# 排反な規則を用いた文節まとめあげ

村田真樹[†] 内元清貴[†]
馬青[†] 井佐原均[†]

本論文では文節のまとめあげを研究の対象とする．文節まとめあげの処理は，係り受け文法に基づく日本語文の構文解析の際に必要となるものである．本研究ではこの文節まとめあげを対象として，既存の 4 手法（決定木学習，最大エントロピー法，用例ベースの手法，決定リストの手法）と，類似度の大きい排反な規則を用いる新手法の比較実験を行った．その結果，今回の問題設定では類似度の大きい排反な規則を用いる提案手法が最も良いことが分かった．排反な規則とは学習データにおいて分類先が排他的になっており，学習データにおいては 100%正しく推定できる規則になっていることを意味する．従来の機械学習のほとんどは排反でない規則，つまり，学習データでの精度が 100%でない規則を用いて解いているものであったが，これはあらかじめ間違う可能性がある規則を用いて解いているということを意味している．今後より高い精度の解析を目指すならば提案手法のような排反な規則を利用して解析するという考え方をとる必要がある．しかし，単純に排反な規則を用いるだけでは偶然その学習データでは排反な規則になったという無意味な規則まで用いてしまうことになる．本論文で提案する類似度と排反な規則を用いる提案手法は，排反な規則の中でも入力との類似度の高いものだけを用いることで，無意味な排反な規則を用いてしまう弊害を解決している．つまり，提案手法は類似度の高い信頼できる排反な規則だけを用いることとなり，類似度の低い信頼性の低い（無意味な）排反な規則を捨てていることとなるのである．提案手法は，類似度を適切に定義できる問題ではきわめて有力な手法となっている．

# Bunsetsu Identification Using Category-exclusive Rules

Masaki Murata,[†] Kiyotaka Uchimoto,[†] Qing Ma[†]
and Hitoshi Isahara[†]

This thesis describes bunsetsu identification using supervised learning. Since Japanese syntactic analysis is usually done after bunsetsu identification, bunsetsu identification is important for analyzing Japanese sentences. In this thesis, we carried out experiments on bunsetsu identification to compare the four existing machine-learning methods (decision tree, maximum-entropy method, example-based approach and decision list) and our new method which uses category-exclusive rules with the highest similarity. In these experiments, our new method using category-exclusive rules with the highest similarity was better than the other learning methods. A category-exclusive rule is defined as a rule where all the data satisfying the category-exclusive rule belong to an exclusive category. The identification by such rules is done at an accuracy rate of 100% in a learning set. Traditional machine learning algorithms use rules that are not category-exclusive. These algorithms use rules that allow for the possibility to make identification errors. In order to improve identification results, it is necessary to use category-exclusive rules, but, if we use category-exclusive rules only, we may incorrectly apply misleading rules that happen to be category-exclusive in a learning set only. To solve this problem we use only category-exclusive rules having the highest similarity. This method is very accurate in cases where we can define the similarity appropriately.

## 1. はじめに

本論文では文節のまとめあげを研究の対象とする．文節のまとめあげとは，たとえば，「今日 (名詞) は (助詞) 天気 (名詞) が (助詞) いい (形容詞) です (助動詞) . (特殊)」というように形態素解析された文を「今日は｜天気が｜いいです.」のように文節にまとめあげることをいう．文節まとめあげの処理は，係り受け文法に基づく日本語文の構文解析[1)]の際に必要となるものである．

文節まとめあげの処理は，文節が自立語＋付属語☆によって構成されるため，自立語や付属語を認定する

† 郵政省通信総合研究所関西先端研究センター
Kansai Advanced Research Center, Communications Research Laboratory Ministry of Posts and Telecommunication

規則を作成しておけば簡単に行うことができると思われがちである．しかし，自立語が連接している場合では区切るかまとめるかは曖昧であるので，自立語，付属語が認定できただけでは文節まとめあげを行うことはできない．

たとえば，以下の 2 例を考える．

（例文 1） 今日　京大　に　行った．
名詞　名詞　助詞　動詞

（例文 2） 京都　大学　に　行った．
名詞　名詞　助詞　動詞

この例では同じ名詞連続でも，例文 1 では「今日」「京大」では文節は区切られ，例文 2 では「京都」「大学」では文節はまとめられることになる．このように自立語，付属語が認定できただけでは文節を区切るかいなかは判定できない場合がある．

ところで，高精度に構文解析を行う knp[2)]では，上記のような現象は,「今日」が時間を表す名詞であることを認識して正しく文節の認定を行っている．また，時間を表すものは「今日」のような 1 つの単語から構成されるものだけではなく,「年度」「世紀」などの接尾辞をともなったものがあり，knp ではこれらの接尾辞が時間を表すものになることを記載した規則を作成することでこれらの場合も正しく解析できるようになっている☆☆．このような現象は時間を表す名詞だけに限ったことではなく,「左」などの相対名詞や「以外」などの副詞の働きをしうる語についても規則化されている．

また，名詞連続に限らず，下記の例のように動詞連続でも曖昧な場合が存在する．

（例文 3） 歌い　踊った．
動詞　動詞

（例文 4） 書き　損じる．
動詞　動詞

例文 3 の「歌い」「踊った」では文節は区切られ，例文 4 の「書き」「損じる」では文節はまとめられることになる．knp では,「損じる」などの補助動詞となりうる動詞に関する規則を作成することでこれらを正しく解析できるようにしている．

knp ではこのような文節にかかわる規則を上記の他,「…しなければならない」は 1 つの文節になるといった例外的な規則を含めて，906 行のファイルに 148 個の規則を記述しており，文節のまとめあげという簡単そうな問題であっても規則が多数になることが分かる．上記の規則は京大コーパスの作成過程において修正・追加を繰り返して作成されたものであるが，人手で作成している以上コストは大きいものだと思われるし，今後も精度向上のためには人手のコストをはらって修正・追加を繰り返す必要がある．また，人手で作成した規則による方法では，異なる分野のデータを解析するときに，それにあうように規則を作成するコストが多大なものとなる．

本研究では，人手コストの軽減を目的として行われている種々の機械学習を用いた研究と同様，上記の人手によるコストを軽減できないかと考え，文節まとめあげを機械学習を用いて行う実験を行った．ただし，機械学習の手法としてはいくつか存在するが，どの方法が良いかははっきりしていないので，なるべく多くの方法を用いて実験を行おうと考え，4 つの既存の教師あり機械学習の手法（決定木学習，最大エントロピー法，用例ベース，決定リスト）で実験を行った．また，機械学習の手法として上記の 4 手法のほかに，類似度の大きい排反な規則を用いる手法といったものを 2 種類提案し，この方法でも文節まとめあげの実験を行った．

