形態素解析のための拡張統計モデル

浅原正幸† 松本裕治†

自然言語処理の分野で最も基本的な処理として形態素解析がある．近年大量のタグ付きコーパスが整備され，コーパスに基づいた統計的形態素解析器が開発されてきた．しかし単純な統計的手法ではコーパスに出現しない例外的な言語現象に対処することができない．この問題に対処するため，本論文ではより柔軟な拡張統計モデルを提案する．例外的な現象に対応するために単語レベルの統計値を利用する．この拡張により，細かく分類された大量のタグを扱う際，必要なコーパスの量は増加する．一般に適切なコーパスの量で学習する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り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，先行する品詞タグ集合と，後続する品詞タグ集合とで，別々の品詞タグの同値類を導入するようにした．コーパスの量が不足する場合に tri-gram モデルを構築すると，学習データへの過学習が起きる．これを回避するために選択的 tri-gram モデルを導入した．一方，これらの拡張のため，語彙化するタグや tri-gram 文脈の選択を人手で設定することは困難である．そこで，この素性選択に誤り駆動の手法を導入し半自動化した．日本語・中国語形態素解析，英語品詞タグ付けについて評価実験を行い，これらの拡張の有効性を検証した．

Extended Statistical Model for Morphological Analysis

Masayuki Asahara† and Yuji Matsumoto†

Recently, large-scale part-of-speech tagged corpora have become available, making it possible to develop statistical morphological analyzers trained on these corpora. Nevertheless, statistical approaches in isolation cannot cover exceptional language phenomena which do not appear in the corpora. In this paper, we propose three extensions to statistical models in order to cope with such exceptional language phenomena. First of all, we incorporate lexicalized part-of-speech tags into the model by using the word itself as a part-of-speech tag. Second, because the tag set becomes fragmented by the use of lexicalized tags, we reduce the size of the tag set by introducing a new type of grouping technique where the tag set is partitioned creating two different equivalent classes for the events in the conditional probabilities of a Markov Model. Third, to avoid over-fitting, we selectively introduce tri-gram contexts into a bi-gram model. In order to implement these extensions, we introduce error-driven methods to semi-automatically determine the words to be used as lexicalized tags and the tri-gram contexts to be introduced. We investigate how our extension is effective through experiments on Japanese, Chinese and English.

1. はじめに

近年，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，言語の使用の多様性や言語そのものの多様性を考えると，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．この拡張統計モデルは可変長マルコフモデル[10]に基づいた統計的形態素解析器『茶筌』[18]に使用されている．現在『茶筌』は日本語形態素解析のほか，内部開発版では英語品詞タグ付けおよび中国語形態素解析を行うことができる．本論文では形態素解析のための拡張統計モデルの概要について述べる．

『茶筌』では日本語の品詞体系として IPA 品詞体系[16]に少し手を加えたものを採用している．そのタグの数は約 650 にもなる．助詞や助動詞などのいくつかの単語については 1 単語を 1 品詞として見なすため，実際のタグの数はさらに多い．タグの数が多いために，単純な tri-gram 連接規則を構築することは困難である．すべてのタグを個別のものとすると bi-gram 連接規則を構築することすら難しい．

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

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[6]がある．本論文ではこの手法を拡張し，マルコフモデルの条件付き確率計算について，先行するタグ集合と後続するタグ集合とで，別々の同値類を導入するようにした．本手法により，統計モデルに対して，日本語の活用形態や縮約形態の特徴を反映させることができる．日本語の活用語は，前の単語の活用形は後の単語に対して重要であるが，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．先行するタグ集合では活用形を個別に扱い，後続するタグ集合ではすべての活用形を含めた同値類を導入することにより，この活用語の特徴を統計モデルに反映させることが可能である．また，話し言葉に頻出する縮約形態として，2つ以上の形態素が1つの形態素へと縮約するという現象がある．たとえば，助動詞「ちゃう」は「て（助詞）」と「しまう（助動詞）」の2つの単語の縮約形態である．このような単語は，前の単語に対してと，後ろの単語に対してとで，別々の品詞に属するような文脈的振舞いを行う．前からの連接規則の場合と，後ろからの連接規則の場合とで，別々の品詞とグループ化することにより，その文脈的振舞いを統計モデルに反映させ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つねに重要な問題である．特に日本語で採用しているタグ集合では，活用形を展開するとタグの数が650を超え，スムージングを導入したとしても，tri-gram モデルを構築することは非現実的である．しかし解析のために tri-gram の文脈が必要になる場合がある．そこで我々は bi-gram モデルを基にし，必要に応じて tri-gram を利用する選択的 tri-gram モデルを導入する．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．選択した tri-gram 連接について，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する．

またこれらの拡張モデルに対し，有用な素性選択を人手で行うことは非常に困難である．上に述べた各拡張は例外的な言語現象に対応するために導入されることを鑑み，これらの例外的な言語現象を素性として抽出するために，誤り駆動の手法を導入する．

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．

2章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．3章では拡張モデルについて詳述する．4章では誤り駆動による素性選択手法について説明する．5章で様々な条件でモデルの評価を行い，最後に6章でまとめと今後の課題について述べる．

本論文では，単語やタグの相対的位置を表すために，ある単語（もしくはタグ）生起位置 c に対して，1つ前の生起位置を p，2つ前の生起位置を p' で表す．c を後件，p を前件，p' を前々件と呼ぶ．位置 c に単語 w が出現する事象を w^c，位置 c にタグ t が出現する事象を t^c と書く．同様に，位置 p に単語 w が出現する事象を w^p，位置 p にタグ t が出現する事象を t^p，位置 p' に単語 w が出現する事象を $w^{p'}$，位置 p' にタグ t が出現する事象を $t^{p'}$ と書く．

また $\langle w,t\rangle$ は品詞タグが t である単語 w が出現する事象を示す．$F(E)$ は事象 E がコーパス中に生起する頻度，$F(E^p,E^c)$ は事象 E^p と E^c が連続して同時に生起する頻度，$F(E^{p'},E^p,E^c)$ は事象 $E^{p'}$ と E^p と E^c が連続して同時に生起する頻度を示す．

