教師あり機械学習による助詞の使い分け

三浦 智　村田 真樹　徳久 雅人
鳥取大学大学院　工学研究科
情報エレクトロニクス専攻

{s072052, murata, tokuhisa}@ike.tottori-u.ac.jp

1　はじめに

日本語の文法を対象とした研究には様々なものがある [1][2][3][4][5]．一般的に，ノンネイティブの日本語学習者にとって，助詞の理解は難しいとされている．その中でも副助詞「は」と格助詞「が」の使い分けや，格助詞「に・へ・を・で」の使い分けは特に困難である．例えば，副助詞「は」と格助詞「が」の使い分けにおいて,「彼は学生だ」と「彼が学生だ」の二文は文法として誤りでなく，かつニュアンスも近い．田中ら [6] は,「は・が」の使い分けについて「は」は既知情報や説明文,「が」は未知情報や描写文を示すと述べているが，明確な分類法については述べていない．そこで本研究では，日本語学習者の支援を行うため，使い分けが困難な助詞の自動推定を行う．これにより，日本語学習者が助詞の使い分けに迷う場合，どちらを使うべきかを示すシステムを構築可能になる．また，助詞に関わるデータの分析を行うことにより，日本語学習者にとって有用な情報を獲得する．

まず，副助詞「は」および格助詞「が」を含む文を京大コーパス 3.0[7] から収集し，これらを教師データとして利用する．次に，獲得した文から副助詞「は」および格助詞「が」を取り除いた文を収集し，これらをテストデータとして利用する．獲得した教師データ，テストデータを利用し，Support Vector Machine(以下 SVM) で取り除いた助詞を再推定する．最後に，教師データを分析し副助詞，格助詞の使い分けの手掛かりを模索する．同様の実験を,「に・へ」「に・を」「に・で」に対しても行った．

SVM を利用した推定の他に，Komori[8] の手法を利用した推定手法を比較手法として用いた．

本研究の主張点は次の 3 つである．

1. 機械学習を用いて格助詞「に・へ」「に・を」「に・で」の分類を初めて行った．(「は・が」の分類は文献 [9] において既に行っている．)
2. 「は・が」「に・を」「に・で」の使い分けの問題において，機械学習により比較手法よりも高い正解率を得た．
3. 実験データを用いた素性の分析によって，多数の使い分けに役立つ表現を獲得した．

表 1: データ数

助詞	教師データ数	テストデータ数
は	4323	5558
が	4653	6009
に	5529	7045
で	2238	3071
を	6432	8329
へ	85	85

2　実験データ

京大コーパスの 1995 年 1 月 1 日～1995 年 1 月 9 日（休刊日のため 1995 年 1 月 2 日を除く）と 1995 年 1 月 10 日～1995 年 1 月 17 日のデータから教師データとテストデータを生成する．まず，対象の助詞が最低 1 つは出現する文を抽出する．次に，対象の助詞を取り除く．対象の助詞を取り除いた文に対して，取り除いた助詞の種類を分類先として与える．文中に対象の助詞が複数存在する文の場合，対象の助詞の出現数分の教師データを獲得する．例えば,「今は鳥取が熱い」の文からは次のような教師データを獲得する．X は取り除いた「は・が」の位置を表す．

副助詞は　今 X 鳥取が熱い

格助詞が　今は鳥取 X 熱い

の 2 つの教師データを獲得する．教師データの素性の情報は京大コーパスの形態素・構文情報から得た．教師データは 1995 年 1 月 1 日～1995 年 1 月 9 日のデータから，テストデータは 1995 年 1 月 10 日～1995 年 1 月 17 日から獲得した．教師データ数を表 1 に示す．

3　提案手法

本研究では，日本語学習者が「は」と「が」,「に」と「へ」,「に」と「を」,「に」と「で」の使い分けに迷った場合を想定し，それらの助詞を 1 つ空白にした文を問題とする．その問題に対し，機械学習を利用し空白に入れるべき助詞を推定する．

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 1 次の多項式カーネルを利用した．

機械学習で利用する素性は村田ら [11] の研究を参考にして以下のものを用いる．分類語彙表 [12] を利用

する素性は，村田ら [13] の手法を利用し素性化する．N(体言の文節に相当) は推定すべき助詞を含む文節を表し，V(述部に相当) は N の係り先の文節を表す．