## 2. 本論文での文節まとめあげの問題設定

本研究で文節まとめあげの実験を行う際には京大コーパスを用いた．この理由は，文節の定義は曖昧な面があって難しいが，構文解析システム knp を作成している研究グループと同じ研究グループが作成したコーパスならばある程度文節の定義が信頼できると考えたためである．本研究では，学習にも評価にも京大コーパスを用い，そこでの文節の区切りかいなかの情報を用いるので，我々自身が文節の定義に詳しい必要性はなくなっている．

本論文での解析では，形態素解析まではなされたものとし，形態素解析の結果から形態素の情報を取り出し，それを用いて文節の認定（文節まとめあげ）を行う（つまり，京大コーパスに記述してある正しい形態素解析までの情報を用いて解析する）．本論文では文節まとめあげの問題を以下のように形態素間に文節を区切るための記号 “|” を挿入する問題として扱う．

（例）文節を｜まとめあげる

☆ 自立語とは名詞や動詞のようにそれ単独でも文節を構成できるもの．付属語とは単独では文節を構成できず自立語に後接することで文節を構成するもの．

☆☆ knp[2)]ではさらに「三時半」などに用いられるような「半」という接尾辞まで規則化して高精度な解析を実現している．

| | 文 | を | 区切る | . |
|---|---|---|---|---|
| 品詞 | 名詞 | 助詞 | 動詞 | 特殊 |
| 品詞細分類 | 普通名詞 | 格助詞 | 基本形 | 句点 |
| 意味情報 | × | none | 217 | × |
| 単語自体 | × | を | 区切る | × |

**図 1** 解析に用いる情報
Fig. 1 Information used in bunsetsu identification.

処理の手順は，文の頭から各形態素の境界に対してまわりの形態素の情報によって文節を区切るための記号 “|” を挿入するかいなかを判定していくということになる（前から後ろに向けて処理を進めていく場合，すでに挿入した区切り記号自体が 1 つの情報であるが，本研究では簡単のため，区切りかいなかの判定には，すでに挿入した区切り記号の情報は用いず，まわりの形態素の情報だけを用いる）．

文節を区切るかいなかの判定に用いる情報は，文節を区切るための記号 “|” を挿入するかいなかを判定する形態素の境界の前 2 形態素と後ろ 2 形態素の形態素情報である．形態素情報として用いるものは，各形態素ごとに下記の 4 つのものである．

(1) 品詞☆
(2) 品詞細分類（or 活用形）☆☆
(3) 意味情報（分類語彙表[4]の分類番号上位 3 桁）
(4) 単語自体（語尾変化する単語の場合は基本形を用いる）

ただし，解析に用いる情報が増え過ぎると解析に時間がかかり実験の不都合が生じるため，本論文では両端 2 形態素については，意味情報，単語自体の情報は用いない．

**図 1** は,「文を区切る.」という文の「を」と「区切る」という形態素の境界に文節区切り記号を挿入するかいなかを判定する際に用いる情報を示している.「文」「を」「区切る」「.」に対してそれぞれ品詞の情報として「名詞」「助詞」「動詞」「特殊」がある．また，品詞細分類として同様に「普通名詞」「格助詞」「基本形」「句点」の情報がある．また，意味情報，単語自体は上で述べたように実験の都合上真中よりの 2 つの形態素「を」「区切る」に対してのみ用いる．意味情報としては分類語彙表の分類番号上位 3 桁を用い,「を」は分類語彙表になく，ないという意味の印 “none” の情報を,「区切る」は分類語彙表にあり上位 3 桁の “217” の情報を用いる．単語自体はそのまま,「を」「区切る」となる．

こういった情報を用いてそれぞれの形態素の境界に対して文節区切り記号を挿入するかいなかを判定していくことで解析を行う．

本研究では，上記の問題設定で以下の教師あり学習手法で比較実験を行い，それぞれの手法の傾向や利点や弱点を調べることになる．

- 決定木学習
- 最大エントロピー法
- 用例ベース（最も類似した用例の利用）
- 決定リスト（確率と頻度を利用）
- 手法 1（排反な規則の利用）
- 手法 2（排反な規則と類似度の利用）☆☆☆

以降それぞれの学習手法ごとの文節まとめあげへの適用方法について述べる．

## 3. 各種学習手法ごとの文節まとめあげへの適用法

### 3.1 決定木学習

決定木学習とは簡単にいうと分類決定のための yes/no の分岐の木を学習するアルゴリズムのことである．本研究の決定木学習には，Quinlan の C4.5[5]を用いる．

解析に用いる情報としては，前章で述べたとおり品詞と品詞細分類と意味（意味情報）と単語（単語自体）の 4 つである．これをそれぞれ決定木学習の際の属性として用いる．ただし，両端 2 形態素については前章で述べたとおり品詞と品詞細分類しか用いないので，学習の際に用いられる属性の数は下記のように 12 個となる．

$$\text{属性数} = 2 + 4 + 4 + 2 = 12$$

$$\underbrace{\left\{\begin{array}{l}\text{品詞}\\ \text{細分類}\end{array}\right\}}_{\text{2 つ前の形態素：2 個}} + \underbrace{\left\{\begin{array}{l}\text{品詞}\\ \text{細分類}\\ \text{意味}\\ \text{単語}\end{array}\right\}}_{\text{1 つ前の形態素：4 個}} + \underbrace{\left\{\begin{array}{l}\text{品詞}\\ \text{細分類}\\ \text{意味}\\ \text{単語}\end{array}\right\}}_{\text{1 つ後の形態素：4 個}} + \underbrace{\left\{\begin{array}{l}\text{品詞}\\ \text{細分類}\end{array}\right\}}_{\text{2 つ後の形態素：2 個}}$$

たとえば，問題となっている部分が図 1 のようになっている場合は，“属性「2 つ前の形態素の品詞」の属性値は名詞” といった情報が 12 個用いられること

☆ 本研究では品詞体系は juman3.5[3]に準拠している．
☆☆ 本研究では解析に用いる情報の項目の個数を 4 つにおさえるために，品詞細分類の欄には，活用する単語については活用形を記述している．しかし，このようにすると品詞細分類を書く場所がなくなってしまうので，活用形がある単語で品詞細分類の項目もある単語については，品詞細分類の情報は品詞の情報と合わせて品詞の欄に記載することにしている．
☆☆☆「概要」「はじめに」「おわりに」で述べている「類似度の大きい排反な規則を用いる手法」はこの手法 2 を意味することに注意．