2. 背　　景

2.1 各国語の形態素解析

日本語形態素解析は，入力テキストを単語単位にわかち書きし，品詞タグを付与する処理である．必要に応じて活用語の処理を行う．我々は日本語形態素解析の品詞体系として IPA 品詞体系[16]を少し改良したものを採用している．この品詞体系は階層構造をなしている．品詞情報とは別に活用型や活用形が品詞体系中に定義されており，活用型，活用形まで個別に見た際のタグの数は約650にもなる．

英語はわかち書きをする習慣があるため，単語境界同定をほとんど必要としない．しかし多品詞語が多いため，1単語あたりの品詞の曖昧性が大きい.『茶筌』では英語の品詞体系として PennTreebank[9]の Tagged Corpus で採用している品詞体系を詳細化して利用している．現在英語のトークナイザの実装や「New York」などといった2語で1語と見なすべき固有表現の辞書登録により実用的な品詞タグ付け器を目指している．

中国語形態素解析は，日本語と同様，わかち書きの作業を必要とする．しかし，中国語は活用しないため，活用語の処理を必要としない．中国語は，わかち書きの基準だけでなく品詞体系の基準の揺れが多く，言語学者の間でも品詞定義が揺れており，品詞同定は他の言語に比べるとより困難である．本論文では実験用のコーパスとして Academia Sinica Balanced Corpus[3]を採用した．解析器が出力する単語わかち書きの単位および品詞体系は，このコーパスに基づいている．

2.2 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マルコフモデルが知られている．以下，マルコフモデルによる統計的形態素解析について説明する．

形態素解析は入力文 S の単語列 $W = w^1, \ldots, w^n$ 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品詞タグ列 T を発見することである☆．

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

ベイズの定理を利用して，$P(W,T)$ は品詞タグ列の生起確率と単語列の生起確率として展開することができる．

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

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．

$$P(W|T) = \prod_{i=1}^{n} P_w(w^i|t^i),$$

$$P(T) = \prod_{i=1}^{n} P_t(t^i|t^{i-1}) \ \ (or\ P_t(t^i|t^{i-2}, t^{i-1})).$$

これらの値をタグ付きコーパスの頻度から最尤推定する．最尤推定時には文中の絶対位置ではなく，以下のように相対位置で頻度を数えあげたものを利用する．

$$P_w(w^c|t^c) = \frac{F(\langle w^c, t^c\rangle)}{F(t^c)},$$

$$P_t(t^c|t^p) = \frac{F(t^p, t^c)}{F(t^p)},$$

$$P_t(t^c|t^{p'}, t^p) = \frac{F(t^{p'}, t^p, t^c)}{F(t^{p'}, t^p)}.$$

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．

2.3 統計モデルの問題点

自然言語には様々な例外的な現象が出現し，品詞に基づく統計的手法のみではすべての言語現象を解決することは不可能である．本論文では品詞統計モデルでは解決することが困難である以下のような問題点に着目した．

まず，同じ品詞の他の単語とは異なる振舞いをする単語がある．特に日本語の場合，助動詞，助詞といった付属語は各単語で連接の挙動が異なり，解析が困難であることが知られている．

次に，日本語では活用語などの活用型や活用形を個別に見ると品詞数は約 650 にもなる．これらをすべて個別に見て統計モデルを作成するとコーパス中には出現しない活用形態が出現する．また，品詞数が膨大な場合には，単純に tri-gram モデルを作成することが困難である．しかし，tri-gram の文脈を見ないと解決できない言語現象が存在する．

最後に，近年形態素解析の話し言葉への対応が求められている．話し言葉のコーパスも整備されつつあるが，依然として量が少ないために，新聞記事などの書き言葉のコーパスによるところが多く，良い精度を達成することが難しい．さらに話し言葉特有の問題として，縮約表現に対応できないという問題がある．

3. 形態素解析のための拡張統計モデル

本章では，前章で述べた問題点に対処するために，統計モデルに対する 3 種類の拡張を提案する．

まず，同一品詞中の例外的な文脈的振舞いを行う単語に対し，**単語レベルの統計値**を利用する．次に，日本語の活用語特有の性質や話し言葉に出現する縮約表現に対応するため，**前件，後件などの文脈に応じて別々のグループ化**を行う．最後に，tri-gram 文脈を必要とする現象に対し選択的に連接規則を利用する**選択的** tri-gram **モデル**を提案する．

3.1 単語レベルの統計値

単語の中には同じ品詞に属する他の単語と異なる文脈的振舞いをする単語がある．特に，日本語の助詞，助動詞，一部の動詞，英語の前置詞，中国語の接頭辞，接尾辞などは，単語ごとに異なる文脈的振舞いをすることが知られている．たとえば「する」「できる」といった動詞は，他の動詞と異なり，前件に品詞が「名詞–サ変接続」である単語をとりやすいという文脈的振舞いを持つ．このような単語に対し，単語を別々の品詞タグとして定義し，個別に統計値をとるように拡張した．

3.1.1 提案手法の詳細

以下，具体的手法を示す．元の品詞タグ集合 $\mathcal{T}$ に対し，いくつかの単語について，その語彙化したタグを新たに加える．また各タグについて，前件に対する

☆ 日本語や中国語の場合には，入力が文字列となり，可能な単語列をすべて展開したうえで品詞列同定と単語列同定を同時に行うことになる．

ものと後件に対するものとを区別し，前件のタグ集合を $\mathcal{T}^p$ ，後件のタグ集合を $\mathcal{T}^c$ とする．

後件に，単語 w^c について定義された，語彙化したタグ $\bar{t}^c$ が現れる場合の単語生起確率は次のようになる：

$$\begin{aligned} P_w(w^c|\bar{t}^c) &= P(w^c|\langle w^c, t^c\rangle) \\ &= \frac{F(\langle w^c, t^c\rangle)}{F(\langle w^c, t^c\rangle)} \\ &= 1. \end{aligned}$$

