

東京工業大学工学部

学士論文

日本語書き言葉における
ガ格項の表出形式の決
定

指導教員 高村 大也 准教授
奥村 学 教授

平成28 2

提出者

学科 情報工学科

学籍番号 12 15375

氏名 山中 智仁

指導教員認定印		
学科長認定印		

日本語書き言葉におけるガ格項の表出形式の決定

指導教員 高村 大也 准教授

奥村 学 教授

情報工学科

12 15375 山中 智仁

機械による日本語文自動生成においては、文中のガ格項の表出形式を適切に 選択する必要がある。本研究ではガ格項を主題化するか、省略するか、あるいはそのまま表出するかという表出形式について分析し、表出形式の自動決定手法を提案する。提案手法では機械学習手法を使用し、共参照連鎖や、対象となるガ格項以外の形態素の出現を素性として用いた。また、ガ格項の表出形式を決める上でどのような手掛かりが有効かを実験により確認した。すでにアノテーションされている新聞記事を使用して提案手法の性能を確認した結果、

70.8%の正解率を得た。また、あるガ格項の表出形式を決定する際、そのガ格項に係る述語と同じ文節中の機能語が有効な手掛かりであることが分かった。残された課題としては、ガ格項以外の述語項への拡張、話し言葉への適用が考えられる。

目次

1 章 序論	1
2 章 関連研究	3
3 章 分析対象とするデータ	6
4 SVMを用いた決定手法	9
4.1 手法の概要	9
4.2 分類に使用する素性	9
5 章 実験と考察	14
5.1 実験設定	14
5.2 決定実験とその結果	15
6 章 結論	20
謝辞	21
参考文献	21

1

序論

自然言語生成とは、話し手の意図、あるいは伝えるべき伝達内容から、それに 対応するテキストを作り出す作業のことである [10]. 自動文章生成においては、深層生成によって伝えたい内容が選択された後、表層生成によって伝達内容が テキストに変換される. この表層生成プロセスにおいて、参照表現の選択や修 辞的關係の表出が行われるが、ある要素を主題化するかどうかという判断もこ こで行われる. 日本語母語話者が日本語文を発話するときには、ほぼ無意識に 何を主題として文を生成するかを決めている.

主題とはその文がなにについて述べるのか示すものである. 例 1.1 の一文目 は「田中さん」について述べたものであり、「田中さん」が主題であることを「は」で示されている. しかし、二文目では「幹事」について述べており、主語 である「田中さん」は主題ではない. なにが主題になるかというのは文の意図 に関わるものであり、機械的に決めるのは難しい.

1.1. 田中さんは 学生です. 田中さんが 幹事 です.

ここで、特に主語であるガ格項が主題化されるか否かという問題に焦点をあ てる. ガ格項が主題となる場合「は」で示され、主題とならない場合、「が」で 示される. 例1.2において、ガ格項である「田中さん」は主題化されて示されて いる. 「田中さんが親切だ」という文では不自然である. 逆に、例 1.3 では、ガ 格項である「男」は主題化されず、「が」で示されるほうが自然である.

1.2. 田中さんは親切です.

1.3. 私の友人に佐藤という男がいる. 私はこの男を尊敬している.

このように, ある要素を主題化して話すかどうかという問題が適切に解かれていなければ, 文がぎこちなくなったり, あるいは日本語としておかしい文に なったりする. そのため, 主題化するかどうかという問題は自然な日本語文生成のための重要な課題である.

また, そもそもガ格項が省略されることもある. 例1.4では「鳴らしつつ」の ガ格項である「オートバイ」が省略されている. 省略せずに書いてしまうと,

1

「オートバイ」がたびたび表記されて不自然である. このような, 格が省略されるか否かという問題は, 自然言語処理の重要な研究分野のひとつになっている.

1.4. 突然, (オートバイが)爆音を鳴らしつつ, 一台のオートバイが現れた.

本研究では, ある述語に係るガ格項を対象に, 主題化するか否か, あるいは 省略するかという 3 値分類を扱う. 焦点となっている述語の前後に出現する形態素や, ガ格項の共参照連鎖など, 大きくわけて 4 つのグループの素性を用いて分類器を構築する.

本論文の構成は以下の通りである. まず, 2 章では関連研究について述べる. 3 章では使用するデータについて述べる. 4 章ではガ格項を分類する手法を提案する. 5 章では提案手法の性能を実験によって検証し, その考察を行う. 6 章では本論文についてまとめ, 今後の課題を述べる.