になる．

### 3.2 最大エントロピー法

最大エントロピー法はデータスパースネスに強い方法で，最近になって多くの研究者によって用いられるようになってきている[1),6)〜12)]．本研究の最大エントロピー法の実験では，文献 13) のシステムを用いる☆．解析は，そのシステムの出力から区切る確率と区切らない確率を計算し，その確率の大きい方であると推定することによって行う．

最大エントロピー法でも決定木学習と同じく解析に用いる情報は，品詞と品詞細分類と意味情報と単語自体の 4 つである．ただし，決定木学習とは異なり最大エントロピー法は素性の AND（組合せ）をシステムが自動的には考えないので，素性として与えるときにあらかじめ組合せを考慮しておく必要がある．

まず，各形態素内の情報，すなわち，品詞と品詞細分類と意味情報と単語自体の組合せを考える．4 つの情報があるので，あらゆる組合せを考える場合 $2^4-1$ 通りの組合せを考えることになる．しかし，これでは組合せの数が多く計算機の負荷が大きくなるので，品詞と品詞細分類と意味情報と単語自体の情報がこの順に細かくなる性質を利用して，本研究では 4 つの情報の組合せとしては以下の 4 種類を用いることにした．

情報 A：品詞

情報 B：品詞と品詞細分類

情報 C：品詞と品詞細分類と意味情報

情報 D：品詞と品詞細分類と意味情報と単語自体

ただし，両端 2 形態素については前章で述べたとおり品詞と品詞細分類しか用いないので，上記の 4 種類のうちの上の 2 つのみとなる．

次に各形態素間の組合せを考えると，以下のように情報の組合せの数は $2\times4\times4\times2=64$ となる．

| 2 つ前の形態素 | | 1 つ前の形態素 | | 1 つ後の形態素 | | 2 つ後の形態素 |
|---|---|---|---|---|---|---|
| { 情報 A | & | { 情報 A | & | { 情報 A | & | { 情報 A |
| 情報 B } | | 情報 B | | 情報 B | | 情報 B } |
| | | 情報 C | | 情報 C | | |
| | | 情報 D } | | 情報 D } | | |
| 2 個 | | 4 個 | | 4 個 | | 2 個 |

これらの組合せのほかにデータスパースネスの対策として，両端 2 形態素の情報がそれぞれ片方しか用いられない場合，両方用いられない場合，真中 2 形態素のうち片方しか用いられない場合を考えると，最大エントロピー法で用いられる素性の数は 1 カ所につき下記のように 152 個となる☆☆．

$$
\begin{aligned}
\text{素性の数} &= 2\times4\times4\times2\\
&+ 2\times4\times4\\
&+ \quad 4\times4\times2\\
&+ \quad 4\times4\\
&+ \quad 4\\
&+ \quad\quad 4\\
&= 152
\end{aligned}
$$

たとえば，問題となっている部分が図 1 のようになっている場合は，4 つの形態素の情報として 2 つ前の形態素の情報 B と，1 つ前の形態素の情報 D と，1 つ後の形態素の情報 C と，2 つ後の形態素の情報 A を用いるような素性は，“名詞：普通名詞 | 助詞：格助詞：none：を | 動詞：基本形：217 | 特殊” といったものとなり，このような素性を 152 個用いることになる．

### 3.3 用例ベース（類似度の利用）

用例ベースのアプローチとは，Nagao[15)]によって 1984 年に機械翻訳の問題において初めて提案された手法で，ある解析を行う際にそれと最も類似した用例を持ってきてその用例を利用して解を得るという手法である．近年になって文献 16)，17) において格フレーム選択や照応解析など機械翻訳以外の問題にも用いられるようになってきた．また，文献 18) も形態素解析の問題を用例ベースのアプローチで解いている研究といえよう．本節ではこの用例ベースのアプローチを利用して文節区切りを行う手法を述べる．

この用例ベースの手法でも公平に用いる情報は最大エントロピー法と同じく各形態素の情報としては，3.2 節にあげた情報 A，情報 B，情報 C，情報 D の 4 種類を用いる．

また，用例ベースでは入力と用例との類似度を定義する必要がある．なるべく微妙な違いまで検出できるような類似度を設定するために，入力と用例の類似するレベルとして，データスパースネスのことを考慮した最大エントロピー法と同じ 152 パターンを考えることにする．本論文では入力と用例の類似度 S を，それらの一致するレベルがこの 152 のパターンのどのレベルになっているかに応じて以下のように定める．

☆ 今は Web 上に存在していない．文献としては 14) を参照のこと．

☆☆ この手法で，たとえば京大コーパス 95 年 1 月 1 日分（形態素の境界の数が 25,814 のもの）で素性を取り出すと，その種類の数は 1,534,701 個となる．

| | | 文 | を | 区切る | . |
|---|---|---|---|---|---|
| | $s(x)$ | $m_{-2}$ | $m_{-1}$ | $m_{+1}$ | $m_{+2}$ |
| 情報なし | 1 | — | — | — | — |
| 情報 A | 2 | 名詞 | 助詞 | 動詞 | 特殊 |
| 情報 B | 3 | 普通名詞 | 格助詞 | 基本形 | 句点 |
| 情報 C | 4 | × | none | 217 | × |
| 情報 D | 5 | × | を | 区切る | × |

**図 2** 類似度の説明のための例
Fig. 2 Example of levels of similarity.

$$S = s(m_{-1}) \times s(m_{+1}) \times 10000 + s(m_{-2}) \times s(m_{+2}) \quad (1)$$

ただし，$m_{-1}$，$m_{+1}$，$m_{-2}$，$m_{+2}$ は，解析している形態素の境界に対して，1 つ前の形態素，1 つ後の形態素，2 つ前の形態素，2 つ後の形態素を意味する．また，$s(x)$ は，形態素 $x$ の形態素単位の類似度で以下のように定義される．

$s(x)$ =1(形態素 $x$ の情報がまったく一致しない場合)
2(形態素 $x$ の情報 A のレベルでのみ一致する場合)
3(形態素 $x$ の情報 B のレベルで一致する場合)
4(形態素 $x$ の情報 C のレベルで一致する場合)
5(形態素 $x$ の情報 D のレベルで一致する場合)

式 (1) は真中の 2 形態素が特に重要であると考えて作成したものである．本論文ではこの類似度の式を用いるが，この式よりも良い類似度の式があるかもしれないし，また，類似度自体なんらかの学習アルゴリズムで定めるということをした方が良いかもしれない．