また，タグ連接確率は次のようになる：

$$\begin{aligned} P_t(\bar{t}^c|t^p) &= P(\langle w^c, t^c\rangle|t^p) \\ &= \frac{F(t^p, \langle w^c, t^c\rangle)}{F(t^p)}. \end{aligned}$$

前件に，単語 w^p について定義された，語彙化したタグ $\bar{t}^p$ が現れる場合単語生起確率は変更しない．タグ連接確率は次のようになる：

$$\begin{aligned} P_t(t^c|\bar{t}^p) &= P(t^c|\langle w^p, t^p\rangle) \\ &= \frac{F(\langle w^p, t^p\rangle, t^c)}{F(\langle w^p, t^p\rangle)}. \end{aligned}$$

本手法では，前件と後件で別々の単語について，語彙化したタグとして定義することを許す．このため，前件におけるタグの母集団 $\mathcal{T}^c$ と後件におけるタグの母集団 $\mathcal{T}^p$ は区別される．各タグは単語の集合と見なすと，本手法は，この単語の集合に対し，前件と後件とで別々の再分割を行っていることに等しい．

注意するべき点として，ある単語について，語彙化したタグとして導入した場合に，その単語の元の品詞タグ集合からその導入した単語を除かれることがある．

後件のタグ集合 $\mathcal{T}^c$ 中で，語彙化したタグを，品詞 t_a に属する単語 $w_{a_1}, \ldots, w_{a_n} (\in t_a)$ について導入した場合，品詞タグ $t_a^c \in \mathcal{T}^c$ は次のようになる：

$$t_a^c = t_a \setminus \{w_{a_1}, \ldots, w_{a_n}\}.$$

同様に，前件のタグ集合 $\mathcal{T}^p$ 中で，語彙化したタグを，品詞 t_b に属する単語 $w_{b_1}, \ldots, w_{b_m} (\in t_b)$ について導入した場合，品詞タグ $t_b^p \in \mathcal{T}^p$ は次のようになる：

$$t_b^p = t_b \setminus \{w_{b_1}, \ldots, w_{b_m}\}.$$

そのうえで，品詞タグ連接 $t_b t_a$ の確率を推定する際，頻度 $F(t_b, t_a)$ ではなく $F(t_b^p, t_a^c)$ を利用する．

3.1.2 単語—品詞間スムージング

語彙化したタグをある単語について導入する際に，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しなければならない．別の手法として，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．連接規則に対し，語彙化した品詞タグを導入した際の統計値のスパースネスを緩和するために，連接確率を計算する際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_{lc} を後件におけるスムージング係数 $(0 \leq \lambda_{lc} \leq 1)$，$\lambda_{lp}$ を前件におけるスムージング係数 $(0 \leq \lambda_{lp} \leq 1)$ とする．

単語 w^c を別個に統計をとった際に導入される語彙化したタグを $\bar{t}^c$ とする．後件の語彙化したタグにスムージングを適用する際，タグ生起確率 $P_t^{lt}(\bar{t}^c|t^p)$ は次のようになる：

$$\begin{aligned} P_t^{lt}(\bar{t}^c|t^p) = \lambda_{lc} P(\langle w^c, t^c\rangle|t^p) \\ +(1-\lambda_{lc}) P(t^c|t^p). \end{aligned}$$

同様に，単語 w^p を別個に統計をとった際に導入される語彙化したタグを $\bar{t}^p$ とする．前件の単語についてスムージングを適用する場合，タグ生起確率 $P_t^{lt}(t^c|\bar{t}^p)$ は次のようになる：

$$\begin{aligned} P_t^{lt}(t^c|\bar{t}^p) = \lambda_{lp} P(t^c|\langle w^p, t^p\rangle) \\ +(1-\lambda_{lp}) P(t^c|t^p). \end{aligned}$$

後件に対し語彙化した品詞タグを導入した際には，単語生起確率に対しても，以下のようなスムージングを考えることが可能である（スムージング係数を $\lambda_{lw} (0 \leq \lambda_{lw} \leq 1)$ とする）：

$$\begin{aligned} P_w^{lw}(\bar{t}^c|t^p) = \lambda_{lw} P(\langle w_i, t_i\rangle|\langle w_i, t_i\rangle) \\ +(1-\lambda_{lw}) P(t^c|\langle w^c, t^c\rangle). \end{aligned}$$

しかし，$1 \gg P(t^c|\langle w^c, t^c\rangle)$ のため，単語生起確率が λ_{lw} に敏感になりすぎ，有用な統計モデルを構成することができなかった．本モデルには，この単語生起確率に対するスムージング手法は採用しなかった．

3.1.3 関 連 研 究

Kim ら[7] は英語品詞タグ付けに対し，単語レベルの統計値を利用している．Kim らの手法では，本手法のように，前件と後件とで別々の単語を語彙化した品詞タグとして導入することは行っていない．

3.2 前件文脈と後件文脈とで別々の同値類の導入

非常に細かいタグ集合を導入する際，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．また，いくつかの品詞（もしくは単語）は，現れる位置によって別々の文脈的振舞いをする．この問題に対処するために，同値類を導入する際に，前件文脈と後件文脈とで別々のグループ化を定義する．

たとえば，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 連接もしくは tri-gram 連接の前件 t^p の位置に現れるもののみ

を考慮に入れればよい．これは活用語の統計値をとる際，その出現位置によって別々のグループ化を導入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．たとえば，助動詞「ちゃう」は「て（助詞）」と「しまう（助動詞）」の 2 つの単語から構成される縮約表現であり，他の助動詞の単語とは異なる振舞いをする．これらの振舞いを統計的に学習する方法として，その単語の様々な使用例を集め，正確にタグ付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，各件で別々の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方法を提案する．単語「ちゃう」についてタグ連接確率 $P(t^c|t^p)$ を計算する際，後件 t^c については「て」と同じ同値類にグループ化し，前件 t^p については「しまう」と同じ同値類にグループ化することにより，コーパス中に低頻度の縮約形態についてもその文脈的振舞いを学習することが可能になる．

