概要

本研究では, 助詞「も」に対して, 教師あり機械学習を用いることにより, いくつかの場合の助詞「も」の使い分けを行い, 助詞「も」に関わる知見を得ることを目指す.

助詞「も」の使い分けを分析しそれに関する知見を得ることは, 以下の二つのことに役立つと思われる. 一つは, 助詞「も」に関わる知見により助詞「も」の誤った使用の検出技術の構築につながる. もう一つは, 日本語文法の課題の一つである助詞「も」に関わる知見を増やすことにより, 日本語文法に関わる研究の推進につながる.

本研究では「も」の使い分けを７割から８割の正解率で行うことができた. また教師あり機械学習に用いた素性を分析することにより, 「も」の文中での使用における特徴を得ることができた.

目 次

図 目 次

表 目 次

第1章　はじめに

既存の日本語文は, 作文のための手本ととらえることができる. 既存の大量の日本語文を, 教師あり機械学習で分析することにより, 日本語文法 [1][2][3][4][5] に関わる様々な知見を得ることができる. 例えば, 林ら [6] は日本語文章における文の順序を教師あり機械学習を用いて研究することにより, 文の順序に関わる知見を得ている. 三浦ら [7][7] は日本語の格助詞の使い分けを教師あり機械学習を用いて研究することにより, 格助詞の使い分けに関わる知見を得ている.

本研究では, 三浦らの研究 [7][8] で取り上げられなかった助詞「も」に対して, 教師あり機械学習を用いることにより, いくつかの場合の助詞「も」の使い分けを行い, 助詞「も」に関わる知見を得ることを目指す.

助詞「も」の使い分けを分析しそれに関する知見を得ることは, 以下の二つのことに役立つと思われる. 一つは, 助詞「も」に関わる知見により助詞「も」の誤った使用の検出技術の構築につながる. もう一つは, 日本語文法の課題の一つである助詞「も」に関わる知見を増やすことにより, 日本語文法に関わる研究の推進につながる.

本研究で強調したいことをあらかじめまとめておくと以下のようになる.

1. 本研究では教師あり機械学習を用いて助詞「も」の使い分け問題に取り組んだ.
2. 「も」の使い分け問題ではデータ数を拡張することにより, 7 割から 8 割の正解率を得ることができた.
3. 文脈素性を使わない実験を行うことで, 「も」の使い分けに文脈素性が必要であることがわかった.
4. 教師あり機械学習に用いた素性の分析により「も」の文中での使用における特徴を得ることができた.

本論文の構成は以下の通りである. 第 2 章では, 本研究に関連する研究としてどのような研究が行われてきたかを記述し, その研究と本研究との関連を説明する. 第 3 章では, 本研究が扱う問題の設定とそれを解決するために提案した手法について説明を行う. 第

４章では, 本研究で行った実験の実験結果とその考察を記述する. さらに, 追加で行った二つの実験の方法とその結果を記述する. 第５章ではまとめを行う.

第2章　関連研究

関連研究としては以下のものがある.

三浦ら [7][8] は教師あり機械学習を用いて副助詞「は」と格助詞「が」の使い分けや, 格助詞「に」「へ」「を」「で」の使い分けを行った. 「は」「が」の使い分けを 0.76 の正解率で行うことができ, また, その使い分けに役立つ表現を獲得した.

林ら [6] は教師あり機械学習を用いて文の順序推定を行った. 比較手法として用いた確率モデルに基づく従来手法での正解率が 0.58 から 0.61 であるのに対し, 林らの提案手法での正解率は 0.72 から 0.77 と高い値を得ることができた. また, 文の順序推定に役立つ表現を獲得した.

木曽ら [9] はとりたて詞において代表的である「も」に特に焦点を当て, 「も」を含む文を計算機で取り扱い可能となるよう関係意味論に基づき「も」の定式化を行った. 「太郎も知っている」という文が用いられた文脈において, この文で生じている影の意味「太郎と異なる X も知っている」の X に相当するオブジェクトの存在を調べた.

また, 文系の日本語学において助詞の研究は数多くなされている [1][2][3][4][5].

沼田 [1] はとりたて詞の「も」とそうでない「も」の用法についてまとめた. とりたて詞の「も」は「単独他者肯定」「意外」「不定他者肯定」の「も」と分け, とりたて詞でない「も」は語中あるいは慣用句中の「も」, 慣用的強調の「も」, 形式副詞の「も」と分けている.

助詞「も」は「他者の存在を暗示する」とされる. 他者とは木曽らの研究での例文「太郎と異なる X も知っている」の X に相当するオブジェクトのことである. 岡野 [2] は特定の他者の想定が意味を持たないため, あるいは特定の他者の想定が難しいために周辺的とされる４つの「も」の用法を示した.

1. 君もしつこいな.

2. お前も大きくなったな.

3. イチローもエラーをすることがあるかぁ.

4. 夜もふけて参りました.

岡野はこれらの「も」の文を論理学的に解釈し, 周辺的とされる要因に関して考察した.

中俣 [3] は日本語のとりたて詞と並列助詞の接点について考察した. とりたて詞によって作られる集合と, 並列助詞によって作られる集合には何か違いがあるのかを, 「も」と「とか」が使われている文を分析し, それらが使われる意味的, 語用論的な条件を探った.

第3章　提案手法

3.1　「も」の使い分け

本論文での「も」の使い分けでは, 体言の末尾付近に用いる助詞「も」であり, 係り先の用言に動詞, 形容詞, 判定詞のどれかを含むもののみを扱う.

「も」には, 前出の体言に対比的に用いる体言を示すもの, 強調を示すもの, 文中の他の語の存在に呼応して用いるものなど, 様々の用法がある. 既存の日本語文を学習データに用いた教師あり機械学習で「も」の使い分けの問題を扱ってこの問題を分析することにより, これらの「も」に関わる様々な用法に関わる知見を取得することを目指す.