類似度の例を示しておく．たとえば，問題となっている部分が**図 2** のようになっている場合，152 パターンのうち，“名詞 | 助詞 | 動詞 | 特殊” といった品詞レベルですべて一致し，より細かいレベルでは一致しないパターンの場合は $2 \times 2 \times 10000 + 2 \times 2$ で類似度は 40004 となる．また，“—| 助詞：格助詞：none：を |—|—” といったパターンで一致する場合は $5 \times 1 \times 10000 + 1 \times 1$ で類似度は 50001 となる．

本手法はこの類似度の値が最も高い用例を取り出してその用例が文節区切りになっていれば，文節区切り記号を挿入すると判定し，そうでない場合は文節区切り記号を挿入しないと判定する．類似度が等しい用例が複数あった場合，どの用例を利用して解析すればよいかが曖昧な場合がある．本論文の実験では，類似度が等しい用例が複数あった場合，その複数の用例において文節区切りになった数とそうでなかった数を調べ，多かった方であると推定するようにしている☆．

### 3.4 決定リスト（確率と頻度の利用）

決定リストは 1987 年に Rivest によって考案された手法で[23]，決定木のように規則を木構造で表現するのではなく，情報をすべて素性の AND などで展開し規則を 1 次元のリスト上に蓄えたものである．規則の優先順序をなんらかの方法であらかじめ決定しておき，この優先順序で規則を並べかえておく．このリストを探索し，一番初めに適用可能な規則のみを利用して解析するというものである．

ここでは展開した情報としては最大エントロピーのときと同様に形態素の境界 1 カ所につき 152 種類のパターンを利用することにする（京大コーパス 1 日分で約 153 万の規則を利用することになる）．

次に決定リスト上での規則の優先順序を決める必要がある．ここでは，Yarowsky[24]，西岡山ら[25]の手法を参考にして，各規則によって区切り記号が挿入される確率と，各規則の頻度を利用して解析することにする．つまり，まず各規則の確率によって規則をソートし，確率が同じ規則については規則の頻度の大きさによってソートする．

たとえば，パターン A“名詞：普通名詞 | 助詞：格助詞：none：を | 動詞：基本形：217 | 特殊：句点” が学習セット中に 13 回出現し，そのうち 10 回は文節の区切り記号を挿入すべき場所であるといったデータが得られたとする．また，パターン B “名詞 | 助詞 | 動詞 | 特殊” が学習セット中に 123 回出現し，そのうち 90 回は文節の区切り記号を挿入すべき場所であるといったデータが得られたとする．

このような情報は下記のような規則のように認識される．

パターン A
⇒ 区切部分になる確率 76.9% (10/ 13) 頻度 13
パターン B
⇒ 区切部分になる確率 73.2% (90/123) 頻度 123

上記のような規則を大量に作成し，これらをまず確率優先でソートし，確率が同一の場合は頻度順にソートしておく．そしてその規則のリストにおいて，先頭から順に調べていき一番最初に適用可能になった規則を用いて区切るかいなかを判定する．

☆ 用例ベースの手法を包含する考え方として，k-nearest neighbor method という手法がある[19)〜22)]．これは，最も類似した 1 個の用例を用いるかわりに最も類似した k 個の用例の多数決により解を求めるものである．$k = 3$ として 4 章の実験 1，実験 2 のデータで実験したところ，いずれも精度が落ち，$k$ の値を増やせば増やすほど精度が悪くなる状態になっていた．このことはデータがスパースになっていたり類似度の定義式がかなり適切になっていたりする場合に生じることなので，本問題がそのような場合になっているということが予想される．

| | | | | | |
|---|---|---|---|---|---|
| 規則 A: | パターン A | ⇒ | 継続部分になる確率 | 100% ( 34/ 34) | 頻度 34 |
| 規則 B: | パターン B | ⇒ | 区切部分になる確率 | 100% ( 33/ 33) | 頻度 33 |
| 規則 C: | パターン C | ⇒ | 区切部分になる確率 | 100% ( 25/ 25) | 頻度 25 |
| 規則 D: | パターン D | ⇒ | 区切部分になる確率 | 100% ( 19/ 19) | 頻度 19 |
| 規則 E: | パターン E | ⇒ | 区切部分になる確率 | 81.3% (100/123) | 頻度 123 |
| 規則 F: | パターン F | ⇒ | 区切部分になる確率 | 76.9% ( 10/ 13) | 頻度 13 |
| 規則 G: | パターン G | ⇒ | 継続部分になる確率 | 57.4% (310/540) | 頻度 540 |
| ..... | ..... | ..... | | | |

**図 3** 得られた規則の例
Fig. 3 Example of obtained rules.

### 3.5 手法 1（排反な規則を利用）

前節までは，文節まとめあげに用いる手法として，有力な既存の 4 つの手法を説明してきた．本節と次節では，我々が考えた手法について説明する．

本節の手法でも解析に用いる情報は先の 3 手法と同じように形態素の境界 1 カ所につき 152 種類のパターンとする．決定リストの節と同様に学習セットのデータのすべての境界において，152 種類のパターンのデータの統計情報をとってみる．その結果，ある境界において**図 3** にあるような情報が得られたとしよう．

今ある問題を解くときに図 3 のパターン A〜G が適用可能であったとしよう．もし，決定リストの方法ならば最初に適用される規則 A のみを利用するので継続部分であると推定してしまう☆1．しかし，規則 B，C，D を見てみると各規則ごとの頻度は規則 A に比べると小さいがそれぞれの頻度を足し合わせると規則 A の頻度よりも大きいので，規則 B，C，D に従って区切り部分であると解析した方がより望ましいと思われる．本節の手法 1 はこの考え方に従うものである．ただし，頻度の足し算の部分は，単純に頻度を足し合わせるのではなく規則の頻度を算出する際に用いた用例の個数に基づいて行う．つまり，最も大きい確率を持つパターンを満足する複数の用例において区切るかいなかの個数の大きい方であると推定する☆2,☆3．

たとえば，上記の例で規則 B，C，D を満足する用例の個数が 65 個であったとする（複数の規則に重複して利用される用例があるので，用例の総数は規則 B，C，D の頻度，33，25，19 の総和の 77 個よりは小さくなる）．この場合，確率 100%の規則を利用する用例において，区切る場合は 65 個，区切らない場合は 34 個となり，区切ると解析することになる．このとき，頻度の総和の計算でなく用例の個数の総和の計算になるので，計算が困難に思えるかもしれないが，各規則に頻度だけでなくどの用例が用いられていたのかを記載しておくことで容易に計算できる．