3.2.1 提案手法の詳細

以下，各件で別々のグループ化について説明する．簡単のため bi-gram モデルについて説明するが，tri-gram モデルについても同様な拡張を行うことができる．$\mathcal{T}^c$ を後件の品詞タグ集合，$\mathcal{T}^p$ を前件の品詞タグ集合とする．これらのタグ集合に対し，各件で別々の同値類集合を導入する．たとえば，後件に対する同値類集合を $\mathcal{G}^c = \{G^c_1 = \{t^c_1, t^c_2\}, G^c_2 = \{t^c_3\}\}$ とし，前件に対する同値類集合を $\mathcal{G}^p = \{G^p_1 = \{t^p_1\}, G^p_2 = \{t^p_2, t^p_3\}\}$ とする．ここで，後件に対する同値類集合 $\mathcal{G}^c$ を定義する際の注意点として，1 つの単語が複数の品詞になりうる場合，導入された各同値類 $G^c_1, G^c_2, \cdots$ が，そのなりうる複数の品詞を 2 つ以上含まないようにする必要がある．そうでない場合，その当該単語についての品詞同定に対し，何ら寄与しない統計モデルになってしまう．

これらのタグ集合間の写像として，後件の同値類を生成する写像 $I^c(\mathcal{T}^c \to \mathcal{G}^c)$ および前件の同値類を生成する写像 $I^p(\mathcal{T}^p \to \mathcal{G}^p)$ を定義する．同値類のクラスを表現するために，後件に出現するタグ t^c が，I^c により写像される先のタグを $[t^c] \in \mathcal{G}^c$，前件に出現するタグ t^p が，I^p により写像される先のタグを $[t^p] \in \mathcal{G}^p$ と書くと，単語生起確率，タグ連接確率は次のようになる：

$$
\begin{aligned}
P_w(w^c|t^c) &= P(\langle w^c, [t^c]\rangle|[t^c]) \\
&= \frac{F(\langle w^c, [t^c]\rangle)}{F([t^c])} \\
&= \frac{F(\langle w^c, t^c\rangle)}{F([t^c])},
\end{aligned}
$$

$$
\begin{aligned}
P_t(t^c|t^p) &= P([t^c]|[t^p]) \\
&= \frac{F([t^p], [t^c])}{F([t^p])}.
\end{aligned}
$$

本手法を利用することにより，表記の揺れ（漢字/かな）もグループ化により吸収することが可能となる．

3.2.2 関 連 研 究

Cutting ら[4] は，可能なタグの集合が同じ単語を同値類と見なしグループ化した．これにより，学習時に再推定されるべき単語生起確率のパラメータの数を減らすことができる．Schmid[11] は，さらに同値類と各単語の間，同値類と各品詞の間でスムージングを導入した．これらの手法はパラメータを減らすための処置で，可能なタグの集合が同じ単語は同じ振舞いをするという仮定を基にしている．これに対し本手法では，言語知識を利用した任意のグループ化の設定を可能にした．さらに条件付き確率の前件と後件とで別々のグループ化を設定することが可能である．また，従来の規則に基づく日本語形態素解析では，前方接続表現，後方接続表現として品詞の分類を越えた素性を与えるという方法が主として使われてきた．本手法は，この考え方を統計モデルに導入したものと考えることができるが，それを一般化することにより，tri-gram モデルにも自然に拡張できるようになっている．

3.3 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現実的には不可能である．しかし，品詞決定に tri-gram の文脈を必要とする場合がある．たとえば，単語「ない」は形容詞か助動詞かで品詞の曖昧性がある．係助詞「は」が先行する場合に後続する「ない」は通常形容詞である．例外として助動詞「だ」の連用形や形容詞の連用テ接続が「は」の前に先行する場合には,「ない」は助動詞になる．このような現象は bi-gram 統計だけでは解析できない．

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．本モデルでは選択的に導入される tri-gram 統計と bi-gram 統計とを混合して利用する．

3.3.1 提案手法の詳細

以下，選択的 tri-gram について詳述する．選択的 tri-gram は tri-gram と bi-gram とを混合させたモデルである．ある bi-gram 文脈が tri-gram 文脈と交わりを持つ場合，tri-gram 文脈は bi-gram 文脈中の例外規則と見なす．すべての文脈は本モデル中で互いに共通の要素を持たないように構成される．bi-gram 文脈が tri-gram 文脈と重なりを持つ場合には，bi-gram 文脈はその tri-gram 文脈を除いて推定される．

(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)

図 1 選択的 tri-gram の例
Fig. 1 Examples of the selective tri-gram.

ある tri-gram 文脈 $t^{p'}t^pt^c$ を本モデルに含める際，次に示すタグ連接確率を利用する：

$$P_t(t^c|t^{p'},t^p)=\frac{F(t^{p'},t^p,t^c)}{F(t^{p'},t^p)}.$$

この際，文脈に重なりがないようにするために，bi-gram 文脈 t^pt^c のタグ連接確率は次のように計算する（以下の式で F はコーパス中の真の頻度を意味し，F' は確率計算に用いられる見なし頻度を意味する）：

$$F'(t^p,t^c)=F(t^p,t^c)-F(t^{p'},t^p,t^c),$$
$$F'(t^p)=F(t^p)-F(t^{p'},t^p),$$
$$P_t(t^c|t^p)=\frac{F'(t^p,t^c)}{F'(t^p)}.$$

図 1 に選択的 tri-gram の例を示す．この例は，先に示した係助詞「は」に後置する「ない」の例である．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．ここで「＊」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」のすべての活用形，すべての表記を同一視したものである．