3.2 問題設定

本論文では, 以下の課題を想定する.

文章とともに体言の箇所が与えられ, その体言において「も」を使うべきか否かを推定する. 課題の文章は既存の文章を用い, 「も」を使うかどうかをわからなくした状態でその文章を与えて, 「も」を使うかどうかを推定する. 「も」の使用についての推定結果が, 元の文章での「も」の実際の使用と同じであれば正解と判定する.

具体的にどういう箇所で上記推定をするとよいかを調べるために, 助詞の「も」の直前の格助詞の個数を計数した. また, 格助詞を伴わない助詞「も」についても, 格の個数を計数した. 京大コーパス 3.0 の毎日新聞の 1995 年 1 月 1 日から 9 日までの記事での出現を調べたところ, 表 3.1 に示すようになった. 格助詞の「に」の後に「も」が生じることが多いことがわかった. また, 京大コーパス 4.0 の毎日新聞の 1995 年 1 月 1 日から 7 日までの記事で格助詞を伴わない助詞「も」の格を調べたところ, 表 3.2 に示すようにガ格とヲ格が多いことがわかった.

そこで, 二格, ガ格, ヲ格の箇所に, 「も」をつけるべきかどうかを判定することとする.

二格の箇所での課題では以下のものを考える. 文中の「に」「にも」である箇所を抜き出し, その箇所が「に」「にも」のどちらであるかはわからないようにして, 元の文章でその箇所が「に」「にも」のいずれであるかを推定する. この課題を「に」「にも」の使い分け問題と呼ぶことにする.

ガ格, ヲ格の箇所では, 「も」を使う際格助詞を使わずに単に「も」だけを使う場合が多い. 例えば, 「太郎 も 来た」という場合の「も」はガ格 (太郎 が 来た) であり, 「本 も 持っている」という場合の「も」はヲ格 (本 を 持っている) である. 以下, これらの場合の「も」をそれぞれ「がも」「をも」[1] と表現する.

ガ格, ヲ格の箇所での課題では以下のものを考える. 文中の「が」「がも」(または「を」「をも」) である箇所を抜き出し, その箇所が「が」「がも」(または「を」「をも」) のどちらであるかはわからないようにして, 元の文章でその箇所が「が」「がも」(または「を」「をも」) のいずれであるかを推定する. この課題を「が」「がも」(または「を」「をも」) の使い分け問題と呼ぶことにする.

実際の文では, 「がも」「をも」は「も」と記載されているだけで, 「がも」なのか「をも」なのかはわからない. しかし, ここでは「も」を使うか否かのみを推定するととして, その体言の格が何であるかはわかっているものとする. そこで, 「が」か「がも」かの推

[1] 「藁をも掴む」などのように「も」が使われていても「を」も表記する場合がある. しかしこのような事例は少数であるため, 本論文では扱わない.

定, または, 「を」か「をも」かの推定を扱うものである.

本論文では, 「に」「にも」, 「が」「がも」, 「を」「をも」の３つの使い分け問題を扱う.

以下に, 「が」と「がも」の使い分け問題の例を示す.

> 太郎と花子は本屋へ行きました. その時太郎はお金を沢山持っていました. また, 花子 も 同様にお金を沢山持っていました.

上記の文章で一番最後の文の「花子も」の助詞「も」を推定箇所とした場合, 一番最後の文は「また, 花子 X 同様にお金を沢山持っていました. 」とする. この文の X の部分に入る助詞として, 「が」と「がも」のどちらを使用するのが元の文通りであるかの推定を行う.

表 3.1: 「も」の前の格助詞の出現回数

格助詞	出現回数
に	199
と	55
から	21
より	15
を	4
まで	1
へ	1

表 3.2: 格助詞を伴わない「も」の格の出現回数

格	出現回数
ガ格	458
ヲ格	79
ニ格	4
ト格	4
カラ格	2
デ格	0
ヘ格	0
マデ格	0
ヨリ格	0

3.3 提案手法

本論文での提案手法では, 「に」「にも」, 「が」「がも」, 「を」「をも」の３つの使い分け問題に対して教師あり機械学習を利用する. 機械学習には, 認識性能が優れている SVM を実装している TinySVM[10] を使用する. カーネル関数には 2 次の多項式カーネルを利用する.

機械学習で利用する素性は村田らの研究 [11] を参考にして表 3.3 のものを用いる. 分類語彙表 [12] を利用する素性は, 村田らの手法 [11] を利用し素性化する.

3.3.1 教師あり機械学習

本研究では提案手法として教師あり機械学習を利用する. 以下に図 3.1 を用いて教師あり機械学習の手順を説明する. 図 3.1 は「に」「にも」の使い分けの推定を行っている.

まず, 機械が学習データの内容を学習する. 図 3.1 の例文「図書館に行きました. また, 体育館**にも**行きました. 」からは

- 「また」を使うと「にも」になりやすいこと
- 前方の動詞と同じ動詞だと「にも」になりやすいこと

などを学習する.

次に推定を行いたい箇所を含む文を, 推定箇所を元が何かわからないようにして入力する. 図 3.1 では「また, おばあさん **X** 同じものを買います」の X の部分を推定したいとする.

学習結果から入力の X の部分に元々入っていたのは「に」と「にも」のどちらであるかの推定を行う. 図 3.1 では学習結果で, 「また」を使うと「にも」になりやすいこと, 前方の動詞と同じ動詞だと「にも」になりやすいことなどを学習していることにより, X の部分は「にも」であるという推定を行う.