ここで，確率が 100%になっている規則を，区切りか継続かのいずれかのみという排他的な集合に分割されている場合の規則という意味で，排反な規則と呼ぶこととする．本問題では適用される規則は 1 カ所につき最大で 152 個もあるので，排反な規則が適用されて推定されることが多い☆4．本節の手法 1 は，この排反な規則を総合的に利用して解析するものなので,（この排反な規則で利用される用例の個数によって解を求めるものなので），排反な規則を利用する手法とも呼ぶこととする☆5．

確率が 100%でない規則を利用して問題を解くとい

☆1 決定リストと本節の手法 1 の最も大きな違いは，決定リストはあらかじめソートしておいた規則のリストの中からただ 1 つの規則を選ぶのに対し，本節の手法 1 は解析の段階で動的に適用可能な規則集合を取り出し，その規則集合を総合的に分析して解析するところにある．

☆2 本節の手法 1 では規則の頻度を利用しないが，3.4 節の確率と頻度による決定リストの方法と同じように確率の他に頻度を利用した方がよいかもしれない．この場合は確率と頻度が最も大きい規則を満足する用例の多数決によって解を求めるとよい．4 章の実験 1，実験 2 のデータを用いて上記の手法の実験を行ったが，頻度を用いた方がよいかどうかははっきりした結果は得ていない．

☆3 本節の手法 1 では最も大きい確率を持つ規則を満足する用例の集合における多数決で解を決定するが，この多数決の際に頻度の大きい規則を満足する用例には相応の重みをつけて多数決を行うということも考えられる．

☆4 排反な規則が適用されて解析された問題の割合は 4 章の実験 1 のデータで 99.30%（16864/16983）であった．このことからもほとんどの問題が排反な規則によって解かれていることが分かる．ただし，4 章からも分かるように解析精度が 99%程度なので排反な規則の利用の割合をさらに高めるように解析に用いる情報を増やさない限りはそれほど精度があがらない状態なのかもしれない．

☆5 3.4 節の確率と頻度による決定リストの方法も確率の最も大きい規則を利用する方法なので，本問題の場合排反な規則が適用されて推定されることが多く，これも排反な規則を用いる手法と呼んでもよさそうであるが，この手法では本来排反な規則を用いることを前提として作成されておらず，確率も頻度も同じ場合のことも考慮されていないものであり（排反な規則が多く適用可能な場合は確率も頻度も同じ規則が適用される場合が多く出現する），排反な規則を用いる手法と呼ぶにはふさわしくない．また，本節の手法 1 のように用例の個数による多数決は，一番最初に適用可能となった規則を 1 つだけ利用する決定リストという枠組みを採用する手法では行うことができず，確率も頻度も同じ規則が複数存在する場合の処理は困難になっている．

うのは，あらかじめ間違う可能性がある規則を用いて解いているということである．文節まとめあげの問題に限らず一般に従来の機械学習のほとんどは確率が100%でない規則を用いて解いているものであったが，これでは高い精度の解析は望めない．今後より高い精度の解析を目指すならば本節のような排反な規則を利用して解析するという考え方をとる必要がある．解析に用いる情報が少ない場合は本節の手法を採用しても排反な規則があまり適用されない場合がある．その場合は解析に用いる情報を増やすことで，多くの排反な規則が適用される状態にする必要がある．

しかし，排反な規則を用いるからといってそれで十分なのではなく，偶然その学習データでは排反な規則になったという無意味な規則も存在する．このような無意味な規則をいかにして除いていくかが今後の課題として残っている．

### 3.6 手法 2（排反な規則と類似度を利用）

本節の手法は用例ベースの方法と手法 1，すなわち，類似度を用いる手法と，排反な規則を用いる手法を組み合わせた方法である．

解析に用いる情報はいままでと同じ 152 パターンである．この 152 パターンを前節と同様に規則として扱い，最も確率の高い規則を利用して解析する．最も確率の高い規則が複数ある場合は，用例ベースの手法の節の類似度の値を用いこの値が最も高い規則を利用して解析する．また，類似度の値が等しい規則が複数ある場合は手法 1 と同様にその複数の規則で用いられる用例における区切るかいなかの個数によって解析する．

ただし，頻度が 2 以上の排反な規則が存在するときは，頻度が 1 の排反な規則はあまり信用できないとして省いて上記の計算をしている．つまり，類似度の高い頻度が 1 の排反な規則よりも類似度の低い頻度が 2 以上の排反な規則を優先して用いるようになっている☆．

本節の手法 2 は，前節の手法 1 で述べた無意味な排反な規則を取り除くために，類似度という尺度を用いている手法であるともとらえることができる．

☆ 頻度 1 の排反の規則を軽視する手法の有効性は，頻度 1 の排反の規則を軽視しない手法，頻度 1，2 の排反の規則を軽視する手法の実験を 4 章の実験 1，実験 2 のデータを用いて行い，それらの精度が頻度 1 の排反の規則を軽視する手法の精度よりも F-measure で 0.1%程度低いことを調べることによって確認している（F-measure の定義については 4 章を参照のこと）．手法 1 でも同様の実験を行ってみたが，手法 1 では頻度 1 の規則を軽視すべきかどうかははっきりしなかった．

**表 1** 実験 1 での学習セット（1 月 1 日）での解析精度
Table 1 Results of learning set of Experiment 1.

| 手法 | F | 再現率 | 適合率 |
|---|---|---|---|
| 決定木 | 99.58% | 99.66% | 99.51% |
| MEM | 99.20% | 99.35% | 99.06% |
| 用例ベース | 99.98% | 100.00% | 99.97% |
| 決定リスト | 99.98% | 100.00% | 99.97% |
| 手法 1 | 99.98% | 100.00% | 99.97% |
| 手法 2 | 99.98% | 100.00% | 99.97% |
| knp 2.0b4 | 99.23% | 99.78% | 98.69% |
| knp 2.0b6 | 99.73% | 99.77% | 99.69% |

