パラメータ推定用訓練コーパス A、素性選択用訓練コーパス B、評価時パラメータ推定用訓練コーパス C、評価用テストコーパス D) それぞれについて、毎日新聞 95 年度の記事約 35,000 文 (約 3000 記事、約 96 万形態素、約 45 万字) の品詞タグ付きコーパスから異なる 3000 文 (約 9 万形態素) ずつを無作為に取り出したものを用いた。

本研究の目的は「なるべく少ないパラメータ数で高い精度を得ること」である。そこで、少ないパラメータ数と高い精度のそれぞれに重点をおいた2通りの実験を行った。それぞれの目的に応じて、初期の素性集合と抽出する素性を変化させ、2種類ずつ学習を行った。

3.2.1 少ないパラメータ数を目的とした実験

はじめに、パラメータ数を極力少なくすることに重点をおき、その上で高い精度を得ることを目的とした実験を行った。実験では、以下の二種類の学習によって段階的に精度を向上させた。

1. 品詞レベルで連接の片方を素性として追加
   まず、前件の品詞と後件の品詞を掛け合わせた品詞分類で高い精度を示すような素性集合を求めるために、品詞レベルで連接の片方を素性として追加していく実験を行った。ただし、助詞と助動詞は語彙レベルのものも追加した。
   初期の素性集合として、品詞大分類のみを区別した品詞分類を用いた。ただし、品詞細分類と活用の両方をもつ動詞と形容詞については、抽出する素性候補があまりにも多くなってしまうため、学習の都合上 (動詞 - 自立), (動詞 - 非自立) などのように品詞細分類までを区別した。
   学習で得られた素性集合を Fa_1 とする。素性を追加するごとに、コーパス C でパラメータ推定しコーパス D を解析した。その再現率と適合率の変化を図 7に示す。
2. 語彙レベルを含む連接のペアを素性として追加
   次に、1の学習によって得られた素性集合 Fa_1 を初期の素性集合として、さらに精度を上げるために語彙レベルで連接のペアを素性として追加していく実験を行った。素性を 8 個追加したところで学習を中止し、学習で得られた素性集合 Fa_2 を用いてコーパス C でパラメータ推定し、コーパス D を解析した。

3.2.2 高い精度を目的とした実験

次に、高い精度を得ることに重点をおいた実験を行った。この実験ではパラメータ数についてはいっさい気にせず、精度の向上のみに目標を絞った。以下の二種類の学習によって段階的に精度を向上させた。

1. 品詞レベルで連接の片方と同レベルの品詞を素性として追加
   まず、前件の品詞と後件の品詞を掛け合わせた品詞分類でなるべく高い精度を示すような素性集合を求めるために、品詞レベルで連接の片方と同レベルの品詞を素性として追加していく実験を行った。

