機械学習による「が」と「は」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」の使い分けは特に困難である．例えば,「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，副助詞「は」および格助詞「が」の自動推定を行う．これにより，日本語学習者が「は・が」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，副助詞「は」と格助詞「が」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」または格助詞「が」を含む文を京大コーパス 3.0[9] から獲得し，これらを教師データとして利用する．獲得した文から副助詞「は」および格助詞「が」を取り除いた文を獲得し，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．実験データを分析し副助詞，格助詞の使い分けの手掛かりを調査する．

本研究の主張点は次の 3 つである．

1　機械学習を用いた副助詞「は」，格助詞「が」の分類を初めて行った．

2　「は・が」の使い分けの問題において，機械学習により正解率 0.715 を得た。

3　実験データを用いた素性の分析によって,「が」「は」の使い分けに役立つ，格助詞「が」になりやすい表現，副助詞「は」になりやすい表現を獲得した．格助詞「が」になりやすい表現として，述部に出現する助詞「の」および「か」，述部の係り先の体言等を獲得した．また，副助詞「は」になりやすい表現として，述部に出現する助詞「だ」および「ない」，述部にかかる主部以外の体言につく格助詞「が」等を獲得した．

2　提案手法

本研究では，日本語学習者が「は」と「が」の使い分けに迷った場合を想定し,「は」,「が」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「は」と「が」のどちらを空白に入れるべきかを推定する．

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 [7] 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 [8] の手法を利用し素性化する．N は主部を表し，V は述部を表す．

例　私（N）【が or は】社長（V）だ

1　述部 V の自立語の品詞

2　述部 V の自立語の基本形

3　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 [8] の表の変更を行っている．

4　述部 V に出現する付属語

5　主部 N の体言の単語の品詞

6　5 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 [8] の表の変更を行っている．

7　述部 V にかかる主部 N 以外の体言の単語

8　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 [8] の表の変更を行っている．

9　述部 V にかかる主部 N 以外の体言がとっている格

10　同一文に共起する語

11　述部 V の係り先に体言が存在するか否か

12　述部 V の係り先の体言の単語

13　12 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 [8] の表の変更を行っている．

3　実験

3.1　教師データ

京大コーパスの 1995 年 1 月 1 日と 1995 年 1 月 3 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，副助詞「は」か格助詞「が」が最低 1 つが出現する文を抽出する．次に，文から副助詞「は」，格助詞「が」を取り除く.「は・が」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．文中に副助詞「は」，格助詞「が」が複数存在する文の場合，副助詞「は」，格助詞「が」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 日分のデータから，テストデータは 1995 年 1 月 3 日分のデータから獲得した．教師データ数を表 1 に示す．

3.2　SVM による「は」および「が」の推定結果

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．結果の正解率を全てを「が」に分類する手法，全てを「は」に分類する手法の正解率と比較した．正解率を表 2 に，F 値を表 3 に示す．推定の結果，SVM の正解率は 0.715 であり，比較手法のなかで最も高い値となった．また，SVM の「が」の推定は F 値 0.747(再現率：0.734，適合率：0.761),「は」の推定は F 値 0.674(再現率：0.690，適合率：0.659) であった．

表 1: 教師データ数

データの日付	合計	「は」	「が」
1995 年 1 月 1 日分	1048	504	544
1995 年 1 月 3 日分	689	294	395

表 2: 各手法正解率

手法	正解率
SVM	0.715
全て「が」に分類	0.573
全て「は」に分類	0.426

表 3: SVM 結果

分類先	F 値	再現率	適合率
が	0.747	0.734(290/395)	0.761(290/381)
は	0.674	0.690(203/294)	0.659(203/308)

4　分析

テストデータを使った素性の頻度分析を行うことにより，どういった素性が出現すると副助詞「は」，あるいは格助詞「が」が使われやすいのかを調査する．調査の結果，次の素性が影響を与えていることがわかった．

素性 1　述部 V に存在する助詞

素性 2　述部 V の係り先に体言が存在するか否か

素性 3　述部 V に存在する判定詞

素性 4　述部 V にかかる主部 N 以外の体言につく格助詞

素性 5　述部の自立語の基本形

表 4: 「が」になる素性 1

素性 1	確率	頻度
の	0.714	14
か	0.833	18

表 5: 「は」になる素性 1

素性 1	確率	頻度
だ	0.795	44
ない	0.777	45

4.1 素性１：述部 V に存在する助詞

分析の結果,「は」および「が」の使い分けに，述部 V に存在する助詞の種類が関係することがわかった．有効な素性の例を表 4 と表 5 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 4 の素性が出現する例文を示す．