選択的に導入された tri-gram 文脈についても，先に述べた単語レベルの統計値と各件で別々のグループ化を導入する．また選択された tri-gram 文脈に対して，bi-gram 文脈とのスムージングの手法も適用する．スムージング係数を $\lambda_{tri}(0\leq\lambda_{tri}\leq 1)$ とするとスムージングを適用したタグ連接確率 $P_t^t(t^c|t^{p'},t^p)$ は次のようになる：

$$P_t^{tri}(t^c|t^{p'},t^p)=(1-\lambda_{tri})P_t(t^c|t^p)+\lambda_{tri}P_t(t^c|t^{p'},t^p).$$

3.3.2 関連研究

関連研究として，Ron ら[10] の Variable-gram モデルがある．Schütze ら[12] はこのモデルを実際に英語品詞タグ付けに採用した．Ron らの手法は，n を変化させた n-gram の混合モデルで，可変長の文脈をマルコフモデルに混在させている．文脈の集合は有限状態集合として定義されるが，このようなモデルでは，有限状態を決定的に明確にするために，長さの異なる文脈の集合を相互に分割する必要がある．これに対し，本手法では少し異なった改良を行った．tri-gram 連接をあくまで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，モデルの中ですべての文脈は相互に独立したものとして考えることができ，そのまま Ron ら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．

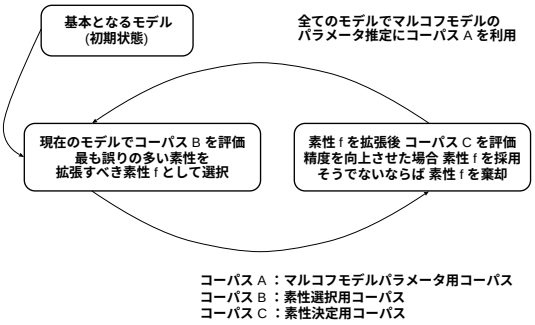

図 2 誤り駆動による素性選択
Fig. 2 Feature selection by the error driven method.

4. 誤り駆動による素性選択

単語レベルの統計量を利用するためには，どの単語を語彙化した品詞として定義するかを決定する必要がある．また選択的 tri-gram は，どの tri-gram 文脈を選択するか決定する必要がある．しかしこれらの決定を人手で行うのは非常に困難である．特に母国語でない言語に対しては，言語知識を利用して有用な素性を選択することはより困難である．この決定を自動化するために誤り駆動による手法を導入する．

4.1 誤り駆動による素性選択手法

誤り駆動による素性選択の方法を**図 2** に示す．本手法では，3 等分したコーパス（コーパス A，コーパス B，コーパス C）を利用する．コーパス A は，マルコフモデルのパラメータ推定に用いる．まず初期モデルを拡張なしに構成し，コーパス B を評価する．コーパス B の解析結果から，誤りの多い単語もしくは tri-gram 文脈を素性として選択する．最後に，選択された素性を拡張したモデル（マルコフモデルのパラメータ推定には最初のコーパスを用いる）を用いて，コーパス C を評価する．コーパス C の精度を改善する場合，その新しい素性を統計モデルに導入する．改善しない場合には，その素性は採用しない．この手順を繰り返すことにより，統計モデルを漸進的に改良していく．

この方法は，多くの誤りの要因になっている素性は詳細化すべきであるという仮定に基づいている．もし

単語が多くの誤りを生成する場合，その単語の拡張は精度を改善できると推測される．また，多くの誤りの要因になっている tri-gram 文脈は，その tri-gram 文脈を追加することによりモデルを改善できると推測される．

4.2 関連研究

誤り駆動による手法として Brill[2] の変形規則による英語品詞タグ付けがある．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り，精度を向上させている．

北内ら[15] は日本語形態素解析に対し誤り駆動により品詞の詳細化を行うことにより精度を向上させている．これは品詞階層構造方向のグループ化と見なすことができる．本手法は，この手法をベースとしているが，彼らの手法では，bi-gram のモデルに限定されており，tri-gram 以上の文脈を含むモデルにまで扱うことをしていない．

単語レベルの統計値のための素性選択手法として，Kim ら[7] の品詞出現分布の異なる順に追加する手法がある．

Haruno ら[5] は，コーパスから文脈木を作り，短い文脈と長い文脈における品詞出現分布の差を見ることによって，長い文脈を選択するかどうかを判断する方法を用いている．ただし，この方法では文脈長を伸ばすことが解析の精度向上に直接つながるかどうかは自明ではない．本論文で用いる方法は，第 3 のコーパスで精度向上を確認するというものであり，精度向上により敏感な手法といえる．

5. 評価

提示してきた拡張が通常の bi-gram モデルをどのように改善することができるかを評価するためにいくつかの実験を行った．

5.1 日本語形態素解析

日本語の素性選択は，解析誤りの情報とそれに対する言語知識を基に人手で行った．まず，評価実験手順と結果を提示し，次に実験結果についての考察を示す．

5.1.1 実験手順

単語レベルの統計値は助詞，助動詞と一部の動詞などに導入した．単語レベルの統計値を導入する際には，各単語に属する品詞とのスムージングを導入した．単語—品詞間のスムージング係数 λ_{lc}，λ_{lp} は，すべての単語について 0.9 に固定した．各件で別々のグループ化については，前件について活用形を個別に統計をとり，後件について活用形の違いを無視するようなグループ化を導入した．また，解析誤りの多いいくつかの縮約表現についても，元の構成語とのグループ化を導入した．選択的 tri-gram については，特に解析誤りの多い 30 個程度の tri-gram 文脈規則を入れて実験を行った．導入した tri-gram 文脈規則についてスムージングを導入し，bi-gram–tri-gram 間スムージング係数 λ_{tri} は 0.9 に設定した．

$$\text{再現率} = \frac{\text{一致した形態素数}}{\text{コーパスの形態素数}}$$

$$\text{適合率} = \frac{\text{一致した形態素数}}{\text{システムが出力した形態素数}}$$

$$\text{F 値} = \frac{(\beta+1)\cdot\text{再現率}\cdot\text{適合率}}{\beta^2\cdot(\text{再現率}+\text{適合率})}$$

図 3 F 値
Fig. 3 F-Value.

