教師あり機械学習による助詞「も」の分析

井上愛実 *1 村田真樹 *1 徳久雅人 *1 馬青 *2
*1 鳥取大学 工学部 知能情報工学科
*2 龍谷大学 理工学部 数理情報学科
{s102003, murata, tokuhisa} @ ike.tottori-u.ac.jp
qma@math.ryukoku.ac.jp

1　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] に関わる様々な知見を得ることができる. 例えば, 林ら [2] 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 [3] 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 [3] 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, 助詞「も」を分析し, 助詞「も」に関わる知見を得ることを目指す. 本研究の助詞「も」の分析では, 「も」の文中での使用における課題 (「も」の生成) と, 「も」の使用法に基づく分類の課題 (「も」の分類) の二つを扱う.

助詞「も」を分析しそれに関する知見を得ることは, 以下の二つ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.

本稿では, 2 節で「も」の生成の研究について, 3 節で「も」の分類の研究について述べる.

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」の使い分け問題と, とりたて詞であるか否かの分類問題に取り組んだ.
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 割りの正解率を得ることができた. また, とりたて詞であるか否かの分類問題では 7 割の正解率を得ることができた.
3. 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と, とりたて詞である場合の「も」の特徴を得ることができた.

2　「も」の生成

文において, 助詞「も」を使うべきかどうかの課題を, 本稿では「も」の生成と呼ぶ. 本稿での「も」の生成では, 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」の生成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.

2.1　問題設定

本稿では, 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 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 2 に示すようにガ格とヲ格が多いことがわかった.

表 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 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

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」を使うかいな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」かの分類, または, 「を」か「をも」かの分類を扱うものである.

本稿では, 「も」の生成として, 「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.

*1 本稿では「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, 本稿では扱わない.

表 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
10	9 の品詞
11	解析対象の助詞の直後の単語
12	11 の品詞
13	文内の単語
14	13 の分類語彙表の分類番号
15	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
16	解析対象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名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名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動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動詞のどれかが前方に存在しているか
22	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
23	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

表 4 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

2.2 提案手法

本稿での提案手法では, 「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[5] 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 [6] を参考にして表 3 のものを用いる. 分類語彙表 [7] を利用する素性は, 村田らの手法 [6] を利用し素性化する.

2.3 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 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 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, 出現確率は表 5 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 5 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 6 「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 全て「に」	「に」	1.00 (7045/7045)	0.50
	「にも」	0.00 (0/ 233)	
ベースライン手法 全て「にも」	「に」	0.00 (0/7045)	0.50
	「にも」	1.00 (233/ 233)	

表 7 「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 全て「が」	「が」	1.00 (480/ 480)	0.50
	「がも」	0.00 (0/ 71)	
ベースライン手法 全て「がも」	「が」	0.00 (0/ 480)	0.50
	「がも」	1.00 (71/ 71)	

表 8 「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 全て「を」	「を」	1.00 (1641/1641)	0.50
	「をも」	0.00 (0/ 28)	
ベースライン手法 全て「をも」	「を」	0.00 (0/1641)	0.50
	「をも」	1.00 (28/ 28)	

2.4 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[*2] (各分類での正解率の平均) を表 6 から表 8 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, 提案手法では「にも」, 「がも」, 「をも」の適合率は 0.06, 0.22, 0.03 ととても低い値なり, それぞれの F 値も 0.12, 0.34, 0.06 と低くなってしまう. これは本来「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ことが原因である.

2.5 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, 2.3 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 9, 出現確率は表 10 である. また, 学習データ追加後の実験結果はそれぞれ表 11 である.

[*2] 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 9 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 10 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 11 学習データ拡張後の「も」の生成における分類結果

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)	

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.

2.6 KNP の性能調査

前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い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 12 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 12 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.

表 12 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)

表 13 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 14 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

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.

2.7 素性分析

「も」の生成において, 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, 2.3 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 13, また得られた有用な素性の例を表 14 に示す.

2.7.1 「に」「にも」の分析

表 14 の素性に関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.

2.7.2 「が」「がも」の分析

表 14 の素性に関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．

2.7.3 「を」「をも」の分析

表 14 の素性に関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．

3 「も」の分類

本節では「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．本稿では沼田の研究 [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．

表 15 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 16 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)	

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 15，結果を表 16 に示す．

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．

4 おわりに

本研究では教師あり機械学習による助詞「も」の分析を行った．具体的には，「も」の生成と「も」の分類を扱った．

「も」の生成では学習データの拡張を行うことで提案手法は，「に」「にも」，「が」「がも」，「を」「をも」の分類を 6 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，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．

「も」の分類ではとりたて詞であるか否かの分類を扱った．データ数や利用した素性が少ないながらもある程度の性能を得ることができ，とりたて詞の場合は直前の形態素が名詞である，とりたて詞でない場合は直前の形態素が「%」であるなどの特徴も得ることができた．

今後は「も」の生成で得られた特徴を種々の事柄に利用していきたい．「も」の分類では，扱う分類を増やしていきたいと思っている．例えば，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などを行いたい．さらに，「も」の生成と分類で得られた知見を利用して，他者を自動推定する研究も行いたいとと思っている．

参考文献

[1] 沼田善子，“現代日本語の「も」とりたて詞とその周辺”，つくば言語フォーラム編，「も」の言語学，pp.13-58，1995.

[2] 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.

[3] 三浦智，“機械学習による助詞の使い分け”，鳥取大学大学院工学研究科修士論文，2012.

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

[5] 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.

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