自然言語処理　144－16
(2001. 7. 17)

機械学習手法を用いた日本語格解析
— 教師信号借用型と非借用型，さらには併用型 —

村田 真樹

独立行政法人 通信総合研究所
けいはんな情報通信融合研究センター
〒619-0289 京都府相楽郡精華町光台 2-2-2
TEL:0774-95-2424　FAX:0774-95-2429　murata@crl.go.jp

あらまし

本稿では教師信号借用型機械学習手法を提案する．この手法は，解析対象としている分類のタグがふっていないコーパスから教師信号を借用する手法である．また，解析対象としている分類のタグがふっているコーパスから得られる教師信号も併せて用いる併用型機械学習手法も提案する．これらの手法はあらゆる省略解析に用いることができる．これらの手法を具体的に日本語格解析に適用し，実際にその有効性を確かめた．

キーワード　機械学習，格解析，教師信号借用型，省略解析

Japanese case analysis based on a machine learning method that uses borrowed supervised data

Masaki Murata

Keihanna Human Info-communication Research Center,
Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
TEL:+81-774-95-2424　FAX:+81-774-95-2429　murata@crl.go.jp

Abstract

We have developed a new machine learning method that uses borrowed supervised data. In this method, supervised data is borrowed from corpola that do not have annotated tags related to the problems we are currently trying to solve. We have also developed a new machine learning method that uses both supervised data received from corpola that have annotated tags related to the problems we are currently addressing and borrowed supervised data that does not have those annotated tags. These methods can both be used with any type of ellipsis resolution. In this paper, we will show how these methods can be applied to Japanese case analysis as well as confirm their effectiveness.

key words　Machine Learning, Case Analysis, Type Borrowing Supervised Data, Ellipsis Resolution

1 はじめに

本稿では教師あり機械学習手法を用いて日本語格解析の問題を扱う．本稿での日本語格解析は，文の一部が主題化，もしくは，連体化などをすることにより隠れている表層格を復元することを意味する．例えば，

「りんごは食べた．」

という文だと「りんごは」の部分が主題化しているが，これを表層格に戻すと「りんごを」である．このような場合はこの「は」の部分を「ヲ格」と解析する．また，

「昨日買った本はもう読んだ．」

という文だと「買った本」の部分が連体化しているが，これを表層格に戻すと「本を買った」である．このような場合はこの連体の部分を「ヲ格」と解析する．

日本語格解析の先行研究としては，黒橋らの既存の格フレームを利用するもの(1)，河原らの格フレームを生コーパスから構築しそれを利用するもの(2)，阿部川らの生コーパスでの頻度情報を利用するもの(3)，Baldwin の格情報つきコーパスを用いた機械学習手法[1]を用いたもの(5) などがある．ただし，阿部川らの研究，Baldwin の研究は連体化における格解析のみを扱っており，主題化などにおける格解析は扱われていない．

本稿の格解析のアプローチは，阿部川らの方法や Baldwin の方法に近く，格情報つきコーパスや格情報のついていないコーパスを用いた機械学習手法を用いる．しかし，阿部川らや Baldwin と異なり，連体化における格解析だけでなく主題化などにおける格解析も扱う．

また，格情報のついていないコーパスを用いた方法は，阿部川らの頻度を求めて最尤推定により格を求める方法と異なり，教師あり機械学習手法を用いるものとなっている．このとき，教師信号を借用するためこの格情報のついていないコーパスなどを用いた機械学習手法のことを，本稿では，**教師信号借用型機械学習手法**とよぶ．また，この格情報のついていないコーパスを格解析に用いることができるのが，格解析が省略解析と等価なためであることを指摘し，このあたりの理論を整理するとともに，動詞省略補完(6)，質問応答システム(7, 8, 9) などより広範な省略解析にもこの種の方法を用いることができることを指摘する．さらに，借用しない教師信号 (非借用型教師信号) も併用して用いる**併用型の機械学習手法**も提案する．

さらに実際にサポートベクトルマシン法といった機械学習手法を用いて借用型，非借用型，併用型の機械学習手法を用いた実験を日本語格解析全般の問題で行なう．

あらかじめ本稿の主張点をまとめておくと以下のようになる．

- 教師信号借用型機械学習手法の提案を行ない，またその周辺理論の整理を行なった．この手法は解析対象用の教師信号のタグがふられていないコーパスでも，問題が省略解析に類似する問題ならば，それを教師信号として用いることができるというものである．この手法は単に格解析に用いることができるだけでなく，省略解析に類似するより広範な問題においても利用できる．さらに借用しない元の教師信号も併用する併用型手法も提案した．
- 借用型機械学習と非借用型機械学習と併用型機械学習の比較実験をはじめて行なった．借用型機械学習はランダムな解析より精度がよくまた分類先ごとの精度を平均した精度では非借用型手法よりも精度がよかった．併用型機械学習は全事例の重みを等価と考える評価基準だけでなく各分類先での精度の平均を評価基準としたものでも精度がよかった．これらのことから，借用型機械学習と併用型機械学習の有効性が示された．
- 本研究は，連体化だけでなく主題化の問題も含めて日本語格解析の問題全般を機械学習手法ではじめて扱ったものである．日本語言語処理の分野では，形態素解析，構文解析までは多くの研究が存在しある一定の成果がでている．今後はその次の段階の意味解析が重要になると思われる．本稿はその重要となる意味解析の主要部分である日本語格解析の問題を，機械学習手法という堅実な研究手段[2]を用いて扱ったものとなっている．