表 3.3: 使用する素性

番号	素性
1	述部における名詞, 動詞, 形容詞, 指示詞の単語の連続
2	述部における最初の名詞, 動詞, 形容詞, 指示詞
3	２の単語の品詞
4	２の分類語彙表の分類番号
5	述部の文節内の２より後続の単語
6	述部の係り先の体言の文節の自立語の連続, 存在, 最後の自立語, その品詞と分類番号
7	体言の文節の自立語の連続, 存在, 最後の自立語, その品詞と分類番号
8	述部にかかる体言以外の体言の文節の自立語の連続, 存在, 最後の自立語, その品詞と分類番号
9	解析対象の助詞の直前の単語, その品詞
10	解析対象の助詞の直後の単語, その品詞
12	文内の単語, その分類語彙表の分類番号
13	解析対象の文内の解析対象の文節以外の文節にある助詞
14	解析対象の文節内の名詞が全て前方に存在しているか
15	解析対象の文節内の名詞のどれかが前方に存在しているか
16	解析対象の文節の係り先の文節内の名詞が全て前方に存在しているか
17	解析対象の文節の係り先の文節内の名詞のどれかが前方に存在しているか
18	解析対象の文節の係り先の文節内の動詞が全て前方に存在しているか
19	解析対象の文節の係り先の文節内の動詞のどれかが前方に存在しているか
20	解析対象の文節の係り先の文節内の形容詞が全て前方に存在しているか
21	解析対象の文節の係り先の文節内の形容詞のどれかが前方に存在しているか
22	解析対象の文節内の名詞が全て前方の同一の文節内で「も」と同時に存在しているか
23	解析対象の文節内の名詞のどれかが前方の同一の文節内で「も」と同時に存在しているか

3.3.2 SVM(サポートベクトルマシン法)

本研究では村田の研究 [13] を参考に教師あり機械学習に SVM(サポートベクトルマシン法) を利用する.

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 3.2 参照[2]) 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 [14][10].

$$\begin{aligned} f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \\ b &= -\frac{max_{i,y_i=-1}b_i + min_{i,y_i=1}b_i}{2} \\ b_i &= \sum_{j=1}^{l}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i) \end{aligned} \tag{3.1}$$

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,

$$\begin{aligned} sgn(x) = \quad & 1 \quad (x\geq 0) \\ & -1 \quad (otherwise) \end{aligned} \tag{3.2}$$

であり，また，各 α_i は式 (3.4) と式 (3.5) の制約のもと式 (3.3) の $L(\alpha)$ を最大にする場合のものである.

$$L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x_i},\mathbf{x_j}) \tag{3.3}$$

$$0\leq\alpha_i\leq C\ (i=1,...,l) \tag{3.4}$$

[2]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.

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.5}$$

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．

$$K(\mathbf{x}, \mathbf{y}) \quad = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.6}$$

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,d はそれぞれ 1 と 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 (3.1) 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 [15].

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[3]) 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．ただし, 本論文での実験では 2 値分類しか行っていないため, ペアワイズ手法は利用していない.