例　私（N）【「が・は」等の対象の助詞】社長（V）だ
１　V における自立語の連続
２　V の最初の自立語の基本形
３　２の単語の品詞
４　２の単語の分類語彙表 [12] の分類番号の 1,2,3,4,5,6,7 桁までの数字．ただし，分類番号に対して文献 [13] の表の変更を行っている．
５　V に出現する付属語
６　N における自立語の連続
７　N の最後の自立語
８　７の単語の品詞
９　７の単語の分類語彙表 [12] の分類番号の 1,2,3,4,5,6,7 桁までの数字．ただし，分類番号に対して文献 [13] の表の変更を行っている．
１０　N に体言が存在するか否か
１１　同一文に共起する語
１２　１１の単語の分類語彙表 [12] の分類番号の 1,2,3,4,5,6,7 桁までの数字．ただし，分類番号に対して文献 [13] の表の変更を行っている．
１３　V の係り先の自立語の連続
１４　V の係り先の文節における最後の自立語の基本形
１５　１４の品詞
１６　１４の単語の分類語彙表 [12] の分類番号の 1,2,3,4,5,6,7 桁までの数字．ただし，分類番号に対して文献 [13] の表の変更を行っている．
１７　V の係り先に体言が存在するか否か
１８　V にかかる N 以外の体言の文節の自立語の連続
１９　V にかかる N 以外の体言の文節の最後の自立語
２０　１９の単語の品詞
２１　１９の単語の分類語彙表 [12] の分類番号の 1,2,3,4,5,6,7 桁までの数字．ただし，分類番号に対して文献 [13] の表の変更を行っている．
２２　V にかかる N 以外の体言が存在するか否か
２３　V にかかる N 以外の体言がとっている格
２４　解析対象の助詞の直前，直後の単語
２５　解析対象の助詞の直前，直後の単語の品詞
２６　解析対象の文内において，解析対象の文節以外にある助詞
２７　解析対象の文節内の名詞がすべて，記事内の前方に存在しているか否か

4　先行研究手法

提案手法と比較する Komori[8] の先行研究手法を説明する．

4.1　先行研究手法１

助詞「は・が・に・へ・で・を」（以下対象の助詞）の直前に出現する名詞の統計情報を利用し，X に入る助詞はどちらであるかを推定する．具体的には，現在解いている箇所の直前の単語を A とするとき，学習データにおいて A の後により高い頻度で出現する助詞をもとめ、それを現在解いている箇所の助詞とする．例えば,「は・が」の分類において，X の直前に「今」と言う名詞が出現しており，学習データにおいて「今」が出現した場合に副助詞「は」を使う確率が格助詞「が」を使う確率よりも大きいのであれば，X に入る助詞は「は」であると推定する．

表 2:「は・が」での正解率

手法	正解率
SVM	0.760
先行研究手法 1	0.615
先行研究手法 2	0.522
全て「が」に分類	0.519
全て「は」に分類	0.480

4.2　先行研究手法２

対象の助詞の直前，直後に出現する品詞を利用し，X に入る助詞はどちらであるかを推定する．具体的には，現在解いている箇所の直前の単語の品詞を A とし，直後の単語の品詞を B とするとき，学習データにおいて A と B に挟まれる個所により高い頻度で出現する助詞をもとめ、それを現在解いている箇所の助詞とする．例えば,「は・が」の分類において，X の直前に名詞が出現しており，X の直後に動詞が出現している状況で，学習データにおいて直前に名詞が出現し直後に動詞が出現する場合に「は」になる確率が「が」になる確率よりも大きいのであれば，X に入る助詞は「は」であると推定する．

5　実験

副助詞「は」，格助詞「が」が取り除かれた文に対し，SVM を利用し取り除かれた助詞の推定を行った．提案手法の正解率の他に，全てを「が」に分類する手法，全てを「は」に分類する手法，先行研究手法の正解率ももとめた．これらの正解率を表 2 に示す．表のように，SVM の正解率は 0.760 であり，比較手法の中で最も高い値となった．また，SVM の「が」の推定は F 値 0.768(再現率 : 0.765，適合率 : 0.772),「は」の推定は F 値 0.751(再現率 : 0.755，適合率 : 0.748) であった．また,「に」と「へ」,「に」と「を」,「に」と「で」が取り除かれた文に対しても，同様に助詞の推定を行った．正解率を表 3 に示す．表のように,「に・で」「に・を」の分類において，SVM の手法が比較手法の中で最も高い値となった．また,「に・へ」においては全て「に」でない以外の全ての手法がほぼ同等の正解率であった.「に・へ」に関しては「へ」の教師数が少ないため，1994 年の毎日新聞の記事一年分の各分類ごとの教師データ数を揃えたデータ（に : 3,339 文，へ : 3,339 文）を教師として利用し，各手法で推定を行う実験を追加で行った．正解率を表 4 に，SVM の教師の違いによる F 値の差を表 5 に示す．正解率は下がったが,「へ」の F 値は上昇した．