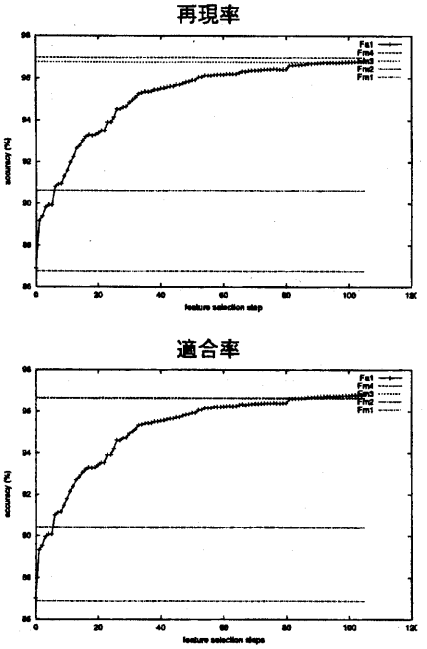

図 7: Fa_1 における実験結果

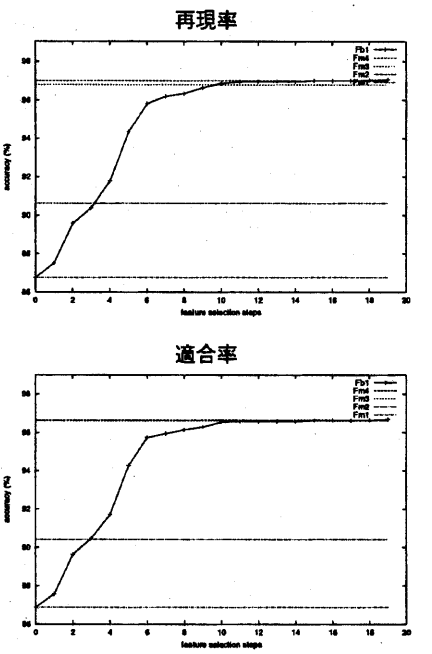

図 8: Fb_1 における実験結果

初期の素性集合として、品詞大分類のみを区別した品詞分類を用いた。段階的に細かい素性を抽出していくことにより最適な素性を決定した。

学習で得られた素性集合を Fb_1 とする。素性を追加するごとに、コーパス C でパラメータ推定しコーパス D を解析した。その再現率と適合率の変化を図 8に示す。

2. 語彙レベルを含む連接のペアを素性として追加

次に、1 の学習によって得られた素性集合 Fb_1 を初期の素性集合として、さらに精度を上げるために語彙レベルで連接のペアを素性として追加していく実験を行った。素性を 17 個追加したところで学習を中止、学習で得られた素性集合 Fb_2 を用いてコーパス C でパラメータ推定し、コーパス D を解析した。

3.2.3 人手で作成した素性集合との比較

人手で適当に決めた素性集合と、学習によって得られた素性集合の比較を表 1に示す。それぞれの素性集合について、再現率、適合率、連接確率のパラメータ数を測定した。

3.3 考察

学習による方法は人手で作成した素性集合に比べ、全般的に連接規則のパラメータ数が少なく、高い解析精度を得ることができた。

品詞レベルで連接の片方を素性として追加した素性集合 Fa_1 は、再現率は Fm_3 と Fm_4 の間、適合率は Fm_4 よりも精度が高く、パラメータ数は Fm_3 より少なかった。また、品詞レベルで連接の片方と同レベルの品詞を素性として追加した素性集合 Fb_1 は再現率、適合率とも $F4$ よりも精度が高かった。パラメータ数は、Fm_3 よりわずかに多く、Fm_4 の 3/4 以下であった。

連接ペアを素性として追加した素性集合 Fa_2, Fb_2 は、Fa_1, Fb_1 よりわずかに精度を向上させることができた。パラメータ数はともに Fm_3 よりわずかに多いが、Fm_4 よりはかなり少ない。

また、Fa_1 や Fb_1 は追加する素性数が少ない段階でかなり高い精度に達しており、これ以上素性を追加してもパラメータ数が増えるだけで、精度はほとんど上がらないことが予測できる。実際、Fa_2 や Fb_2 で素性を追加してもあまり精度が上がっていない。

このように、追加する素性数やパラメータ数の変化が精度の向上とどのように関わっているかなど、全体的な様子が分かるという点も品詞分類を自動的に学習することの利点としてあげられる。

素性集合	再現率	(誤り / 総形態素数)	適合率	(誤り / 総形態素数)	パラメータ数
Fm_1	86.765%	(10335 / 78087)	86.882%	(10230 / 77982)	111
Fm_2	90.619%	(7325 / 78087)	90.414%	(7502 / 78264)	1389
Fm_3	96.783%	(2512 / 78087)	96.610%	(2652 / 78227)	2942
Fm_4	96.996%	(2346 / 78087)	96.645%	(2629 / 78370)	4103
Fa_1	96.870%	(2444 / 78087)	96.838%	(2470 / 78113)	2873
Fa_2	96.959%	(2375 / 78087)	96.916%	(2409 / 78121)	2881
Fb_1	97.064%	(2293 / 78087)	96.676%	(2606 / 78400)	2986
Fb_2	97.084%	(2277 / 78087)	96.682%	(2602 / 78412)	3000