形態素の境界の数 25,814．区切り部分の数 9,523．

**表 2** 実験 1 でのテストセット（1 月 3 日）での解析精度
Table 2 Results of test set of Experiment 1.

| 手法 | F | 再現率 | 適合率 |
|---|---|---|---|
| 決定木 | 98.87% | 98.67% | 99.08% |
| MEM | 98.90% | 98.75% | 99.06% |
| 用例ベース | 99.02% | 98.69% | 99.36% |
| 決定リスト | 98.95% | 98.43% | 99.48% |
| 手法 1 | 98.98% | 98.54% | 99.43% |
| 手法 2 | 99.16% | 98.88% | 99.45% |
| knp 2.0b4 | 99.13% | 99.72% | 98.54% |
| knp 2.0b6 | 99.66% | 99.68% | 99.64% |

形態素の境界の数 16,983．区切り部分の数 6,166．

## 4. 実験および考察

実験は京大コーパス[26]の毎日新聞 95 年 1 月 1 日〜95 年 1 月 5 日の記事で行った．システムの入力となる形態素の情報はコーパスに付与されているものを用いた．

どの教師あり学習の手法が最も有効かどうかを調べるために，以下の 2 つの実験を行った．

- 実験 1
  95 年 1 月 1 日を学習セット，95 年 1 月 3 日をテストセットとする実験
- 実験 2
  95 年 1 月 4 日を学習セット，95 年 1 月 5 日をテストセットとする実験

3.5 節，3.6 節で述べた手法 1，2 は上記の実験 1 を試行して作成したので，実験 1 は若干クローズデータの意味合いがあるため，実験 2 を行っている．

実験結果を**表 1**〜**表 4** に示す．表では，3 章で述べた 5 つの手法の他に比較のために人手ルールベースによる構文解析システム knp 2.0b4[27]と knp 2.0b6[2] の精度も示しておいた．knp の結果については表に示す「学習セット」「テストセット」に意味はない．knp で実験する際も knp の入力となる形態素の情報はコーパスから得ている．95 年 1 月 5 日の実験では knp の

**表 3** 実験 2 での学習セット（1 月 4 日）での解析精度
Table 3 Results of learning set of Experiment 2.

| 手法 | F | 再現率 | 適合率 |
|---|---|---|---|
| 決定木 | 99.70% | 99.71% | 99.69% |
| MEM | 99.07% | 99.23% | 98.92% |
| 用例ベース | 99.99% | 100.00% | 99.98% |
| 決定リスト | 99.99% | 100.00% | 99.98% |
| 手法 1 | 99.99% | 100.00% | 99.98% |
| 手法 2 | 99.99% | 100.00% | 99.98% |
| knp 2.0b4 | 98.94% | 99.50% | 98.39% |
| knp 2.0b6 | 99.47% | 99.47% | 99.48% |

形態素の境界の数 27,665．区切り部分の数 10,143.

**表 4** 実験 2 でのテストセット（1 月 5 日）での解析精度
Table 4 Results of test set of Experiment 2.

| 手法 | F | 再現率 | 適合率 |
|---|---|---|---|
| 決定木 | 98.50% | 98.51% | 98.49% |
| MEM | 98.57% | 98.55% | 98.59% |
| 用例ベース | 98.82% | 98.71% | 98.93% |
| 決定リスト | 98.75% | 98.27% | 99.23% |
| 手法 1 | 98.79% | 98.54% | 99.43% |
| 手法 2 | 98.90% | 98.65% | 99.15% |
| knp 2.0b4 | 99.07% | 99.43% | 98.71% |
| knp 2.0b6 | 99.51% | 99.40% | 99.61% |

形態素の境界の数 32,304．区切り部分の数 11,756.

出力が不完全な文があったが，この文はすべての手法の実験で除いた．

表中の “F” は F-measure を意味し，再現率と適合率の調和平均である☆1．再現率，適合率は区切り部分に対するもので，それぞれシステムの正解した区切り部分の個数をすべての区切り部分の個数で割ったもの，システムの正解した区切り部分の個数をシステムが区切り部分と推定したものの個数で割ったものを意味する．

最大エントロピー法（MEM）での実験では文献 13) のシステムを用いたが，低頻度の素性が存在している場合システムが動かなかったので，“ある素性” と “その素性が存在するときに区切るかいなかの情報” の組合せ☆2 の頻度が 11 以下のものは消して実行した（頻度が 10 以下のものを消して実行した場合システムが動かなかったので，しかたなくこのようにした）．このため，これらの素性を削除しない場合最大エントロピー法による方法ではここに示した精度以上のものが出る可能性がある．また，このシステムではパラメータの繰返し学習が行われるが，本研究では繰り返し回数はとりあえず 200 回に固定して実験を行った．

また，決定木学習においては，C4.5 のシステムを用いたが，実験においては属性値のグループ化を行う -s オプション☆3をつけて実行した．しかし，-s オプションをつけると，属性値の種類の数が大きい場合に計算にかなりの時間がかかる．そこで，頻度が 10 未満の属性値については「その他」という属性値を用意しすべてその属性値を割り当てて実験を行った．

表 1〜表 4 の実験結果から以下のようなことが分かる．

- テストセットにおいては，決定木よりも最大エントロピー法の方が若干ではあるが精度が良かった☆4．最大エントロピー法の欠点として，自動的に素性の組合せを学習しないというものがあるが，本実験での最大エントロピー法ではあらゆる素性の組合せを用いていたので，その欠点が克服され，決定木よりも精度が良くなったと思われる．
- 現時点では，決定リストの方が最大エントロピー法よりも精度が高かった．ただし，最大エントロピー法はシステムの都合で素性を削除しないとシステムが動作しなかっただけなので，これらの素性も利用できるようになると精度が向上する可能性がある．
- 用例ベースは既存の 4 手法の中で最も精度が良かった．
- 手法 1（単純に排反な規則を用いる手法）は用例ベースよりも精度が低かったが，決定リストよりは精度が良かった．手法 1 が決定リストよりも良かった大きな理由は，決定リストでは確率と頻度が等しい規則の集合においてそれらのうちどれを選ぶのかがランダムになっていることによる．

☆1 つまり，再現率，適合率の逆数を足して 2 で割ったものを逆数にした値のこと．

☆2 最大エントロピー法のプログラムの入力として，“素性” と “その素性が存在するときに区切るかいなかの情報” の組合せの頻度が必要となる．この頻度が小さい組合せを省くとシステムが動作するようになったので，本研究ではそのようにして実験を行った．しかし，動作させるためにはここでの対処よりも良い方法があるかもしれない．
また，素性と区切るかいなかの情報の組合せで捨てているので，単純に素性の頻度で捨てる場合に比べて，より多くの素性を捨てていることになる．

☆3 このオプションは，たとえば，属性 A の属性値が { 名詞，副詞，形容詞，動詞 } の 4 つであったとすると，これを {(名詞，副詞)，(形容詞，動詞)} の 2 つの属性値にグループ化して決定木の分岐を行うものである（この場合 4 分岐のところが 2 分岐になる）．