[3]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[10] を利用している．

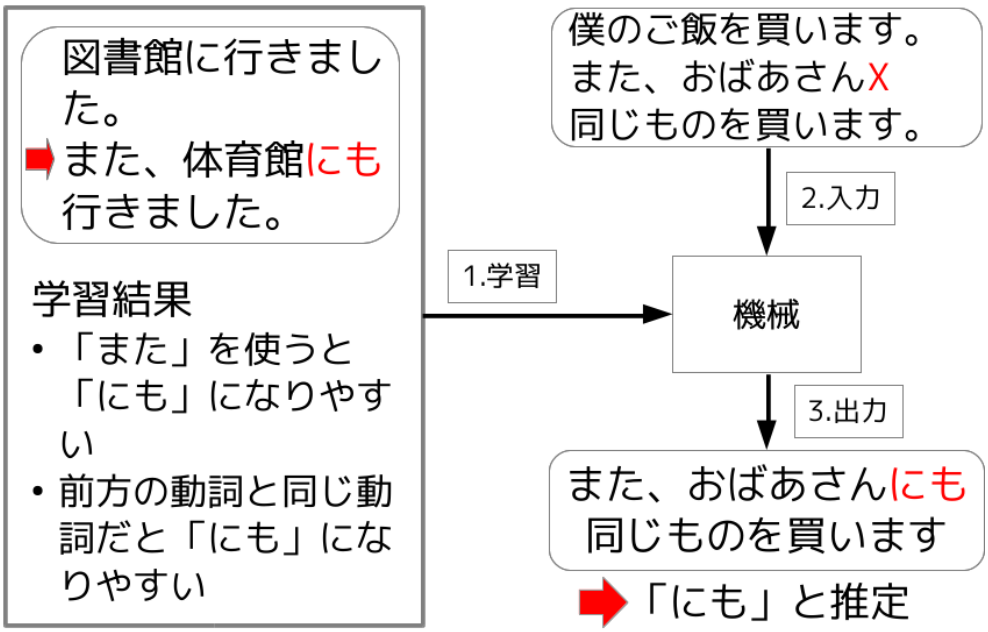

図 3.1: 教師あり機械学習による推定

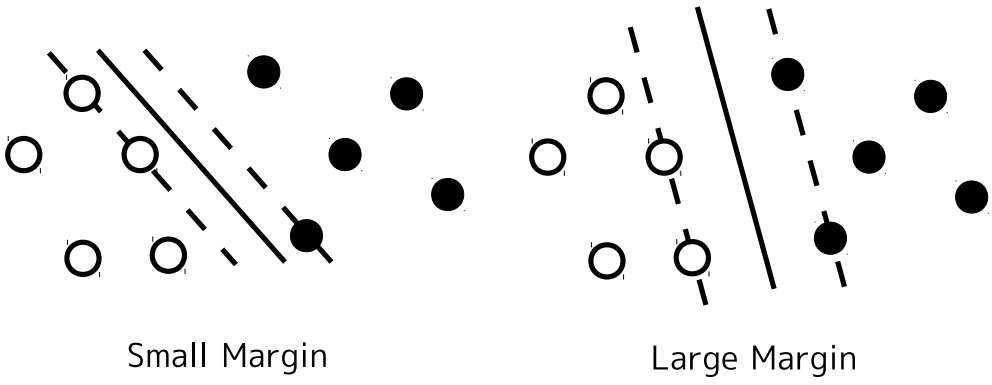

図 3.2: マージン最大化

第4章　実験

4.1　実験データ

「に」「にも」の使い分けの実験において, 学習データは京大コーパス 3.0 の毎日新聞 1995 年 1 月 1 日から 9 日 (2 日は休刊で除く) の記事, テストデータは京大コーパス 3.0 の毎日新聞 1995 年 1 月 10 日から 17 日までの記事を使用する. 「が」「がも」は学習データとして, 京大コーパス 4.0 の 1 月 1 日から 5 日 (2 日は休刊で除く), テストデータとして京大コーパス 4.0 の 1 月 6 日と 7 日を使用する. 「を」「をも」は学習データとして, 京大コーパス 4.0 の 1 月 1 日から 4 日 (2 日は休刊で除く), テストデータとして京大コーパス 4.0 の 1 月 5 日から 7 日を使用する. 京大コーパスでは, 「がも」であるか「をも」であるかの情報も付与されており, その情報を利用して学習データ, テストデータを作成する. それぞれから「に」と「にも」, 「が」と「がも」, 「を」と「をも」を含む文を獲得し, 実験を行う. それぞれのデータ数は表 4.1, 出現確率は表 4.2 である.

ただし, 機械学習法ではデータ数に偏りがある場合, 正しく動作しないことがある. 今回はどの場合もデータ数に差があるため, 学習データはデータ数が多い方をランダムにデータ数が少ない方の数だけ抽出して, データ数の偏りをなくしたものを使用する. () の数字はデータ数を揃えたときの数である.

表 4.1: データ数

使い分け問題に/にも	全データ数	「に」の数	「にも」の数
学習データ	5698　(338)	5529　(169)	169
テストデータ	7278	7045	233
使い分け問題が/がも	全データ数	「が」の数	「がも」の数
学習データ	2235　(562)	1954　(281)	281
テストデータ	551	480	71
使い分け問題を/をも	全データ数	「を」の数	「をも」の数
学習データ	2011　(80)	1971　(40)	40
テストデータ	1669	1641	28

4.2 実験結果

機械学習を利用して「に」「にも」, 「が」「がも」, 「を」「をも」の使い分けの推定を行った. 提案手法の他に, 全て「に」, 「にも」, 「が」, 「がも」, 「を」, 「をも」を推定先とするベースライン手法による分類推定も行った. 各分類での正解率, マクロ平均[1] (各分類先での正解率の平均) を表 4.3 から表 4.5 に示す.

表からわかるように, 提案手法は, 「に」「にも」, 「が」「がも」, 「を」「をも」のすべてのマクロ平均において, ベースライン手法の 0.5 よりも高い値を得ることができた. 「に」「にも」, 「が」「がも」の使い分けでは 7 割近いマクロ平均を得た. しかし, 「を」「をも」では 0.56 とありあまり良い結果ではなかった. これは「を」「をも」の学習データ数が少ないことが原因として考えられる.

また, 提案手法での再現率, 適合率, F 値での評価を表 4.6 から表 4.8 に示す. これらの評価では再現率は 5 割から 7 割の値となっているが, 適合率と F 値は提案手法の「にも」「がも」「をも」に関して極端に低い値となってしまう.

これは, 分類先の事例数が他の分類先よりも極端に小さい場合に一般的に生じやすい現象である. 事例数が他の分類先よりも極端に小さい分類先は, 高い F 値を取得するのが難しい.

本研究では本来「にも」「がも」「をも」であるものを正しく「にも」「がも」「をも」と推定できた数より, 本来「に」「が」「を」であるものを誤って「にも」「がも」「をも」と推定してしまう数の方が圧倒的に多いため, 適合率の値が非常に低くなってしまう.

[1]全事例での正解率 (マイクロ平均) による評価では, 事例数が大きく異なる分類があるとき, 常に事例数の多い分類先と推定する手法が良いと判定されることがある. その場合残りの分類先の正解率が 0%となり, 良い手法と言えない. このためここでは評価にマクロ平均を用いる. マクロ平均では全分類を適切に推定できているかを判定できる.

表 4.2: 出現確率

使い分け問題に/にも	「に」の出現確率	「にも」の出現確率
学習データ	0.97	0.03
テストデータ	0.97	0.03
使い分け問題が/がも	「が」の出現確率	「がも」の出現確率
学習データ	0.87	0.13
テストデータ	0.87	0.13
使い分け問題を/をも	「を」の出現確率	「をも」の出現確率
学習データ	0.98	0.02
テストデータ	0.98	0.02

表 4.3: 「に」「にも」の分類結果

手法	分類先	正解率	マクロ平均
提案手法	「に」	0.64 (4494/7045)	0.69
	「にも」	0.74 (172/ 233)	
ベースライン手法	「に」	1.00 (7045/7045)	0.50
全て「に」	「にも」	0.00 (0/ 233)	
ベースライン手法	「に」	0.00 (0/7045)	0.50
全て「にも」	「にも」	1.00 (233/ 233)	

表 4.4: 「が」「がも」の分類結果

手法	分類先	正解率	マクロ平均
提案手法	「が」	0.65 (312/ 480)	0.67
	「がも」	0.69 (49/ 71)	
ベースライン手法	「が」	1.00 (480/ 480)	0.50
全て「が」	「がも」	0.00 (0/ 71)	
ベースライン手法	「が」	0.00 (0/ 480)	0.50
全て「がも」	「がも」	1.00 (71/ 71)	

表 4.5: 「を」「をも」の分類結果

手法	分類先	正解率	マクロ平均
提案手法	「を」	0.51 (841/1641)	0.56
	「をも」	0.61 (17/ 28)	
ベースライン手法	「を」	1.00 (1641/1641)	0.50
全て「を」	「をも」	0.00 (0/ 28)	
ベースライン手法	「を」	0.00 (0/1641)	0.50
全て「をも」	「をも」	1.00 (28/ 28)	