2 教師信号借用型機械学習手法

本節では，教師信号借用型機械学習手法の理論的背景，理論の整理を行なう．

筆者は過去に用例ベースを用いた様々な照応省略解析を行なっており，その集大成として用例ベースに基づく照応省略解析手法に関する論文(10) をすでに執筆している．その論文においても指摘していることだが，一般に照応省略解析には照応省略に関する情報が付与されていないコーパスを利用することができる．例えば，以下の例を考える．

(例) みかんを買いました．これを食べました．
用例：「ケーキを食べる．」
用例：「りんごを食べる．」

このとき，「これ」の指示先を推定したいとする．この場合「ケーキを食べる」「りんごを食べる」といった用例を使って「を食べる」の前には食べ物がきそうだと予想し「みかん」を指示先と推定できる．これらの用例は照応省略に関する情報が付与されていないただの文でよいのである．次に照応省略に関する情報が付与された用例を利用して解くことを考える．そのような用例は例え

[1] Baldwin は機械学習手法としては k 近傍法の一種の TiMBL(4) を用いている．

[2] 機械学習手法を用いる研究アプローチは，データの拡張，手法の改良が今後期待されるために自然な発展が予想される堅実なものである．これに対し人手の規則に基づく方法は，規則の拡張や規則の調節を人手で行なわねばならずコストが高く発展が厳しいものになると予想される．

ば以下のような形をしている．

用例：りんごを買いました．これを食べました．
「これ」が「りんご」を指す．

「りんごを買いました．これを食べました．」という文に対して，その文の「これ」が「りんご」を指すと，照応省略に関する情報を付与しておくのである．このような用例を用いることでも，「りんご」を指す例があるのなら，「みかん」も指すだろうと判断して指示先を推定することができる．しかし，このような照応省略に関する情報を付与することは大変労力のいることでありこのような情報を用いず，前者のような照応省略に関する情報が付与されていない用例でも解くことができるのならば，その方がコストが小さく，その意味で照応省略に関する情報が付与されていない用例を解析に利用できることは価値があることである．

この種の照応省略に関する情報が付与されていない用例を用いた省略解析の研究は数多く存在する．

1. 指示詞・代名詞・ゼロ代名詞照応解析(11)

(例) みかんを買いました．そして(φを)食べました．
用例：「りんごを食べる．」

2. 間接照応解析

「A の B」の形をした用例を利用することで「屋根」が前文の「家」の屋根であると推定する(12)．

(例)「家がある．屋根は白い．」
用例：「家の屋根」

3. 動詞の省略補完

「そううまくいくとは」の後ろに省略されている動詞部分を「そううまくいくとは」を含む文を集めてきてそれを用いて推測する(6)．

(例)「そううまくいくとは」
用例：「そんなにうまくいくとは思えない．」

4. 「A の B」の意味解析

「A の B」の意味関係は多様である．この意味関係には動詞で表現できるものがある．そのような動詞は，名詞 A,B と動詞との共起情報から推測できる(13)．

(例)「写真の人物」⇒「写真に描かれた人物」
用例：「写真に人物が描かれる」

5. 換喩解析

「漱石を読む」の「漱石」は「漱石が書いた小説」を意味する．そのような省略された情報を「A の B」「C を V する」という形をした用例を組み合わせて用いることで補完する(14, 15)．

(例)「漱石を読む．」⇒「漱石の小説を読む．」
用例：「漱石の小説」「小説を読む」

6. 連体化した節の格解析

名詞と動詞の共起情報を用いて隠れている連体化した節の格を推定する(3)．

(例)「オープンする施設」⇒ 格関係＝ガ格
用例：「施設がオープンする」

7. 質問応答システム

質問応答では疑問詞の部分が省略しておりこの部分を補完する問題であると考えられる．この場合よくにた文を集めてその文の疑問詞にあたる部分を解答として出力する(7, 8, 9)．