表 1 日本語形態素解析の評価実験結果（F 値 %）
Table 1 Results of Japanese morphological analysis (F-Value).

モデル	レベル 1	レベル 2	レベル 3
単純 bi-gram	99.006	98.440	97.356
拡張統計モデル	99.128	98.704	97.812

評価は 5-fold cross evaluation による．タグ付きコーパスを学習データ（80%）と評価データ（20%）に分割し，評価実験を 5 回繰り返し，結果を平均した．全データサイズは 37,490 文 922,932 単語である．

評価は次の 3 つのレベルで行った．

- レベル 1：単語境界のみ一致
- レベル 2：単語境界と品詞のトップレベルが一致
- レベル 3：品詞の全情報が一致

モデルを評価するために，F 値（**図 3**）を利用した．F 値を求めるに際し β の値を 1 とした．

5.1.2 考察

各レベルについて評価した結果を**表 1** に示す．

今回の実験により，より柔軟に言語知識を統計モデルに反映させることが可能となり，各レベルで精度を向上させることが可能となった．単語レベルの統計値の利用により，機能語などの頻度が多く同一品詞内で振舞いが異なる単語について，適切な拡張を行えるようになった．特に，助詞や助動詞はひらがな表記が多く，ひらがな表記の普通名詞との曖昧性による解析誤りが多かったが，これについても解消することができた．

グループ化により適切な大きさのコーパスから活用語や縮約表現に対応できた．「ちゃう（て+しまう）」「でる（で+いる）」などや，元コーパスにほとんど出てこない四段動詞の連接などが，グループ化により解

析できるようになった．

特に現在採用している品詞タグ集合が大きいために，通常の tri-gram モデルを作成することは非現実的であったが，選択的 tri-gram により必要な tri-gram 文脈の情報を利用できるようになった．単純 bi-gram モデルで解析できなかった「た/こと/で」「しよ/う/と」「こと/は/ない」といった，ひらがな表記の機能語の連接について，精度の向上が見られた．

5.2 英語品詞タグ付け・中国語形態素解析

英語品詞タグ付けと中国語形態素解析には，単語レベルの統計値と選択的 tri-gram の 2 種類の拡張について評価実験を行った．各拡張の素性選択は誤り駆動による手法で自動化し，いっさいの言語知識を利用しなかった．英語の評価実験には Penn Treebank（52,725 文）[9] の Tagged Corpus を利用した．中国語の評価実験には Academia Sinica Balanced Corpus（284,888 文）[3] を利用した．

最初に共通する実験手順を示し，次に単語レベルの統計値と選択的 tri-gram 個別の実験について述べ，最後に実験結果に対する考察を示す．

5.2.1 実 験 手 順

以下に単語レベルの統計値と選択的 tri-gram の各素性選択に共通する評価実験手順を示す．まず最初に，コーパスを同じ大きさの 5 つのコーパス（A，B，C，D，E）に分割する．コーパス A，B，C を素性選択に用い，コーパス D，E を評価に用いる．ここで素性とは，語彙化した品詞として定義する単語や，選択的に導入する tri-gram 連接を意味する．

- 素性選択用データ
  - A：マルコフモデルパラメータ推定用コーパス
  - B：素性選択用コーパス
  - C：素性決定用コーパス
- 評価用データ
  - D：マルコフモデルパラメータ推定用コーパス
  - E：評価用コーパス

以下の手順を繰り返す．

(1) 初期化
まず，通常の bi-gram モデルをコーパス A を用いて作成する．その後このモデルを用いてコーパス B を評価する．

(2) 素性選択
最も多くのエラーの原因となっていると考えられる素性をコーパス B の解析結果から選択する．この選択は解析誤りの情報から自動的に決定される．

(3) 素性決定
選択された素性をモデルに一時的に追加する．この素性選択に基づき，マルコフモデルのパラメータをコーパス A から推定する．その後コーパス C を新しいモデルで評価する．もしコーパス C が改善された場合，その素性を採用する．改善されなかった場合，その素性を破棄する．破棄された素性は再び選択されることはない．

(4) 評価
決定された素性を基にして，マルコフモデルのパラメータをコーパス D から推定する．このモデルを用いてコーパス E を評価する．

コーパス D，E はより一般的な評価をするために用いる．以下に示す評価はコーパス E によるものである．

5.2.2 単語レベルの統計値

誤り駆動による語彙化する単語の選択を評価するために 3 種類の実験を行った．

- 前件についてのみ単語を拡張する実験
- 後件についてのみ単語を拡張する実験
- 前件と後件を同時に単語を拡張する実験

本実験では単語—品詞間スムージングのスムージング係数は λ_{lc}，λ_{lp} は 0.9 に固定した．

図 4 に英語コーパスによる実験結果，**図 5** に中国語コーパスによる実験結果を示す．

5.2.3 選択的 tri-gram

まず，個別の tri-gram 文脈の単位で追加する実験を行った．しかしこの単位では，精度の変化が小さく，有用な素性選択をすることができなかった．そこで，我々は 2 つの単位を定義した．1 つは前件と後件を共有する tri-gram 連接の集合 $\{t_1^{p'}t^pt^c, t_2^{p'}t^pt^c, \ldots, t_n^{p'}t^pt^c\}$（単位 $P-C$ と呼ぶ）であり，もう 1 つは前々件と前件を共有する tri-gram 連接の集合 $\{t^{p'}t^pt_1^c, t^{p'}t^pt_2^c, \ldots, t^{p'}t^pt_n^c\}$（単位 $P'-P$

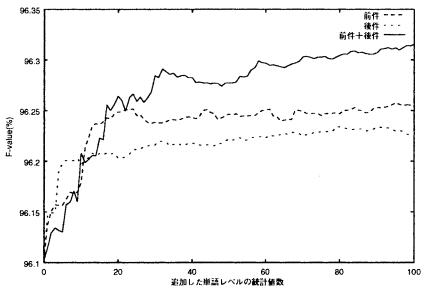

図 4 実験結果：単語レベルの統計値（英語）
Fig. 4 Results: Lexicalized POS tags (English).