☆4 本実験ではともに素性を捨てているので正確な比較になっていない．しかし，決定木では形態素ごとの情報で捨てていて，最大エントロピー法は 4 形態素の AND をとった後の情報（厳密にはさらに区切るかいなかの情報との組合せをとった情報）で捨てており，同じ頻度で捨てる場合であっても最大エントロピー法の方がより多くの素性を捨てることになっている．このため，若干でも精度が良い最大エントロピー法の方が決定木よりもよさそうであると推測される．

- 手法 2（排反な規則と類似度を用いる手法）は学習アルゴリズムの中では最も高い精度を得た☆.
- 用例ベース，決定リスト，手法 1，手法 2 の 4 つの手法は学習セットにおいては 100%にきわめて近い精度を出している．これらの手法は特に学習セットにおいては強いことが分かる．
- 類似度を用いる 2 手法（用例ベース，手法 2）は，類似度を用いない他の手法よりもつねに精度が良かった．適切な類似度を設定できるときは類似度の利用がかなり有効であることを示している．
- knp でも解析してみたが，テストセットにおける精度は knp のものが最も高かった．
- knp においては，knp 2.0b4 と knp 2.0b6 の 2 つのバージョンのもので実験を行ったが，これは knp 2.0b6 の方が断然精度が良かった．人手によるシステム向上もかなり効果的であることが分かる．しかし人手による規則のメンテナンスには量的限界があるので，人手によるシステム向上がいつも効果的とは限らない．

以上の実験から今のところ手法 2 がいろいろな条件つきではあるが，文節まとめあげのための機械学習の手法としては最も優れていそうであると思われる☆☆.

本研究では用例ベースや排反な規則を用いる手法（手法 1，手法 2）が良さそうな結果を得た．しかし，問題が難しくなり考慮すべき情報がたくさんある問題では，あらゆる情報の組合せを扱うことが難しくなり，用例ベースや排反な規則の方法を用いることが困難になってくる☆☆☆．また，12 個の属性しか用いていない決定木の手法では，まだまだ属性を増やして実験をすることが可能であり，解析に用いる情報を増やすことで精度向上を行える可能性があるが，用例ベースや排反な規則を用いる手法では現在の計算機のパワーでは難しい．それでもこれらのことは計算機技術の向上により克服されるであろうから，解析に用いる情報が等しい場合に精度がどのようになるかを調べる本研究のようなアプローチは重要である．

コツコツ |不足我慢し，
余裕を | 持って |不足退けた
会社を | グループ分け |×して
最も | 慣れ |×親しんでいる

**図 4** knp で誤ったが手法 2 で正解した例
Fig. 4 Cases which were analyzed incorrectly with KNP but correcly with Method 2.

また，knp で誤ったが手法 2 では正解した例を**図 4** に示しておく．図中「不足」印がついている部分は knp が誤って区切らなかったものを意味し，×印がついている部分は knp が誤って区切ったものを意味する．実験 1 のテストセットでは knp2.0b6 の F-measure は 99.66%であったが，knp2.0b6 と手法 2 のどちらかが正解していれば正解とするとき F-measure は 99.83%となった．knp2.0b6 は精度が良いが，手法 2 で正解して knp2.0b6 で誤るものも少しはあること（0.17% = 99.83% − 99.66%，つまり knp2.0b6 の誤り部分の 0.33%の約半分あること）が分かる．knp2.0b6 と手法 2 を併用すると精度が上がる可能性があることが分かる．

また，文節まとめあげの問題を機械学習の手法で解いている先行研究として，Zhang らのもの[28]がある．この研究は決定木学習の手法により文節区切り位置を推定するものである．そこでの解析に用いている情報は区切るか否かを判定する境界の前後 2 つの形態素しか用いておらず，さらにその形態素の品詞の情報のみを用いているだけであり，本研究で用いている情報よりも少ない．実験は，ATR コーパスと EDR コーパスの 2 種類で行っており，精度は ATR コーパスで 98.9%（=(521+763)/(521+763+13+1)），EDR コーパスで 96.2%（=(2,502+4,341)/(2,502+4,341+62+ 205)）の精度を得ている（F-measure などの精度では，ATR コーパスで再現率 97.6%（=521/(521+13)），適合率 99.8%（=521/(521+1)），F-measure 98.7%で，EDR

☆ ただし，用例ベースの方法では類似度の式を式 (1) よりも適切なものに作成し直すことで精度が向上し手法 2 の精度を上回る可能性がある．しかし，その可能性は以下の理由により小さいと思われる．
– 3.3 節の最後の脚注のように現時点でも類似度の定義式がかなり適切になっていると予想させる事実がある（ただし，最適とは限らない）.
– 手法 2 では用例ベースの方法とまったく同じ類似度の式 (1) を用いているためこの式が改善されると用例ベースの方法だけでなく手法 2 の精度も向上することが予想される．
また，類似度の式が最適なときに用例ベースの方法の方が手法 2 よりも精度が高くなるとしても，類似度の式を最適なものにすることは難しいことなので，用例ベースの方法の方が手法 2 よりも有用であるとはいいがたい．

☆☆ 本研究の実験での各手法間の精度の差は微小であった．しかし，実験 1，2 と 2 回の実験を行っていること，実験に用いたデータとしてあらかじめ客観的にタグがふられた万のオーダーのコーパスを用いていること，精度が 99%前後あってそこでの 0. 数%の精度差は大きいことから，精度の差を議論できるデータであると考えている．

☆☆☆ 手法 1 では情報が増えると組合せの数が爆発的に増え手におえなくなる．このようなときには，規則の展開の際にすでに排反になっている規則は展開せずにそこでとどめ，排反でない規則のみ情報を追加して展開するといったことを行うと，少しは情報の増加による組合せの爆発をおさえれるかもしれない．

コーパスで再現率 97.6%（=2502/(2502+62)），適合率 92.4%（=2502/(2502+205)），F-measure 94.2% である）．ただし，knp などの高精度で文節まとめあげを行うシステムとの比較実験を行っておらず，コーパスも本研究のものと異なるので，本研究と比較するのは難しい．