(例)「日本の首都はどこですか」⇒ 解答＝東京
用例：「日本の首都は東京です」

ところで，これらの研究は教師あり機械学習の問題に落すことができる．例えば，動詞の省略補完の場合だと，

用例：「そんなにうまくいくとは思えない．」

を

文脈：そんなにうまくいくとは　分類先：思えない

という教師信号と考えると，文脈から分類先を学習する教師あり機械学習の問題になっている．指示詞・代名詞・ゼロ代名詞照応解析の場合だと上にあげた用例を

文脈：「を食べる」　分類先：「りんご」

という教師信号に，また，間接照応解析の場合だと

文脈：「の屋根」　分類先：「家」

という教師信号に，また，「A の B」の意味解析の場合だと

文脈：「写真」「人物」　分類先：「描かれる」

という教師信号に，また，換喩解析ならば

文脈：「漱石の」　分類先：「小説」
文脈：「を読む」　分類先：「小説」

という教師信号に，連体化における格解析の場合は

文脈：「施設」「オープンする」　分類先：ガ格

という教師信号に，質問応答システムの場合は

文脈：「日本の首都は」　分類先：「東京」
文脈：「の首都は東京です」　分類先：「日本」

という教師信号にとらえれば機械学習の問題になっている．以上にあげたもの以外にも省略解析と解釈できる問題は同様に解析対象用のタグのついていないコーパスを機械学習の教師信号にすることができる．また，単純な省略補完だけでなく，「オープンする施設」を「施設がオープンする」ととらえる格解析のように，言葉を少し補いながら言い換えて解釈するような問題も同様に解析対象用のタグのついていないコーパスを機械学習の教師信号にすることができる[3]．

機械学習の問題に落すことには以下のような価値がある．

- 機械学習手法にはさまざまな高度な手法が提案されている．機械学習の問題に落すことで，そのときに応じた最もよい機械学習手法を選択して問題をとくことができる．
- 機械学習手法では，かなり自由に，解析に用いる情報

[3] 意味解釈の問題は，おおかたの場合言い換えた文によってその答えを表現する．このため，言葉を少し補いながら言い換えて解釈するような問題一般も，手法の適用範囲に含める本提案手法の適用範囲の広さがいかに大きなものであるかが理解できるであろう．

を定義でき，広範な情報を利用できる．このため，多くの情報を利用でき，精度があがりやすい．

これで解析対象用のタグのついていないコーパスを用いた機械学習が，省略補完処理で使えるようになった．本稿ではこの手法を教師信号を本来の教師信号でないところから持ってきたという意味で**教師信号借用型機械学習手法**とよぶ．

この借用型の教師信号は，教師信号の形になっているため，本来の解析対象用のタグのついたコーパスのデータからとった教師信号 (本稿では**非借用型教師信号**とよぶ) と同時に併用して用いることができる．この場合は本来の教師信号と，解析対象用のタグのついていない大規模なデータを扱える借用した教師信号の両方を用いるということでかなり強力である．この手法のことを本稿では**併用型教師あり機械学習**とよぶことにする．

ところで，照応解析などの場合は指示先が本文にあり用例だけの情報だけで指示先を特定するのは困難であり，借用した教師信号だけを用いて解析を行なうことはできない．このような場合は非借用型教師信号も用いる併用型を用いればよい．また，格解析でもつねに表層格を補完するのではなく，外の関係など借用信号では扱えない課題もある[4]．その場合は併用型を用いるとよい．本稿では実際に併用型を用いた実験も行なっている．

3 機械学習手法

教師信号借用型機械学習手法，非借用型もしくは併用型機械学習手法を用いるにしても，機械学習を行なうには，なんらかの機械学習手法を使わなければ実現できない．本稿では機械学習手法としては，以下の五つの方法を利用した[5]．

- TiMBL 法 (k 近傍法)
- シンプルベイズ法
- 決定リスト法
- 最大エントロピー法
- サポートベクトルマシン法

本節では紙面の都合上，一番性能の高いサポートベクトルマシン法についてのみ詳しく説明する[6]．

[4] これは外の関係が扱えないのは表層格を用いた文に変形できないためである．しかし，ここで格解析というしばりをなくして言い換えによる文の解釈という立場をとるのならば，外の関係も借用型機械学習で扱える．例えば，外の関係の文「さんまを焼くけむり」は「さんまを焼く時に出るけむり」と言い換えることによって解釈する場合がある．そのような解釈を正解とする問題設定ならば，連体節とその係り先の名詞との間に「時に出る」という表現が省略されていて，それを補完するということにすれば，省略補完の問題となり，本稿の借用型機械学習で扱える問題となる．

[5] 機械学習手法としては，他に C4.5 などの決定木学習を利用する方法があるが，本稿では，種々の問題で決定木学習手法が他の手法に比べて劣っていること[16, 17, 18]，また，本稿で扱う問題は属性の種類の数が多く C4.5 が走るまで属性の数を減らすと精度が落ちるであろうことの二つの理由により，用いていない．