表 4.6: 「に」「にも」の分類結果（F 値）

手法	分類先	再現率	適合率	F 値
提案手法	「に」	0.64 (4494/7045)	0.99 (4494/4555)	0.78
	「にも」	0.74 (172/ 233)	0.06 (172/2723)	0.11
ベースライン手法 全て「に」	「に」	1.00 (7045/7045)	0.97 (7045/7278)	0.98
	「にも」	— (0/ 233)	— (0/ 0)	—
ベースライン手法 全て「にも」	「に」	— (0/7045)	— (0/ 0)	—
	「にも」	1.00 (233/ 233)	0.03 (233/7278)	0.06

表 4.7: 「が」「がも」の分類結果（F 値）

手法	分類先	再現率	適合率	F 値
提案手法	「が」	0.65 (312/ 480)	0.93 (312/ 334)	0.77
	「がも」	0.69 (49/ 71)	0.23 (49/ 217)	0.35
ベースライン手法 全て「が」	「が」	1.00 (480/ 480)	0.78 (480/ 551)	0.93
	「がも」	— (0/ 71)	— (0/ 0)	—
ベースライン手法 全て「がも」	「が」	— (0/ 480)	— (0/ 0)	—
	「がも」	1.00 (71/ 71)	0.13 (71/ 551)	0.23

表 4.8: 「を」「をも」の分類結果（F 値）

手法	分類先	再現率	適合率	F 値
提案手法	「を」	0.51 (841/1641)	0.99 (841/ 852)	0.67
	「をも」	0.61 (17/ 28)	0.02 (17/ 817)	0.04
ベースライン手法 全て「を」	「を」	1.00 (1641/1641)	0.98 (1641/1669)	0.99
	「をも」	— (0/ 28)	— (0/ 0)	—
ベースライン手法 全て「をも」	「を」	— (0/1641)	— (0/ 0)	—
	「をも」	1.00 (28/ 28)	0.02 (28/1669)	0.03

4.3 素性分析

「も」の使い分けにおいて, それぞれどのような素性が役に立つのかを明らかにするために, 素性の分析を行う. 素性分析には, 4.1 節で使用した学習データの各データを同数にしないものから得られるものを使用する.

素性が全データでの出現率より偏って多くどちらかの分類先に出現しているかを, 二項検定に基づく片側検定により求め, 有意確率 p 値を求める. 有意水準は「に」「にも」, 「が」「がも」の使い分け問題では 5%, 「を」「をも」では 10%とする. 得られた有用な素性の数は表 4.9, また得られた有用な素性の例を表 4.10 に示す.

表 4.9: 有用な素性の数

使い分け問題	分類先	獲得ルール数
「に」「にも」	「に」	60
	「にも」	52
「が」「がも」	「が」	54
	「がも」	21
「を」「をも」	「を」	18
	「をも」	12

表 4.10: 得られた有用な素性の例

使い分け問題	分類先	素性	p 値
に/にも	「に」	直後単語：対する	0.0149
	「にも」	共単：接続詞：また	0.0027
が/がも	「が」	連体単：こと	0.0001
	「がも」	直後単語：ある	0.0007
を/をも	「を」	格に存在	0.0016
	「をも」	用形列：ない	0.0104

4.3.1 「に」「にも」の分析

表 4.10 の素性に関わる例文を以下に示す.

用全単：対する 漂流する政治**に** 対して, 「官」がますます強大になっているように見えます.

共単：接続詞：また 警察当局によると, この化合物はサリン発生後, 土中に残留する物質で, また, この化合物の一種は, サリン生成の際**にも**できるが, 自然に生成することはないという.

分析の結果, その述部の最初の自立語が「対する」「よる」「つく」などである場合「に」であることが多いことがわかった. これは「に対して」「によって」「について」などが格助詞相当句であるためと思われる. 格助詞相当句とはいくつかの語で構成される句であり, 全体として格助詞に相当する働きをする. またこの他にも, 同じ文中に助詞の「から」という表現がある場合は「に」になりやすいことがわかった.

文中に接続詞の「また」が存在する場合, 「にも」である場合が多いことがわかった. これは「また ～ にも ～」という並列表現が多く使われるためであると思われる. またこの他にも, 述部における最初の自立語が「ある」の場合や, 解析対象の文節内の名詞のどれか, もしくは全てが前方に存在する場合, 「にも」になりやすいことがわかった.

4.3.2 「が」「がも」の分析

表 4.10 の素性に関わる例文を以下に示す.

連体単：こと これは新進党副党首の羽田孜氏を挙げた議員**が**少なからずいた こと が大きな要因.

直後単語：ある 厚生省は「晩婚化による比較的高年齢の独身女性層が数年前から結婚し始め, 出産に結び付いたと思われる. 第二次ベビーブーム世代の結婚がこれに続けば, 少子化傾向がストップする可能性**が** ある」と分析している.

分析の結果, 係り先の文節の最後の自立語が「こと」である場合, 「が」である場合が多いことがわかった. これは形式名詞「こと」を使用することで「名詞相当表現＋格助詞」という形の補足節を作ることができ, その補足節の中の主語の格助詞に「が」が用いられることが多いことが要因として考えられる. またこの他にも, 「外国人の雇用**が**従来通り認められるの か」や「本人**が**新年の辞を発表する か どう か」などの述部の文節内の最初の自立語の後続部分が「か」である場合は「が」である場合が多いことがわかった.