表 1: 人手で作成した素性集合と学習で得られた素性集合の比較

4 関連研究

本研究の特徴として 1. タグ付コーパスを用いた統計的学習に基づく形態素解析を行う、2. 形態素解析の精度を上げるのに有効な素性を抽出、選択し、モデルに組み込んでいく、の 2 点があげられる。このような手法を用いた研究はほかにもいくつかある。

Brill [1] は、変形規則と呼ばれる操作を適用していくことにより形態素解析を行い、変形規則の中で解析誤りを最も減らすような変形規則を追加していくことによって精度を向上させている。この研究は、品詞タグ付きコーパスを用いた学習を行い、解析誤りをもとに素性を追加していくという点では本研究の手法と似ているが、形態素解析の方法は確率モデルに基づくものではない。また、品詞分類は最初から決められており、品詞分類の細かさをどうするかはこの論文では問題とされていない。

柏岡ら [4] は、形態素の構成の特徴、単語の分類体系上の特徴、文脈による特徴という 3 つの観点からとらえた属性を利用して学習した確率付決定木を用いて日本語形態素解析を行っている。形態素解析の手法は本研究と同様確率モデルに基づいているが、この手法では前後の単語の情報だけではなく、部分的な文字列の特徴や文頭の形態素の品詞など、一般的な N-gram モデルの形態素解析では扱われないような情報を扱っている。属性の選択の基準としてはエントロピーを用いており、本研究の誤り駆動の手法とは異なる。

春野ら [2] は、誤りが多い部分に注目してコーパスの分布を変えることを繰り返すことにより複数の確率モデルを学習し、それらのモデルを混ぜ合わせることで日本語形態素解析の精度を向上させている。確率モデル学習には文脈木を用いたマルコフモデル学習を行っている。この手法は、誤り率が高い単語についてコーパス中の頻度に重みづけしており、解析誤りの頻度が大きいものから素性を追加する本研究の手法と異なる。また、本研究が bi-gram のマルコフモデルのみを扱っているのに対し、この研究では tri-gram 以上を含む variable-gram のマルコフモデルを用いている。この研究は品詞分類の詳細化については語彙化のみを行っているが、本研究では品詞階層や活用型・活用形の組合わせによってより細かな抽象化が行える。また詳細化する品詞を選択する方法も異なる。

5 おわりに

本研究では、誤り駆動の素性選択による日本語形態素解析の確率モデル学習を行った。解析誤りをもとに品詞分類を素性として抽出し追加していくことで、パラメータ推定のための有効な品詞分類を自動的に学習した。これにより、人手で品詞分類を調整するわずらわしさを解消することができた。

学習によって得られた品詞分類を使って解析を行った結果、少ないパラメータ数で高い解析精度を得られることが分かった。また、品詞分類の細かさや分類の方法によってパラメータ数と解析精度がどのように変化するかといった、品詞分類の性質や特徴をつかむことができた。

今後の予定として、素性の抽出と追加の方法を改善すること、抽出する素性を様々に変化させて実験を行うこと、variable-gram の確率モデル学習を行うことなどを考えている。

参考文献

[1] Eric 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, December 1995.

[2] M. Haruno and Y. Matsumoto. Mistake-driven mixture of hierarchical tag context trees. In *Proceedings of the 35th Annual Meeting of ACL and the 8th Conference of EACL*, pp. 230–237, 1997.

[3] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』 version 1.0 使用説明書. Information Science Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology, 1997.

[4] 柏岡秀紀, Stephen G. Eubank, Ezra W.Black. 確率付決定木を用いた日本語形態素解析. 言語処理学会第 3 回年次大会論文集, pp. 433–436. 言語処理学会, March 1997.