[6] シンプルベイズ法，決定リスト法，最大エントロピー法については文献[19] でも説明しているので，それを参照して欲しい．TiMBL[4] は，Daelemans らが開発したシステムで，類似する k 個の事例でもとめる k 近傍法を用いるものになっている．さらに TiMBL では事例間の類似度はあらかじめ定義しておく必要はなく，素性を要素とした重みつきのベクトルの間の類似度という形で自動的に算出される．また本稿では k=3 を用いその他はデフォルトの設定で利用した．

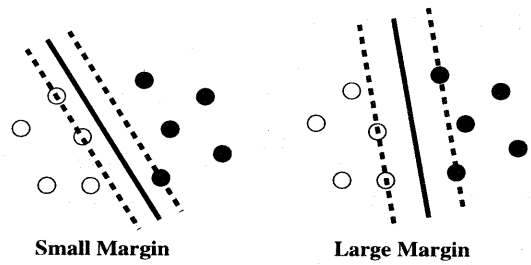

図 1: マージン最大化

サポートベクトルマシン法は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン) が大きいもの (図 1参照[7]) ほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行なう．基本的には上記のとおりであるが，通常，学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる[20, 21]．

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \quad (1)$$

$$b = -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2}$$

$$b_i = \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i (i = 1, ..., l, y_i \in \{1, -1\})$ は学習データの文脈と分類先を意味し，関数 sgn は，

$$sgn(x) = \begin{array}{ll} 1 & (x \geq 0) \\ -1 & (otherwise) \end{array} \quad (2)$$

であり，また，各 α_i は式 (4) と式 (5) の制約のもと式 (3) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \quad (3)$$

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \quad (4)$$

[7] 図の白丸，黒丸は，正例，負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する．

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{5}$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本稿では以下の多項式のものを用いる．

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+1)^d \tag{6}$$

C, d は実験的に設定される定数である．本稿ではすべての実験を通して C を 1 に d を 2 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (1) の和をとっている部分はこの事例のみを用いて計算される．つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない．

サポートベクトルマシン法は分類の数が 2 個のデータを扱うもので，通常これにペアワイズ手法を組み合わせて用いることで，分類の数が 3 個以上のデータを扱うことになる(22)．

ペアワイズ手法とは，N 個の分類を持つデータの場合，異なる二つの分類先のあらゆるペア (N(N-1)/2 個) を作り，各ペアごとにどちらがよいかを 2 値分類器 (ここではサポートベクトルマシン法[8]) で求め，最終的に N(N-1)/2 個の 2 値分類器の分類先の多数決により，分類先を求める方法である．

本稿のサポートベクトルマシン法は，上記のようにサポートベクトルマシン法とペアワイズ手法を組み合わせることによって実現される．

4 問題設定と素性 (解析に用いる情報)

本節では，本稿の格解析の問題設定と素性 (解析に用いる情報) について説明する．つまり，機械学習に用いる文脈 (素性の集合) と分類先を説明する．

格解析を行なう対象は以下のものとした．

- 連体化した節の用言とその係り先の体言の間の関係
- 格助詞のみがつく体言，助詞が一切つかない体言を除く体言が用言にかかる場合のその体言と用言の関係
  (例：「この問題さえ解かれた．」)

分類先としては以下の 7 分類を用いた．

- ガ格，ヲ格，ニ格，デ格，ト格，カラ格．(6 分類)
- その他 (外の関係，格関係にならない主題など)

このとき，受け身の文の場合でも受け身の文型のまま表層格の推定を行なう．例えば，

「解かれた問題」

の場合，「問題が解かれた」なので，ガ格として扱う．受け身を能動態に直して「問題を解く」と解釈して，ヲ格とするようなことはしない．また，外の関係とは関係節の用言と係り先の体言が格関係にならない場合のことをいう．例えば，

「さんまを焼くにおい」

の文の「焼く」と「におい」は格関係が成立しないので，このような文は外の関係と呼ばれる．また，連体化以外で「その他」の分類としたものに以下の「九一年も」がある．

「九一年も出生数が前年より千六百六十人多かった」

これはガガ文としてガ格としてもよいようなものである．このあたりはまだ定義が曖昧になっている．今後分類の修正が必要かもしれない．また，以下の「三度も」のような副詞も「その他」の分類とした．

「過去一年間に三度も首相が代わる」

ここで助詞「も」がなければ解析の対象としない[9]．

文脈としては以下のものを定義した．ただし，体言 n と用言 v の間の格関係を求める場合のものである．

1. 問題が連体節か主題化のものか
   主題化の場合は体言 n についている助詞．
2. 用言 v の品詞
3. 用言 v の単語の基本形
4. 用言 v の単語の分類語彙表の分類番号の 1,2,3,4,5,7 桁までの数字．ただし，分類番号に対して文献の表の変更を行なっている．
5. 用言 v につく助動詞列
   (例：「れる」「させる」)
6. 体言 n の単語
7. 体言 n の単語の分類語彙表の分類番号の 1,2,3,4,5,7 桁までの数字．ただし，分類番号に対して文献の表の変更を行なっている．
8. 用言 v にかかる体言 n 以外の体言の単語列
   ただし，どういった格でかかっているかの情報を AND でつける．
9. 用言 v にかかる体言 n 以外の体言の単語集合の分類語彙表の分類番号の 1,2,3,4,5,7 桁までの数字．ただし，分類番号に対して文献の表の変更を行なっている．
   また，どういった格でかかっているかの情報を AND でつける．
10. 用言 v にかかる体言 n 以外の体言がとっている格
11. 同一文に共起する語