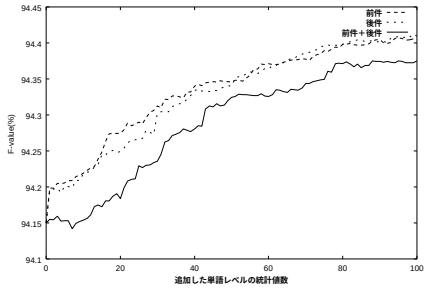

図 5 実験結果：単語レベルの統計値（中国語）
Fig. 5 Results: Lexicalized POS tags (Chinese).

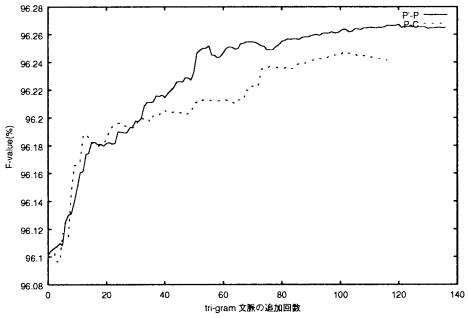

図 6 実験結果：選択的 tri-gram（英語）
Fig. 6 Results: Selective tri-gram (English).

と呼ぶ）である．

これらの文脈はエラーを生成する文脈のみを追加した．エラーを生成しない場合その tri-gram 文脈は追加されない．このためエラーを生成しない文脈は bi-gram 文脈のルールとして定義される．

選択的 tri-gram の評価について，単語レベルの統計値の利用との関係を調べるために複数の評価を行った．各言語に対し，単語レベルの統計値を用いないモデルと，単語 50 個について単語レベルの統計値を導入したモデルについて評価実験を行なった．この単語の選択には，前節の実験で得られたものを利用した．また，bi-gram–tri-gram 間スムージング率 λ_{tri} は 0.9 に固定した．

図 6，**図 8** に英語コーパスの実験で得られた結果，**図 7**，**図 9** に中国語コーパスの実験で得られた結果を示す．

5.2.4 考　　察

表 2 に英語コーパスの各拡張の精度と連接規則数を示す．単語レベルの統計値の導入は精度改善に有効であることが分かる．品詞を語彙化しかつ全 tri-gram を利用するモデルの場合，学習データへの過学習により精度が落ちる．これに対し選択的 tri-gram は，品詞の語彙化の精度を維持したまま必要な tri-gram 文脈を追加することができた．

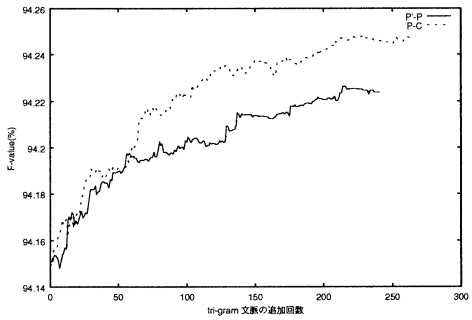

図 7 実験結果：選択的 tri-gram（中国語）
Fig. 7 Results: Selective tri-gram (Chinese).

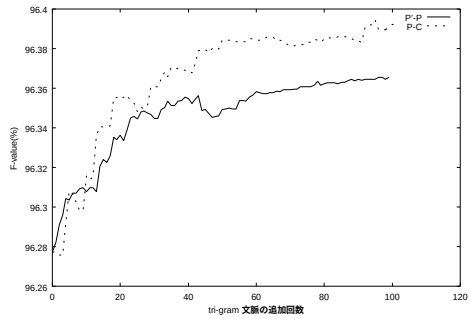

図 8 実験結果：単語レベルの統計値（50 語）を導入した際の選択的 tri-gram（英語）
Fig. 8 Results: Selective tri-gram with 50 word-level statistics (English).

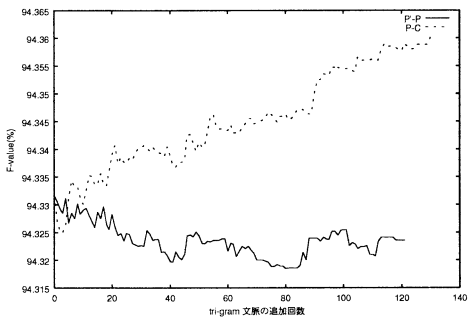

図 9 実験結果：単語レベルの統計値（50 語）を導入した際の選択的 tri-gram（中国語）
Fig. 9 Results: Selective tri-gram with 50 word-level statistics (Chinese).

図 4 を見ると，前件と後件に対し 30 語程度選択するまで精度が向上していることが分かる．前件では次の単語が固有名詞が来るというマーカとして機能している “the（冠詞）” や他の動詞と異なる連接振舞いをする be 動詞が語彙化された．後件では “in（不変化詞）”，“much（形容詞）” などといった機能語が語彙化された．また，tri-gram 連接の手法で単位 $P' - P$

表 2 各拡張と連接規則数（英語）
Table 2 The number of connection rules (English).

モデル	語彙化	F 値 (%)	全連接規則数
bi-gram	なし	96.38	1,309
全 tri-gram	なし	96.60	12,859
tri-gram(P'-P)	なし	96.55	2,023
tri-gram(P-C)	なし	96.53	2,006
bi-gram	50 語	96.56	2,189
全 tri-gram	50 語	95.65	16,985
tri-gram(P'-P)	50 語	96.65	2,488
tri-gram(P-C)	50 語	96.68	2,753

表 3 各拡張と連接規則数（中国語）
Table 3 The number of connection rules (Chinese).