6　分析

どういった素性が出現すると「は」,「が」,「に」,「へ」,「で」,「を」が使われやすいのかを明らかにするために，素性の頻度分析を行った.「は・が」,「に・を」は本研究で用いた教師データを利用して分析を行う.「に・で」は教師データ数が偏っているため，データ数を揃えたデータ（に : 2,238 文，で : 2,238 文）を利

表 3: 各分類の正解率

手法	正解率		
	にへ	にで	にを
SVM	0.987	0.812	0.889
先行研究手法 1	0.985	0.736	0.617
先行研究手法 2	0.988	0.696	0.582
全て「に」	0.988	0.696	0.458
全て「に」でない	0.011	0.303	0.541

表 4: 94 年データ利用「に・へ」

手法	正解率
SVM	0.862
先行研究手法 1	0.867
先行研究手法 2	0.180
全て「に」	0.988
全て「に」でない	0.011

用し分析を行う.「に・へ」においては「へ」の教師が少ないため，1994 年の毎日新聞の記事一年分のデータ数を揃えたデータ（に : 3,339 文，へ : 3,339 文）を利用する.

素性の出現頻度が 50 回以上であり，テストデータにおいてその素性が出現した場合にその分類先が出現する確率が 0.75 以上の素性を使い分けに有用な素性として獲得する．獲得された使い分けに有用な素性の数を表 6 に示す.

6.1 分析 1 :「は・が」使い分け

分析の結果,「は」および「が」の使い分けに，述部 V に存在する判定詞「だ」が関係することが確認できた．この結果は我々の先行研究 [9] と同様の結果である．有効な素性の例を表 7 に示す．確率はテストデータにおいてその素性が出現した場合にその分類先が出現する確率であり，頻度はテストデータでのその素性の頻度である．以下に表 7 の素性を含む例文を示す.

述部 V に存在する判定詞「だ」 二番目**は**、この制度は大きな政党同士に政権交代を可能ならしめるもの だ。

述部 V のかかり先に体言が存在 八五年四月に「電電公社」**が**民営化された 際、電気通信事業法が成立。

また，この他にも，推定する助詞の直後の単語に「記号 :「」が出現する場合，直後に「首相」が出現する場合，述部 V に出現する最初の自立語が「話す」の場合に「は」が使われやすく，推定する助詞の直後の単語に「あう・ある・できる・出る」が出現する場合に「が」が使われやすい等のルールも獲得できた.

6.2 分析 2 :「に・へ」での使い分け

分析の結果,「に」および「へ」の使い分けに，述部 V における最初の自立語が関係することがわかった．有効な素性の例を表 8 に示す．以下に表 8 の素性を含む例文を示す.

表 5: 「に・へ」教師ごとの F 値

手法	教師	分類先	F 値	再現率	適合率
SVM	95 年 1 月 1〜9 日	に	0.993	0.998	0.988
		へ	0.042	0.023	0.200
	94 年 全日	に	0.924	0.863	0.996
		へ	0.109	0.717	0.059
先行研究手法 1	95 年 1 月 1〜9 日	に	0.992	0.997	0.988
		へ	*	0.000	0.000
	94 年 全日	に	0.928	0.872	0.992
		へ	0.073	0.447	0.040
先行研究手法 2	95 年 1 月 1〜9 日	に	0.993	1.000	0.988
		へ	*	0.000	0.000
	94 年 全日	に	0.291	0.171	0.997
		へ	0.025	0.964	0.013
全て「に」	なし	に	0.993	1.000	0.988
		へ	*	0.000	0.000

表 6: 有用な素性の数

分類問題	分類先	獲得ルール数
は・が	は	40
	が	49
に・へ	に	63
	へ	127
に・で	に	28
	で	34
に・を	に	74
	を	114

述部 V における最初の自立語「つく」 生命保険**に** ついて 思うこと

述部 V における最初の自立語「行く」 東京・二子玉川園の「ナムコ・ワンダーエッグ」**へ** 行って みた。

また，この他にも，推定する助詞の直前の単語に「こと」が出現する場合，直後に「よる」が出現する場合に「に」が使われやすく，同一文中に格助詞「から」が存在する場合，推定する助詞の直前の単語に「そこ・ところ・外」が出現する場合,「へ」が使われやすい等のルールも獲得できた.

6.3 分析 3 :「に・で」使い分け

分析の結果,「に」および「で」の使い分けに，助詞の直後の単語，述部 V における最初の自立語が関係することがわかった．有効な素性の例を表 9 に示す．以下に表 9 の素性を含む例文を示す.