実験は，以上の素性のいくつかを用いて行なわれる．

教師信号借用型の手法を用いる場合は 1 の素性は用いることができない．

5 実験

まず非借用型教師あり学習手法を用いた実験を行なった．データは京大コーパス(23) 中の毎日新聞(24)95 年 1 月 1 日の一日分を用いた．このデータに 4節で定義した問題設定で分類先を付与した．京大コーパスの構文タグが誤っていると判明した部分はデータから除いた．事例数は 1530 個であった．全事例における分類先の出現の分布

[8] 本稿の 2 値分類器としてのサポートベクトルマシンは，工藤氏が作成した TinySVM(21) を利用している．

[9] 助詞の脱落現象の少ない分野のデータならば，助詞が一つもついていなければ副詞と判断してもよいだろうが，助詞の省略が存在するとなると，助詞のついていない体言も係り先の用言と格関係を持つ可能性があるために，それらの体言もすべて解析対象とする必要があるだろう．

表 2: 非借用型教師あり学習手法の精度

説明	省いた素性	TMBL	SB	DL	ME	SVM
全素性を利用		67.12%	53.66%	67.19%	77.91%	79.61%
共起語を利用せず	11 の素性	67.06%	62.88%	67.65%	79.22%	80.85%
共起語・用言 v の品詞情報を用いず	2,11	66.93%	61.44%	67.65%	79.22%	81.05%
共起語・用言 v につく助動詞列情報を用いず	5,11	66.47%	59.80%	67.06%	78.43%	80.85%
共起語・推定の種類の情報を用いず	1,11	66.47%	61.18%	67.39%	78.76%	80.52%
共起語・分類語彙表の分類番号情報を用いず	4,7,911	66.67%	66.73%	68.17%	79.08%	80.65%
共起語・他の格の情報を用いず	8,9,11	63.33%	63.73%	70.33%	79.67%	81.37%
共起語・他の格の情報・動詞の分類語彙表の分類番号を用いず	4,8,9,11	64.05%	70.00%	72.35%	80.46%	82.55%
共起語・分類語彙表の分類番号の情報を用いず	4,7,9	66.80%	65.49%	67.06%	75.03%	77.06%

表 1: 事例の分布

	主題化	連体化
ガ格	526	499
ヲ格	29	46
ニ格	14	45
デ格	16	10
ト格	0	2
カラ格	0	1
その他	75	267
合計	660	870

表 3: 格助詞の再推定問題 (借用型教師あり学習手法の精度)

TMBL	SB	DL	ME	SVM
27.40%	50.22%	54.70%	66.93%	70.25%

表 4: 主題化・連体化現象における表層格復元実験

	TMBL	SB	DL	ME	SVM
併用型	9.85%	71.13%	75.34%	82.24%	87.04%
借用型	10.61%	44.11%	33.42%	51.18%	55.39%
非借用型	86.03%	82.07%	84.68%	86.87%	88.22%

を表 1に示す．ガ格が圧倒的に多くついで連体における外の関係が多いことがわかる．このデータで記事を分割の単位とした 10 分割のクロスバリデーションを用いて実験した．この結果を表 2に示す．表の値は全事例での正解率を意味し，TMBL, SB, DL, ME, SVM はそれぞれ TiMBL 法，シンプルベイズ法，決定リスト法，最大エントロピー法，サポートベクトルマシン法を意味する．さまざまな素性の設定で実験を行なったが用いる素性で精度がかわった．現在の機械学習の手法ではその手法自体で正しく素性を選択する能力はまだ低いと思われる．以降の実験では，精度のもっともよかった 4,8,9,11の素性を省く素性の設定を用いる．また，機械学習の手法としてはサポートベクトルマシンがもっともよかった．

次に借用型教師あり学習手法を用いた実験を行なった．借用する教師データ用の用例は京大コーパス中の毎日新聞 95 年 1 月 1 ～ 17 日の 16 日分 (約 2 万文) を用いた．このデータのうち，体言と用言を係り受け関係を格助詞のみで結んでいるもののみを教師データとした．全事例数は 57,853 個であった．このとき，4節で述べる素性のうち，1の素性は，主題化・連体化していないものからデータをもってくるために，用いることはできない．