直後の単語が「ある」, 述部の文節内の最初の自立語の後続部分が「ない」の場合, 「がも」であることが多いことがわかった. これは人やものの存在を表す表現である「(場所) ニ格＋ (存在の主体) ガ格＋アル/ナイ」という構文のガ格が助詞「も」であることが多いことが要因として考えられる. また, 否定の事態が当該の対象のすべてについて成り立つことを強調する「疑問語＋も」や, 「１＋助数辞＋も」という構文でガ格の助詞「も」が否定の接辞「ない」と同時に用いられることによると考えられる. またこの他にも, 「将来への不安から海外に移住する人 も いれば、経済の活況にひかれて香港にやってくる中国人や外国人 も 多い」という同じ文の対象の文節以外に「がも」が存在する場合や, 解析対象の文節の係り先の文節内の動詞 (動詞「する」は除く) のどれか, もしくは全てが前方に存在している場合は, 「がも」になりやすいことがわかった.

4.3.3 「を」「をも」の分析

表 4.10 の素性に関わる例文を以下に示す.

格に存在 村山内閣となり, 長い懸案だった政治改革, 税制改革, 被爆者援護法など困難な課題 に 大きな区切り**を**つけることができた.

用形列：ない 同日の全国紙「インクワイアラー」によると、例年この時期になると酔っぱらった軍人らが空に向けて銃を発砲することが多く、警察当局は昨年末、銃を発砲した者は逮捕、解任**も**辞さ ない と警告.

分析の結果, 文内に助詞「に」, 助詞「が」, 助詞「は」が存在する場合「を」になりやすいことがわかった. これは相手側へのものの移動を表す動詞を使用するとき「(主体) ガ格＋ (相手) ニ格＋ (対象) ヲ格＋動詞」や, 使役表現での「ガ格 (使役の主体) ＋ニ格 (動きの主体) ＋ヲ格＋動詞の使役形」という構文が使われるため, 助詞「に」, 助詞「が」, 助詞「は」が「を」と同時に使用される場合が多いことによると考えられる.

述部の最初の自立語の後続部分が「ない」である場合や, 助詞の「や」が解析対象の文節以外の文節に存在する場合は「をも」を使用する場合が多いことがわかった.

4.4 学習データの拡張を行った追加実験

前節の実験結果において性能があまりよくなかった一因に, 学習データの不足が考えられる. そこで前節の学習データの拡張を行い, 性能の向上を目指す.

毎日新聞 91 年の 1 年分の記事を KNP で構文解析を行い, 4.1 節で使用した学習データに追加した. KNP には格解析の機能もついており, 「がも」「をも」の認識も可能である. 学習データ追加後のデータ数はそれぞれ表 4.11, 出現確率は表 4.12 である. また, 学習データ追加後の実験結果はそれぞれ表 4.13 である.

前節の実験結果と比べると, 学習データの拡張によって「に」「にも」, 「を」「をも」において高いマクロ平均を得ることができた. 「に」「にも」ではマクロ平均は 0.80 となった. 前節であまり良い性能が得られなかった「を」「をも」でも, マクロ平均が 0.56 から上昇し 0.71 となった. しかし, 「が」「がも」のマクロ平均は下ってしまう結果となった. これは, 拡張したデータは KNP で作成しており, KNP では誤った処理をする可能性があり誤ったデータを追加で用いたことが影響している可能性がある.

表 4.11: 学習データ拡張後のデータ数

使い分け問題に/にも	全データ数	「に」の数	「にも」の数
学習データ	507318 (28348)	493144 (14174)	14174
テストデータ	7278	7045	233
使い分け問題が/がも	全データ数	「が」の数	「がも」の数
学習データ	426185 (76478)	387946 (38239)	38239
テストデータ	551	480	71
使い分け問題を/をも	全データ数	「を」の数	「をも」の数
学習データ	572437 (10916)	566979 (5458)	5458
テストデータ	1669	1641	28

表 4.12: 学習データ拡張後の出現確率

使い分け問題に/にも	「に」の出現確率	「にも」の出現確率
学習データ	0.97	0.03
テストデータ	0.97	0.03
使い分け問題が/がも	「が」の出現確率	「がも」の出現確率
学習データ	0.91	0.09
テストデータ	0.87	0.13
使い分け問題を/をも	「を」の出現確率	「をも」の出現確率
学習データ	0.99	0.01
テストデータ	0.98	0.02

表 4.13: 学習データ拡張後の「も」の使い分けにおける分類結果

使い分け問題	分類先	正解率	マクロ平均
に/にも	「に」	0.76 (5377/7045)	0.80
	「にも」	0.85 (197/ 233)	
が/がも	「が」	0.79 (378/ 480)	0.65
	「がも」	0.51 (36/ 71)	
を/をも	「を」	0.78 (1282/1641)	0.71
	「をも」	0.64 (18/ 28)	

4.5　文脈素性の削除を行った追加実験

教師あり機械学習で使用した素性のうち, 文脈に関する素性 (表 3.3 の素性番号 14 番から 23 番) がどの程度提案手法の性能に影響しているかを調べる.

実験に使用するデータは 4.1 節のものと同じだが, そのうちの文脈に関する素性を削除して「に」「にも」, 「が」「がも」, 「を」「をも」の使い分けの実験を行った.

その結果を表 4.14 に示す. 4.2 節の実験結果と比べると, 提案手法「がも」の正解率が 0.69 から 0.66 へ下っている. これより本研究の提案手法で用いた文脈に関する素性が「がも」の使い分けにおいて有効であることがわかった.

以下に, 素性に文脈素性を含む場合の実験では正しく「がも」と推定できたが, 文脈素性を削除した場合の実験では「がも」と正しく推定できなかった文を示す. この例文は, 4.1 節の「が」「がも」の使い分け問題における各分類先のデータ数を揃えた学習データを使用し, 10 分割のクロスバリデーションによる「が」「がも」の使い分けの推定を行った結果から得たものである.

解析対象の「がも」である「も」を含む文節は一番最後の文にあり, 対象の「も」は太字で示している.