モデル	語彙化	F 値 (%)	全連接規則数
bi-gram	なし	94.15	2,213
全 tri-gram	なし	94.14	27,494
tri-gram(P'-P)	なし	94.22	4,033
tri-gram(P-C)	なし	94.24	4,159
bi-gram	50 語	94.33	3,581
全 tri-gram	50 語	93.28	35,351
tri-gram(P'-P)	50 語	94.32	4,068
tri-gram(P-C)	50 語	94.36	4,331

と単位 $P-C$ の両方についてほぼ同等の精度向上が見られた．英語の選択的 tri-gram については，追加単位の構成にかかわらず精度の向上が見られた（図 6，図 8）．単語レベルの統計値を用いたうえで選択的 tri-gram を導入した場合，通常の tri-gram モデルと比べて，連接規則数を 5 分の 1 以下に削減することができる．これにより解析速度も改善させることができた．通常の tri-gram モデルでは，10,545 文（267,619 単語）の解析に 15.04 秒かかっていたが，拡張モデルでは 14.11 秒で解析することができた．

中国語コーパスについては以下のことがいえる．図 9 の結果から，語彙化した品詞を導入した場合に，単位 $P'-P$ の単位で tri-gram 連接規則を追加する場合にはあまり精度が伸びなかった．さらに，**表 3** から，中国語の場合には，tri-gram の規則自体があまり有効でないことが分かる．文脈長を伸ばすよりも語彙化した品詞を導入する方が精度が伸びていることから，品詞体系自体の詳細化が精度の向上に寄与することが予測される．実際，利用したコーパスの品詞体系は副詞について細分類化され，各副詞に対応する動詞が細分類化されている一方で，接頭辞，接尾辞といったものが細かく定義されていない．これらの機能語の細分類を行うことにより精度の向上を見込むことができると考えられる．

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ることにより例外的な振舞いをする単語にも対応できるようになった．また，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．条件付き確率の各件ごと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．また，これらの統計モデルの拡張のための素性選択に，誤り駆動の手法を導入し半自動化することができた．

今後の課題として未知語処理があげられる．欧米語のようにわかち書きのされる言語では，接頭辞，接尾辞からの品詞生起確率を用いて未知語の品詞推定を行うこと[11]が可能である．また，Brants[1]はこの接頭辞，接尾辞の語長間でスムージングを行い，未知語の品詞推定の精度を向上させている．

しかし，日本語，中国語のようなわかち書きしない言語の場合，未知語境界を決定することすら難しく，この手法を導入することは困難である．『茶筌』では字種により未知語境界を制限する手法をとっているだけで，今後，未知語に対する対応が必要になってくると考えている．

形態素解析器『茶筌』は以下の URI から入手できる．
`http://chasen.aist-nara.ac.jp/chasen/`

謝辞 本研究の一部は，平成 13 年度科学研究費補助金（特別研究員奨励費）の援助を受けている．ここに記して謝意を表す．また，示唆に富むご指摘をいただきました査読者の方々に記して謝意を表す．

参 考 文 献

1) Brants, T.: TnT — A Statistical Part-of-Speech Tagger, *Proc. 6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics ANLP-NAACL 2000*, and *Proc. ANLP-NAACL 2000 Student Research Workshop*, pp.224–231 (2000).
2) Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, *Compu-*

tational Linguistics, Vol.21, No.4, pp.543–565 (1995).
3) Chen, K., Huang, C., Chang, L. and Hsu, H.: SINICA CORPUS: Design Methodology for Balanced Corpora, *PACLIC 11: Language, Information and Computation Selected Papers from the 11th Pacific Asia Conference on Language, Information and Computation*, Seoul, pp.167–176 (1996).
4) Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P.: A Practical Part-of-Speech Tagger, *Proc. 3rd Conference on Applied Natural Language Processing* (1992).
5) Haruno, M. and Matsumoto, Y.: Mistake-Driven Mixture of Hierarchical Tag Context Trees, *35th Annual Meeting of the Association for Compuational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp.230–237 (1997).
6) Jelinek, F.: *Statistical Methods For Speech Recognition*, The Mit Press (1998).
7) Kim, J.D., Lee, S. and Rim, H.: HMM Specialization with Selective Lexicalization, *the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.121–127 (1999).
8) Manning, C.D. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (1999).
9) Marcus, M., Santorini, B. and Marcinkiweicz, M.: Building a large annotated corpus of English: PennTreebank, *Computational Linguistics*, Vol.19, No.2, pp.313–330 (1993).
10) Ron, D., Singer, Y. and Tishby, N.: Learning Probabilistic Automata with Variable Memory Length, *COLT-94*, pp.35–46 (1994).
11) Schmid, H.: Improvements In Part-of-Speech Tagging With an Application To German, *EACL SIGDAT Workshop*, pp.47–50 (1995).
12) Schütze, H. and Singer, Y.: Part of Speech tagging using a variable memory Markov model, *Proc. Association for Computational Linguistics* (1994).
13) 北 研二，中村 哲，永田昌明：音声言語処理，森北出版 (1996).
14) 北 研二：確率的言語モデル，東京大学出版会 (1999).
15) 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 (1999).
16) データベースワークショップテキストグループ：テキストデータベース報告書，技術研究組合新情報処理開発機構 (1995).
17) 長尾 真（編）：岩波講座ソフトウェア科学 15 自然言語処理，岩波書店 (1996).
18) 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 (1999).

(平成 12 年 11 月 24 日受付)
(平成 13 年 12 月 18 日採録)

浅原 正幸（学生会員）

1998 年京都大学総合人間学部基礎科学科卒業．同年，奈良先端科学技術大学院大学情報科学研究科博士前期課程入学．2001 年同大学博士後期課程進学．同年より日本学術振興会特別研究員，現在に至る．自然言語処理の研究に従事．言語処理学会学生会員．

松本 裕治（正会員）

1955 年生．1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984〜85 年英国インペリアルカレッジ客員研究員．1985〜87 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．京都大学工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．