表 5: 主題化・連体化現象における表層格復元 (ガヲニデの四つの格のそれぞれでの精度の平均で評価)

	TMBL	SB	DL	ME	SVM
併用型	21.90%	45.74%	37.33%	51.35%	62.16%
借用型	28.63%	48.31%	31.37%	54.40%	59.11%
非借用型	24.93%	52.79%	31.95%	42.90%	44.96%

ここではまず借用型教師あり学習手法の基本性能を調べるために，表層格の再推定という問題を解くことにする．これは文中の表層格を消してそれをもう一度推定できるかと試すものである[10]．この問題を対象として，さきほどの借用教師信号 (57,853 個) で記事ごとの 10 分割のクロスバリデーションを用いて実験した．結果を表 3に示す．表のようにサポートベクトルマシン法がもっともよく，7 割の精度を得た．このことから，文生成における助詞の生成が少なくともこの精度でできることがわかる[11]．また，一般的な助詞脱落補完問題がこの精度の周辺で解けるであろうことがわかる．

次に借用型教師あり学習手法を用いて，最初に用意した主題化・連体化したデータで，表層格復元の実験を行なった．ここでは，借用型教師では外の関係などの「その他」の分類を推定することができないので，「その他」の分類の事例を除いて実験を行なった．この結果，評価用のデータの事例数は 1530 から 1188 に減少した．学習にはさきほど集めた借用教師信号 (57,853 個) を用いる．この実験の結果を表 4に示す．また，この実験はガ

[10] この種の考え方を用いた研究は生成の分野[25, 26, 27, 28]ではよくある．解析対象とするタグがふられていないコーパスを用いることができるという点で，省略解析と生成がにていることについては文献[10]ですでに指摘している．

[11] 文生成の場合は深層格などなんらかの格に対する情報を入力としても与えられると思われるため，本当の精度はここでのものよりは高くなると思われる．また，この深層格の情報も用いた学習は本稿でいう併用型に近いものとなる．

ヲニデの四つの格のそれぞれでの精度の平均でも評価した．この結果を表5に示す．ここでは比較のために，この1188事例を学習に用いた非借用型教師あり学習手法による結果も示す．また，この1188個の非借用教師信号と，57,853個の借用教師信号の両方を併用する併用型教師あり学習手法による結果も示す．ただし，これらの実験では記事を単位とする10分割のクロスバリデーションを行ない，解析対象の事例と同じ記事の借用教師信号と非借用教師信号は用いないようにしている．

結果より以下のことがわかる．

- まず，表4の全事例での精度で議論する．手法としてはサポートベクトルマシン法が一般的にもっともよい．以降の議論ではサポートベクトルマシン法の結果のみを使う．
- 借用型での精度は55.39%であった．主な格の出現がガヲニデの四つであったので，ランダムな選択の場合精度は25%であるのでこれよりはよい結果となる．借用した教師信号を用いた場合の精度としてはよいものと思われる．
- 併用型，借用型，非借用型の中では非借用型がもっともよかった．借用したデータは実際の問題とは異なる性質を持っている可能性があり，これを少しでも用いると精度が低下する可能性は十分ありうる．この結果はこのことを反映したものと考えられる．
- この実験の評価に用いたデータは1188事例でそのうちガ格は1025であり，ガ格の出現確率は86.28%であり，なにも考えずすべてガ格であると判定しただけでも86.28%の精度を得る．しかしこれでは他の格の解析精度は0%でありこの結果は利用先によってはなにも役に立たない可能性がある．そこで，表5に示したガヲニデの四つの格のそれぞれでの精度の平均での評価も行なっている．この評価では最も頻度の高い分類に決め打ちにする手法だと精度は25%となる．併用型，借用型，非借用型ともにこの精度よりは高いことがわかる．
- 平均での評価では，精度の順は併用型，借用型，非借用型となっている．非借用型は，問題に密接な教師信号を用いるために高い精度を得やすいとはいえ，本稿のように事例数が少ない場合は他の手法よりも精度が低くなる場合があることがわかる．
- 併用型の手法は表4の評価で借用型に1%負けているだけで，表5の平均での評価では圧倒的によく両方の評価基準ともによい結果を得ている．

以上のことから，借用型がランダムな選択より有効でありかつ分類先の平均を評価基準とすると非借用型より有効であること，また，併用型が複数の評価基準で安定してよい結果を示したことがわかる．このため，借用型と併用型の有効性が示されたことになる．

次に外の関係などの「その他」の分類も含めた実験を行なった．これは評価用のデータ(1530事例)すべてを用いた．この実験は併用型と非借用型の二つで実験した．借用教師信号だけでは「その他」の分類を特定できないため，借用型は用いなかった．この結果を表6に示す．また，この実験はガヲニデ,“その他”の五つの分類先のそれぞれでの精度の平均でも評価した．この結果を表7に示す．結果はサポートベクトルマシンの精度がもっともよくまた，併用型は全事例での精度で1%だけ非借用より低いだけで平均精度では併用型の方が圧倒的に高かった．