メコンの旅のささやかな冒険は, ラオス国境から 中国$_{15}$・雲南省に入ることだった. ラオス北部は山が深いうえ, 外国人による国境越えは不可能, とされていた. 不安は残ったが, まず, 空路ルアンプラバンへ. ここからモーターボートでメコン川と支流のハン川を約三時間. ヘルメットにライフジャケットを着込んで, 岩場だらけの渓流を一気に上り, 通称「中国$_{15}$ 橋」の架かる山村に着いた. そこからトラックに揺られ約十時間. 国境の村で一泊. 翌朝, 検問所で恐る恐るパスポートを出すと, 両国いずれの係官も, 首をひねった末, 通過を許可してくれた. 国境から約百七十キロの景洪は「中国$_{15}$ 雲南省・西双版納（シーサンパンナ）ダイ族自治州」の州都で, タイ人の元祖, ダイ族が住む, いわばタイ人の故郷だ. 一年前に 中国$_{15}$ 側から訪れたことがあった辺境の街は, 外資系ホテル, レストランが急増. 中国$_{15}$ からの流民も 増え$_{18,19}$, 一大観光都市として脚光を浴びていた. 「インドシナ半島から世界市場へ」を目指す雲南省は, ベトナム, ラオス, ミャンマーとの国境貿易で 中国$_{15}$ 商品の売り込みに成功. 上海や広東方面からの 中国$_{15}$ 人旅行者**も**, うなぎ登りに 増えた$_{18,19}$.

この文脈から生成される文脈素性は表 3.3 の素性番号 15 番と 18,19 番に該当する. 二重下線が文脈素性を生成する条件の元となる要素で, 下線部分がその文脈素性を生成する条件に該当する部分である. また, 表 3.3 の素性番号に当る数字をそれぞれ下線部分の右下に示す.

4.6 KNP の性能調査

4.4 節で学習データを拡張すると提案手法は「が」「がも」のマクロ平均が下るという結果となった. これは学習データに追加した KNP を用いたデータに誤りが含まれていたことが考えられる. そこでガ格の「も」とヲ格の「も」において, KNP がどのくらいの性能で正しく格解析を行えているのかを調査した.

毎日新聞 91 年の 1 年分を KNP4.01 で格解析を行い, 「がも」「をも」と判定された箇所を含む文をそれぞれランダムに 10 文ずつ取り出し人手で評価を行った. 人手で評価を行った結果を表 4.15 に示す. それぞれの人手判定で不正解であった文を以下に示す. また, 下線部分は, 対象の助詞が含まれている文節である.

KNP が「がも」と判定した箇所での誤り例

- 対ソ支援も この枠内で考えていこうという姿勢だ.
- ２人は「一緒に 入場行進も やらせてもらいたいぐらい」と開会式に心をはせた.

KNP が「をも」と判定した箇所での誤り例

- 野村証券のある営業マンは「株価が高値を続けていた当時は, 野村が推奨した株は必ず値段が上がるという神話があった. 自分たちも 価格は野村がつくる, と信じてそれをプライドに株を売りまくった. それを株価操作と言われると……」と言葉を濁した.
- この時点で多国籍軍筋は, 戦争で打撃を受けたイラク軍には反政府運動を乗り切る力はないとの見方を流し, 米当局者も フセイン政権は数カ月で崩壊すると予測した.
- これだけなら絵はくそ面白くもないが, ちゃんと中心は ずしも 心得ていた.

表 4.15 より, KNP の正解率は「がも」は 8 割, 「をも」は 7 割という結果を得た. それぞれの誤りの文を見ると, 「がも」の場合はガ格でなくヲ格, 「をも」の場合はヲ格でなくガ格という判定が正しいと思われる. また, 「をも」の最後の誤りの文は, 「中心はずし」という単語を正しく認識できなかったことによる間違いである.

KNP の格解析に誤りが含まれていることがわかった. これが, 提案手法「がも」の学習データを拡張した際のマクロ平均の低下の理由の一つである可能性がある.

表 4.14: 文脈素性を削除した「も」の使い分けにおける分類結果

使い分け問題	分類先	正解率	マクロ平均
に/にも	「に」	0.64 (4485/7045)	0.69
	「にも」	0.74 (173/ 233)	
が/がも	「が」	0.63 (301/ 480)	0.65
	「がも」	0.66 (47/ 71)	
を/をも	「を」	0.51 (838/1641)	0.58
	「をも」	0.64 (18/ 28)	

表 4.15: KNP が「がも」「をも」と判定した箇所での正解率

	「がも」	「をも」
正解率	0.8 (8/10)	0.7 (7/10)

第5章　おわりに

本研究では教師あり機械学習による助詞「も」の使い分けを行った.

学習データの拡張を行うことで提案手法は,「も」の使い分けを6割から8割のマクロ平均で行えた. 素性分析により, 助詞の「も」がどのような文で使用されやすいのかのその特徴を得ることができた.

具体的には,「また」や「や」などの文構造を並列にする助詞が用いられる場合や,「だれも」など直前に疑問語がきて不定語となる場合など,「にも」「がも」「をも」になりやすいという特徴を得た.

また, 文脈に関する素性を削除して実験を行うことで, 提案手法の「がも」の使い分けにおいて文脈に関する素性の有効性を確認することができた.

今後は「も」の使い分けで得られた特徴を種々の事柄に利用していきたい. この研究で得られた知見を利用して,「も」の使い分けの研究以外の「も」に関係する研究も行いたいとと思っている.

謝辞

本研究を進めるにあたり, 終始に渡り研究の進め方や本論文の書き方など, 細部にわたる御指導を頂きました, 鳥取大学工学部知能情報工学科計算機工学講座 C の村田真樹教授に心から御礼申し上げます. また, 本研究を進めるにあたり, 御指導, 御助言を頂きました, 村上仁一准教授, 徳久雅人講師に心から御礼申し上げます. その他様々な場面で御助言を頂いた計算機工学講座 C 研究室の皆様に感謝の意を表します.

参考文献

[1] 沼田善子, “現代日本語の「も」”, つくば言語フォーラム編,「も」の言語学, pp.13-58, 1995.