助詞の直後の単語「対する」 犯罪者**に** 対して「ムチ打ち」の刑を導入する法案が四日までに米ミシシッピ州議会に提出された。

述部 V における最初の自立語「行われる」 ルワンダからの報道によると、交換は各州中心都市の銀行**で**三日と四日に 行われ、それ以降旧紙幣は無価値となる。

この他にも，推定する助詞の直後の単語に「つく・なる・よる」が出現する場合「に」が使われやすく，同

表 7: 「は・が」での素性分析

分類先	素性	確率	頻度
は	述部 V に判定詞「だ」が存在	0.754	809
が	述部 V の係り先に体言が存在	0.873	1255

表 8: 「に・へ」での素性分析

分類先	素性	確率	頻度
に	述部 V の最初の自立語「つく」	1.000	189
へ	述部 V の最初の自立語「行く」	0.948	466

表 9: 「に・で」での素性分析

分類先	素性	確率	頻度
に	対象の助詞の直後の単語が「対する」	1.000	52
で	述部 V の最初の自立語が「行われる」	0.903	62

表 10: 「に・を」での素性分析

分類先	素性	確率	頻度
に	述部 V の最初の自立語が「よる」	1.000	371
を	述部 V の最初の自立語が「持つ」	0.844	123

一文中に格助詞「に」が存在する場合，推定する助詞の直前の単語に「中」が出現する場合「で」が使われやすい等のルールも獲得できた．

6.4 分析４ :「に・を」使い分け

分析の結果,「に」および「を」の使い分けに，述部 V における最初の自立語が関係することがわかった．有効な素性の例を表 10 に示す．以下に表 10 の素性を含む例文を示す．

述部 V における最初の自立語「よる」 わが国にふさわしい国際貢献**に** よる 世界平和の創造

述部 V における最初の自立語「持つ」 アンケートでは「好感**を** 持つ 各政党の首脳、幹部名」も挙げてもらった。

この他にも，同一文中に格助詞「を」が出現する場合，推定する助詞の直前の単語に「前・日」が出現する場合,「に」が使われやすく，推定する助詞の直前の単語に「開く・求める・見る」が出現する場合「を」が使われやすい等のルールも獲得できた．

7 おわりに

日本語学習者の支援のために，本研究では，機械学習を利用した助詞の推定を行った．

実験の結果,「は・が」の分類における SVM の正解率は 0.760,「が」の推定は F 値 0.768(再現率 : 0.765，適合率 : 0.772),「は」の推定は F 値 0.751(再現率 : 0.755，適合率 : 0.748) であった．また,「に・で」「に・を」の分類において，SVM の手法が比較手法の中で最も高い値となった．また,「に・へ」においては全て「に」でない以外の手法がほぼ同等の正解率であった．

実験データにおいて素性の分析を行い,「が・は」の使い分けに役立つ表現を獲得した．副助詞「は」になりやすい表現として，述部に存在する助詞「だ」があった．また,「に・へ・で・を」の使い分けに役立つ表現を多数獲得した．これらの知見は，今後の助詞に関わる研究に役立つと思われる．

参考文献

[1] H. Suzuki and K. Toutanova: 2006, “Learning to Predict Case Markers in Japanese,” ACL-COLING.

[2] 内元 清貴, 村田 真樹, 馬 青, 関根 聡, 井佐原 均 : 2000, “コーパスからの語順の獲得”, 自然言語処理, 7 巻, 4 号, 163-180.

[3] 蓮池 いずみ: 2004, “場所を示す格助詞選択のストラテジー —韓国語母語話者と中国語母語話者の比較—”, 言葉と文化. v.5, 105-117.

[4] 若生 正和: 2012, “韓国人日本語学習者による場所の格助詞「に」と「で」の選択に関する研究”, 大阪教育大学紀要　第□部門人文科学 60(2), 91-99.

[5] 杉村 泰: 2005, “上級・超上級日本語学習者に見る格助詞「に」と「へ」の使い分け”, 2005, 言語文化論集 26, 91-102.

[6] 田中 稔子: 1990, 田中　稔子の日本語の文法—教師の疑問に答えます—, 近代文藝社.

[7] 京大コーパス: http://nlp.ist.i.kyoto-u.ac.jp

[8] Y. Komori: 2003, “Disambiguating between ‘wa’ and ‘ga’ in Japanese,” Proceedings of the Class of 2003 Senior Conference, pp.30-34.

[9] S. Miura, L. Fan, M. Murata and M. Tokuhisa: 2012, “Machine Learning for Analysis and Detection of Redundant Sentences Toward Development of Writing Support Systems”, SCIS-ISIS, PT-4.

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

[11] 村田 真樹: 2001, “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—”, 情報処理学会自然言語処理研究会, 2001-NL-144, 113-120.

[12] 分類語彙表: http://www.ninjal.ac.jp/products-k/kanko/goihyo/

[13] 村田 真樹, 神崎 享子, 内元 清貴, 馬青, 井佐原 均: 2000, “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, 89-96.