表6: 格解析全般での実験

	TMBL	SB	DL	ME	SVM
併用型	8.95%	65.42%	51.50%	71.63%	81.57%
非借用型	64.05%	70.00%	72.35%	80.46%	82.55%

表7: 格解析全般での実験(ガヲニデ,“その他”の五つの分類先のそれぞれでの精度の平均で評価)

	TMBL	SB	DL	ME	SVM
併用型	24.35%	43.57%	29.28%	46.57%	56.93%
非借用型	22.90%	50.23%	33.67%	46.29%	47.03%

最後に格解析の他の研究と比較しながら議論する．

- 黒橋らの既存の格フレームを用いる方法だが，河原らの格フレームを構築する研究があることからわかるように，既存の格フレームでは語彙数，格フレームの緻密さなどに問題がある．
- 次に河原らの格フレームを構築してそれを用いて格解析する方法だが，これはコーパスから格フレームを学習しているともとらえることができ，本稿の借用型機械学習とよくにたことをしていることになる．しかし，実際の解析では併用型の方がよく，対象とする問題である主題化・連体化に対するなんらかの知見・情報を用いた方がよいことがわかる．格フレームを用いる方法では主題化・連体化に関する情報を扱うことが困難である．実際，河原らの論文にも「ガガ文」「外の関係」の場合の取り扱いを今後の課題としている．これに対し，併用型機械学習である場合，「外の関係」，もしくは「ガガ文」用の分類先を定義することで非借用教師信号として扱うことができる．また，コーパスからの格フレームの学習についても，機械学習手法を用いると広範な情報を容易に用いることができるため，借用型機械学習の方がよいと思われる．
- 次に阿部川らの研究だが借用信号だけを利用している点に問題がある[12]．Baldwinの研究は逆に借用信号を用いていないという問題がある．また，双方ともに格

[12] さらには，阿部川らの方法ではk=5の類似度尺度固定のk近傍法を用いているようなものだが，この手法では広範な情報を解析に用いることが困難で例えば「られる」などの受け身の情報を扱うことが困難である．このため，「解かれた問題」などでは，「問題が解かれる」のガ格でなく，「問題を解く」のヲ格のように受け身を戻して格解析をしてしまうという問題もある．

解析全般を扱っておらず，連体化しか扱っていない．現在の言語処理では形態素，構文がルールベース手法，機械学習手法によりとりあえず現時点ではもうこれ以上はそれほど改善しないであろうというところまできた．今後は構文の次の意味解析が主な研究分野になると思われる．その意味で格解析の重要性は高い．このため，格解析全般を扱っている本研究は重要なものとなっている[13]．

6 おわりに

本稿では日本語格解析全般の問題を機械学習手法により扱った．このとき，新たに教師信号借用型機械学習手法や併用型機械学習手法を提案した．また，これらの手法を省略解析に関係する広範な問題で用いることができることも述べた．また，実験の結果，教師信号借用型機械学習手法がランダムな解析よりも精度が高くまた分類先ごとの精度を平均した精度では非借用型手法よりも精度が高いことがわかった．また，併用型機械学習手法が全事例での精度だけでなく，分類先ごとの精度を平均した精度でも高く複数の評価基準において安定して高い精度を得ることを確認した．これらのことから，教師信号借用型機械学習手法と併用型機械学習手法の有効性が確認された．

本稿の格解析の研究は，連体化や主題化した文をもとに戻す言い換えの研究[29, 30, 31]にも役に立つ．例えば「りんごは食べた」の「は」がヲ格とわかると「りんごを食べた」に言い換えることができるとわかる．本稿の次の課題としては受け身 / 使役における格の交替現象を考えている．例えば「問題が解かれた」は「問題を解いた」となる．このときに変更される格を推定するのである．これができると受け身 / 使役から能動文への言い換えも扱えることになる．今後は，このような言い換えにかかわる意味解析を中心に研究をすすめる予定である．

[13] 精度の比較としては Baldwin が連体の解析で 88% を出しており，これはわれわれの実験を連体に限ったものよりも高い．しかし，データの違い，素性の違いなどがあるため比較は困難である．本稿は全般的な大枠を議論することに主眼をおいておりどのような素性を用いればよいかについては扱わない．ここでは，よりよいより多くのデータ，よりよいより多くの素性，よりよい手法を用いるとよいだろうということを記述するにとどめる．

参考文献

(1) Sadao Kurohashi and Makoto Nagao, A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary, *IEICE Transactions on Information and Systems*, Vol. E77–D, No. 2, (1994), pp. 227–239.

(2) 河原大輔, 黒橋禎夫, 用言と直前の格要素の組を単位とする格フレームの自動獲得, 情報処理学会 自然言語処理研究会 2000-NL-140-18, (2000).