[2] 岡野ひさの, “助詞「も」の周辺的用法はなぜ周辺的なのか”, 福岡大学研究部論集, A, 人文科学編 10(7), 213-222, 2010-12.

[3] 中俣尚己, “日本語のとりたて助詞と並列助詞の接点”, 言語文化学研究, 言語情報編 3, 153-176, 2008-03.

[4] 益岡隆志, 田窪行則, “基礎日本語文法-改訂版-”, くろしお出版.

[5] 小泉保, “日本語教師のための言語学入門”, 大修館出版.

[6] 林裕哉, 村田真樹, 徳久雅人, “教師あり機械学習を用いた文の順序推定”, 言語処理学会第 18 回年次大会, P1-12, pp239-242, 2012.

[7] 三浦智, 村田真樹, 徳久雅人, “教師あり機械学習による助詞の使い分け”, 言語処理学会第 19 回年次大会, P1-1, pp.322-325, 2013.

[8] Satoshi Miura, Liangliang Fan, Masaki Murata, and Masato Tokuhisa, “Automatic Selection and Contextual Analysis of the Japanese Particles ”Ga” and ”Wa” Using Machine Learning”, The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012), Nov. 20-24, 2012, Kobe, Japan, pp.2221-2224.

[9] 木曽宏顕, 森辰則, 中川裕志, “関係意味論に基づく取り立て助詞「も」の定式化”, 自然言語処理研究会報告, 33-38, 1995-01-20.

[10] TinySVM, `http://chasen.org/taku/software/TinySVM/`

[11] 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均, “意味ソート msort 意味的並びかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例 ”, 自然言語処理, 7 巻, 1 号, pp.89-96, 2000.

[12] 分類語彙表, http://www.ninjal.ac.jp/productsk/kanko/goihyo/

[13] 村田真樹, “機械学習手法を用いた日本語格解析”, 自然言語処理研究会報告 2001(69), 113-120, 2001-07-16.

[14] Nell Cristianini and John Shawe-Taylor, “An Inttoduction to Support Vector Machines and Other Kernel-based Learning Methods”, Cambridge University Press, 2000.

[15] 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.

付 録A　「も」の分類

付録 A では「も」の分類を扱う.

助詞の「も」を含む文からは, 直接伝わる情報と, 助詞の「も」を使うことでそこから読み取れる間接的な情報を得ることができる. 例えば「太郎も来た. 」という文からは,「太郎が来た」ことと,「太郎以外の誰か (例えば二郎) が来た. 」ことが同時に示される. 本論文では沼田の研究 [1] を参考にし,「も」の直前にある名詞句を自者といい, 自者に対する他の名詞句を他者という. 上の例では「太郎」が自者で,「他の誰か (例えば二郎)」が他者にあたる. このように他者が想定できる場合, その「も」はとりたて詞の「も」といわれる. この他者は文中に存在していなくても他者が想定できる場合はとりたて詞の「も」とされる.

とりたて詞である場合ととりたて詞でない場合の例を以下に示す.

とりたて詞である ドルの評価も下がり, 対ドルレートが大きく変わらない割には, 円の評価が落ちる.

とりたて詞でない 政策金利を一年間に三%, 四%も動かすのは過激すぎる.

本節では, 対象の「も」がとりたて詞であるか否かの分類の問題を扱う. とりたて詞であるか否かの推定を, 教師あり機械学習により行う.

とりたて詞であるか否かの推定では, とりたて詞であるとき, 前方の文脈中に他者が存在しない場合もあるが他者は必ず存在するため, 確実に他者が存在しないものを省くことができる. これは今後行う予定である文脈中に他者が存在するとりたて詞であるか否かの推定実験に役立つと思われる.

実際にとりたて詞であるか否かの分類を行った. 毎日新聞 91 年, 92 年の助詞「も」を含むはじめの 100 文に, 対象の「も」がとりたて詞であるか否かのタグ付けを行い, 91 年のものを学習データ, 92 年のものをテストデータとして使用した. 機械学習の素性には対象の「も」の前後の文字, 形態素, 形態素の品詞を利用した. また, ベースライン手法として全てをとりたて詞とする場合と全てをとりたて詞でないとする場合の二種類を求

めた. それぞれのデータ数を表 A.1, 結果を表 A.2 に示す. また, 表 A.3 に実験の結果を F 値で表す.

対象の「も」の前後の情報しか素性として利用していないのに, ある程度の性能を得ることができた. 得られた特徴としては, 解析対象の「も」の直前の形態素が名詞である場合はとりたて詞でることが多い, また, 解析対象の「も」の直前の形態素が「%」である場合や直後の形態素が動詞である場合はとりたて詞でないことが多いなどである.

今後文脈の情報などを素性に追加して性能の向上を目指す.

表 A.1: とりたて詞に関わる実験データ数

	全データ数	とりたて詞である数	とりたて詞でない数
学習データ	100	83	17
テストデータ	100	61	39

表 A.2: とりたて詞に関わる分類結果

手法	分類先	正解率	マクロ平均
提案手法	とりたて詞である	1.00 (61/61)	0.74
	とりたて詞でない	0.49 (19/39)	
ベースライン手法 全てとりたて詞である	とりたて詞である	1.00 (61/61)	0.50
	とりたて詞でない	0.00 (0/39)	
ベースライン手法 全てとりたて詞でない	とりたて詞である	0.00 (0/61)	0.50
	とりたて詞でない	1.00 (39/39)	

表 A.3: とりたて詞に関わる分類結果 (F 値)

手法	分類先	再現率	適合率	F 値
提案手法	とりたて詞である	1.00 (61/61)	0.75 (61/81)	0.86
	とりたて詞でない	0.49 (19/39)	1.00 (19/19)	0.66
ベースライン手法 全てとりたて詞である	とりたて詞である	1.00 (61/61)	0.61 (61/100)	0.76
ベースライン手法 全てとりたて詞でない	とりたて詞でない	1.00 (39/639)	0.39 (39/100)	0.56