## 5. おわりに

本研究では，文節まとめあげを対象として種々の教師あり機械学習の比較を行った．本研究の実験の結果では，既存の 4 手法の間には下記のような優劣があった．

用例ベース ≥ 決定リスト
≥ 最大エントロピー ≥ 決定木学習

また，類似度の大きい排反な規則を用いるという新しい手法を提案し，それが用例ベースの手法と同程度かもしくは若干高い精度をあげた．適切な類似度を人手で容易に与えることができる問題の場合は，本論文の類似度と排反な規則を用いる手法を用いるのが最も良さそうである．

また，人手で詳細に作成した規則を用いて高精度に文節まとめあげを行う knp とも比較実験を行った．今回の実験では knp の方が精度が良かったが，人手で詳細な規則を作成する方法ではメンテナンスの問題や異なる分野のデータに適応させるのが困難であるという問題があり，また機械学習の手法ではコーパスを増やすことや手法自体を改良することで精度が向上する可能性があるため，即座に，人手で規則を作成する方が良いとはいえない．また，knp で誤って手法 2 で正解したものが少なからず存在しているので，両方の手法をうまく組み合わせて精度向上を図るということも考えられる．

## 参考文献


1) 内元清貴，関根　聡，井佐原均：ME による日本語係り受け解析，情報処理学会自然言語処理研究会 NL128-5 (1998).
2) 黒橋禎夫：日本語構文解析システム KNP 使用説明書 version 2.0b6，京都大学大学院情報学研究科 (1998).
3) 黒橋禎夫，長尾　真：日本語形態素解析システム JUMAN 使用説明書 version 3.5，京都大学大学院工学研究科 (1997).
4) 国立国語研究所：分類語彙表，秀英出版 (1964).
5) キンラン，J.R.：AI によるデータ解析，トッパン (1995).
6) Berger, A.L., Pietra, S.A.D. and Pietra, V.J.D.: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).
7) Ratnaparkhi, A.: A Maximum Entropy Model for Part-Of-Speech Tagging, *Proc. Empirical Method for Natural Language Processings*, pp.133–142 (1996).
8) Ratnaparkhi, A.: A Linear Observed Time Statistical Parser Based on Maximum Entropy Models, *Proc. Empirical Method for Natural Language Processings* (1997).
9) Borthwick, A., Sterling, J., Agichtein, E. and Grishman, R.: Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition, *Proc. 6th Workshop on Very Large Corpora*, pp.152–160 (1998).
10) 江原暉将：最大エントロピー法を用いた日本語文節間係り受け整合度の計算，言語処理学会第 4 回年次大会，pp.382–385 (1998).
11) 白井清昭，乾健太郎，徳永健伸，田中穂積：統計的構文解析における構文的統計情報と語彙的統計情報の統合について，自然言語処理，Vol.5, No.3, pp.85–106 (1998).
12) 乾　裕子，内元清貴，村田真樹，井佐原均：文末表現に着目した自由回答アンケートの分類，情報処理学会自然言語処理研究会 NL128-25 (1998).
13) Ristad, E.S.: Maximum Entropy Modeling Toolkit, Release 1.6 beta, http://www.mnemonic.com/software/memt (1998).
14) Ristad, E.S.: Maximum Entropy Modeling for Natural Language, ACL/EACL Tutorial Program, Madrid (1997).
15) Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, *Artificial and Human Intelligence*, pp.173–180 (1984).
16) Kurohashi, S. and Nagao, M.: A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary, *IEICE Trans. Information and Systems*, Vol.E77–D, No.2, pp.227–239 (1994).
17) 村田真樹，長尾　真：表層表現と用例を用いた照応省略解析手法，言語理解とコミュニケーション研究会 NLC97-56 (1998).
18) 山下達雄，松本裕治：品詞タグ付きコーパスを直接利用した形態素解析，言語処理学会第 4 回年次大会 C4-4 (1998).
19) Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, Academic Press Inc. (1972).
20) 富浦洋一，日高　達：k-NN 推定法に基づく統語的曖昧さの解消法，言語理解とコミュニケーション研究会 NLC96-7，pp.39–45 (1996).
21) 岡本青史，太田唯子，湯上伸弘：k–最小近傍法におけるノイズの影響，人工知能学会全国大会予稿集 (1997).

22) Okamoto, S. and Yugami, N.: An Average-Case Analysis of the k-Nearest Neighbor Classifier for Noisy Domains, *IJCAI '97* (1997).
23) Rivest, R.L.: Learning Decision Lists, *Machine Learning*, Vol.2, pp.229–246 (1987).
24) Yarowsky, D.: Decision Lists For Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *32th Annual Meeting of the Associtation of the Computational Linguistics*, pp.88–95 (1994).
25) 西岡山滋之，宇津呂武仁，松本裕治：コーパスからの日本語従属節係り受け選好情報の抽出，言語理解とコミュニケーション研究会 NLC98-11, pp.31–38 (1998).
26) 黒橋禎夫，長尾　真：京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会，pp.115–118 (1997).
27) 黒橋禎夫：日本語構文解析システム KNP 使用説明書 version 2.0b4，京都大学大学院情報学研究科 (1997).
28) Zhang, Y. and Ozeki, K.: The Application of Classification Trees to Bunsetsu Segmentation of Japanese Sentences, *Journal of Natural Language Processing*, Vol.5, No.4, pp.17–33 (1998).

(平成 11 年 1 月 13 日受付)
(平成 11 年 10 月 7 日採録)

**村田　真樹**（正会員）

1993 年京都大学工学部卒業．1995 年同大学院修士課程修了．1997 年同大学院博士課程修了，博士（工学）．同年，京都大学にて日本学術振興会リサーチ・アソシエイト．1998 年郵政省通信総合研究所入所．研究官．自然言語処理，機械翻訳の研究に従事．言語処理学会，ACL 各会員．

**内元　清貴**（正会員）

1994 年京都大学工学部卒業．1996 年同大学院修士課程修了．同年郵政省通信総合研究所入所，郵政技官．自然言語処理の研究に従事．言語処理学会，ACL 各会員．

**馬　　青**

1983 年北京航空航天大学自動制御学部卒業．1987 年筑波大学大学院理工学研究科修士課程修了．1990 年同大学院工学研究科博士課程修了．工学博士．1990〜93 年（株）小野測器勤務．1993 年郵政省通信総合研究所入所，主任研究官．人工神経回路網モデル，知識表現，自然言語処理の研究に従事．日本神経回路学会，言語処理学会，電子情報通信学会各会員．

**井佐原　均**（正会員）

1978 年京都大学工学部電気工学第二学科卒業．1980 年同大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所関西支所知的機能研究室室長．自然言語処理，機械翻訳の研究に従事．言語処理学会，人工知能学会，日本認知科学会，ACL 各会員．