(3) 阿部川武, 白井清昭, 田中穂積, 徳永健伸, 統計情報を利用した日本語連体修飾節の解析, 言語処理学会 年次大会, (2001), pp. 269–272.

(4) Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, Timbl: Tilburg memory based learner version 3.0 reference guide, Technical report, (1995), ILK Tchnical Report-ILK 00-01 (http://ilk.kub.nl/ ilk/papers/ilk0001.ps.gz).

(5) Timothy Baldwin, Making lexical sense of japanese-english machine translation: A disambiguation extravaganza, Technical report, (Tokyo Institute of Technology, 2001), Technical Report, ISSN 0918-2802.

(6) 村田真樹, 長尾真, 日本語文章における表層表現と用例を用いた動詞の省略の補完, 言語処理学会誌, Vol. 5, No. 1, (1998).

(7) Masaki Murata, Masao Utiyama, and Hitoshi Isahara, Question answering system using syntactic information, (1999), http://xxx.lanl.gov/abs/ cs.CL/9911006.

(8) 村田真樹, 内山将夫, 井佐原均, 類似度に基づく推論を用いた質問応答システム, 自然言語処理研究会 2000-NL-135, (2000), pp. 181–188.

(9) 村田真樹, 内山将夫, 井佐原均, 質問応答システムを用いた情報抽出, 言語処理学会第 6 回年次大会ワークショップ論文集, (2000), pp. 33–40.

(10) 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).

(11) 村田真樹, 長尾真, 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定, 言語処理学会誌, Vol. 4, No. 1, (1997).

(12) 村田真樹, 長尾真, 意味的制約を用いた日本語名詞における間接照応解析, 言語処理学会誌, Vol. 4, No. 2, (1997).

(13) 田中省作, 富浦洋一, 日高達, 統計的手法を用いた名詞句「NP の NP」の意味関係の抽出, 言語理解とコミュニケーション研究会 NLC98-4, (1998), pp. 23–30.

(14) 村田真樹, 山本専, 黒橋禎夫, 井佐原均, 長尾真, 名詞句「a の b」「ab」の用例を利用した換喩解析, 人工知能学会誌, Vol. 15, No. 3, (2000).

(15) 内山将夫, 村田真樹, 馬青, 内元清貴, 井佐原均, 統計的手法による換喩の解釈, 言語処理学会誌, Vol. 7, No. 2, (2000).

(16) 村田真樹, 内元清貴, 馬青, 井佐原均, 学習による文節まとめあげ — 決定木学習, 最大エントロピー法, 用例ベースによる手法と排反な規則を用いる新手法の比較 —, 情報処理学会 自然言語処理研究会 NL128-4, (1998).

(17) 村田真樹, 内元清貴, 馬青, 井佐原均, 排反な規則を用いた文節まとめあげ, 情報処理学会論文誌, Vol. 41, No. 1, (2000), pp. 59–69.

(18) 平博順, 春野雅彦, Support vector machine によるテキスト分類における属性選択, 情報処理学会論文誌, Vol. 41, No. 4, (2000), pp. 1113–1123.

(19) 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均, 種々の機械学習手法を用いた多義解消実験, 電子情報通信学会言語理解とコミュニケーション研究会 NLC2001-2, (2001).

(20) Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, (Cambridge University Press, 2000).

(21) Taku Kudoh, Tinysvm: Support vector machines, (http://cl.aist-nara.ac.jp/ taku-ku//software/TinySVM/index.html, 2000).

(22) 工藤拓, 松本裕治, Support vector machine を用いた chunk 同定, 自然言語処理研究会 2000-NL-140, (2000).

(23) 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, (1997), pp. 115–118.

(24) 毎日新聞社, 毎日新聞 1991-1998, (1998).

(25) Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, Vol. 19, No. 2, (1993), pp. 263–311.

(26) Adwait Ratnaparkhi, Trainable methods for surface natural language generation, *Proceedings of the ANLP-NAACL 2000*, (2000).

(27) Francis Bond Minnen, Guido and Ann Copestake, Memory-based learning for article generation, *the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop (CoNLL-2000 and LLL-2000)*, (2000), pp. 43–48.

(28) 内元清貴, 村田真樹, 馬青, 関根聡, 井佐原均, コーパスからの語順の学習, 言語処理学会誌, (2000).

(29) 近藤恵子, 佐藤理史, 奥村学, 格変換による単文の言い換え, 情報処理学会自然言語処理研究会 2000-NL-135-16, (2000).

(30) 村田真樹, 井佐原均, 同義テキストの照合に基づくパラフレーズに関する知識の自動獲得, 自然言語処理研究会 2001-NL-142, (2001).

(31) 村田真樹, 井佐原均, 言い換えの統一的モデル — 尺度に基づく変形の利用 —, 言語処理学会第 7 回年次大会ワークショップ論文集, (2001).