素性１：の 山梨・上九一色村で昨年七月、異臭がすると の 届け出があり、長野、山梨両県警など警察当局が土壌を鑑定したところ有機リン系化合物が検出されたことが二日、分かった。

素性１：か 内閣不信任案が可決される状況が生まれる か どうか。

以下に表 5 の素性を含む例文を示す．

素性１：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。

素性１：ない だが、それでは、本筋の民意は問えない。

結果から，述部につく助詞が「の」および「か」の場合，格助詞「が」になりやすく,「だ」および「ない」の場合，副助詞「は」になりやすいことがわかった．

4.2 素性２：述部 V の係り先に体言が存在

分析の結果,「は」および「が」の使い分けに，述部 V の係り先に体言が存在するかどうかが関係することがわかった．有効な素性の例を表 6 に示す．以下に表 6 の素性を含む例文を示す．

表 6: 「が」になる素性 2

素性 2	確率	頻度
述部 V の係り先に体言が存在	0.903	104

表 7: 「は」になる素性 3

素性 3	確率	頻度
判定詞	0.790	43

表 8: 「が」になる素性 4

素性 4	確率	頻度
格助詞「から」存在	0.760	25

素性２：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。

結果から，体言にかかる動詞にかかる主部では「が」用いられやすいことがわかった．

4.3 素性３：述部 V に判定詞が存在

分析の結果,「は」および「が」の使い分けに，述部に判定詞が存在することが関係することがわかった．有効な素性の例を表 7 に示す．以下に表 7 の素性を含む例文を示す．

素性３：述部に判定詞が存在 特に駒井を警戒、タックルはバックス で なく、より当たりの強いＦＷがいくよう確認した。

結果から，判定詞（だ，です，である等）が述部に存在すると「は」になりやすいことがわかった．

4.4 素性４：述部 V にかかる主部 N 以外の体言につく格助詞

分析の結果,「は」および「が」の使い分けに，述部 V にかかる主部 N 以外の体言につく格助詞が関係することがわかった．有効な素性の例を表 8 と表 9 に示す．以下に表 8 の素性を含む例文を示す．

素性４：格助詞「から」存在 大みそかの第四十五回紅白歌合戦の視聴率が二日、ビデオ・リサーチ社 から 発表された。

表 9: 「は」になる素性４

素性４	確率	頻度
格助詞「が」存在	1.00	29

表 10: 「は」になる素性５

素性５	確率	頻度
ある	0.818	22

以下に表 9 の素性を含む例文を示す.

素性４：格助詞「が」存在 しかし、朝鮮中央テレビは共同社説をアナウンサーが読み上げる場面と、新年を迎える行事で金書記をたたえる歌や踊りの場面 が 放映されただけだった。

結果から，主部にかかる述部以外の主部のとっている格が,「から」の場合「が」になりやすく,「が」の場合「は」になりやすいことがわかった.

4.5 素性５：述部の自立語の基本形

分析の結果,「は」および「が」の使い分けに，述部の自立語の基本形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.

素性５：ある 「民意を問い直せ」の声が高まれば、避けられない状況はあり 得るが、今はそこまでいっていない。

結果から，述部の自立語の基本形が,「ある」の場合（〜であり等）「は」になりやすいことがわかった.

5 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「は」および「が」の推定を行った．結果，正解率は 0.715,「が」の推定は F 値 0.747(再現率 : 0.734，適合率 : 0.761),「は」の推定は F 値 0.674(再現率 : 0.690，適合率 : 0.659) であった．また，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．格助詞「が」になりやすい表現として，述部に存在する助詞「の」および「か」、述部の係り先の体言，述部にかかる主部以外の体言につく格助詞「から」があった．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」および「ない」，述部に存在する判定詞，述部にかかる主部以外の体言につく格助詞「が」，述部の自立語「ある」があった．これらの知見は，今後の「は」「が」に関わる研究に役立つと思われる．

参考文献

[1] 久野 ススム, 染谷 方良: 1989, 日本語学の新展開, くろしお出版.

[2] 森田 良行: 1995, 日本語の視点　ことばを創る日本人の発想, 創拓社.

[3] 益岡 隆志, 田窪 行則: 1992, 基礎日本語文法―改訂版―, くろしお出版.

[4] 玉村 文郎: 1992, 日本語を学ぶ人のために, 世界思想社.

[5] 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.

[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] 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.

[8] 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.

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

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

[11] Wikipedia: http://ja.wikipedia.org/wiki/

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