関連研究

佐藤 [7] は, 日本語の文生成器として広く使われているツールは存在しないと指摘し, 機械による自動文生成を行うにあたって, 文生成とはなにかということについて考察した. 語の実際に現れる表記を「表層形」, 表層形に関して活用を抽象化したものを「書字形」, さらに書字形の表記を抽象化したものを

「語彙素」と呼んだ. 表層形としては「食べる」「食べた」「たべる」「たべた」などが考えられ, それらの表層形は活用に関して「食べる」と「たべる」という書字形に整理される. これらの書

字形は「食べる」という語彙素に抽象化される(表 2.1). これは語彙素と表記と活用形というパラメータから表層形を導くことができるということであり, 佐藤はこのパラメータのうち明示的に与えられていないものを何らかの方法で決定することが文生成の本質であると指摘している.

2.1: 書字形と語彙素

書字形	表層形
食べる たべる	食べる(辞書形), 食べた(タ形) たべる(辞書形), たべた(タ形)
語彙素	書字形
食べる	食べる(標準表記), たべる(かな表記)

また, 文生成器で使用する日本語文のデータ構造として日本語文節木(JBT を提案している. JBT は文節をノードとする木構造であり, それぞれのノードは文節属性リスト(BAL)で表される. 佐藤は「太郎が重い荷物を軽々と運ぶ」を生成する JBT として, Ruby Hash BAL を, Array JBT を記述した

2.1 を挙げている. string は表層形を, order は姉妹ノードにおける順序を示している. JBT のすべてのノード BAL において string order が決定されれば, その JBT に対応する表層文字列が生成可能である.

3

```
2.1. [{string: ' 運ぶ'},  
      [{string: ' 太郎が', order: 1}],  
      [{string: ' 荷物を', order: 2}, [{string: ' 重い'}]],  
      [{string: ' 軽々と', order: 3}]]
```

佐藤の考察に基づき, 緒方ら [5] Haori(羽織)と呼ぶ文生成器を作成している. Haori は表層文字列化が自明でない JBT を表層文字列化が自明である JBT に変換するプログラムであり, JBT のノード BAL の属性のうち string を定める. BAL には string を抽象化した属性として mp(主要語)や fp(機能語), fp を規定することのできる f(文節機能)などがある.

Haori は例えば例 2.2 から fp の生成と mp の活用形設定を行って「食べながら」という string を生成したり, 例 2.3 から形式名詞「こと」を補足して「食べる ことが」という string を生成することができる. しかし, 例 2.3 において, f type が不明であるとき, 他のノードによっては「食べることは」という string を生成するほうが自然である場合も考えられ, 他のノードによる制約を設ける 必要がある.

2.2. {f: {type: '副詞節', conj: 'ナガラ接'},
mp: {m: '食べる'}}

2.3. {f: {type: '補足節', case: 'ガ格'},
mp: {m: '食べる'}}

日本語文生成のうち、特に表層生成プロセスを対象としたものでは参照表現の選択に関する研究が盛んに行われている。自然な文章を生成するためには指示詞やゼロ代名詞といった参照表現を適切に選択することが重要である。

参照生成に関するヒューリスティックな手法を用いた研究には中岩ら [11]、滑川ら [12]、橋本ら [14] の研究がある。中岩ら [11] は接続語のタイプ、用言意味属性、様相表現のタイプの三つの制約に着目することで自動文内照応解析を行った。日英機械翻訳用例文 3718 文中の文内照応のゼロ代名詞 139 件、文章外照応のゼロ代名詞 175 件、および受け身変形することによって照応解析が不要となるゼロ代名詞 173 件の、計 487 件のゼロ代名詞に対して、再現率 98 %、適合率 100 %の性能を出せたと報告している。滑川ら [12] はヒューリスティックな手法を用いて、新聞の社説 25 文に対して省略生成実験を行った。その結果、原文で省略されている要素 9 個に対して 5 個の省略が得られ、また省略されていない要素に対して省略されることはなかった。さらに、橋本ら [14] 55 個の新聞記事に対して参照表現の分析を行い、それを元に参照表現選択モデルを構築し

4

た。このモデルでは人手によって 96%以上の精度を出せたと報告している。しかしゼロ代名詞化の選択においては不十分なモデルであり、焦点の主題へのなり変わりについての分析の必要性を指摘している。

また、飯田ら [3] 2449 個の新聞記事という大規模なデータを利用し、特に参照表現を省略すべきか否かという 2 値問題を解いた。共参照連鎖と、センタリング理論を拡張した Nariyama の手法 [2] を用いて参照表現の自動省略を行い、F 値で 0.550 の性能を出せたと報告している。彼らはさらに、評価データをあらためて人手で判断し、判断が一致した箇所を分析した。その結果、自動省略の性能を向上させるためには、対象を主題化するかしなにか、省略するかとい 3 値分類を扱う必要があることを明らかにした。

このように参照表現については多くの研究があるものの、機械学習を用いた述語項の表出形式の決定について扱った研究は少ない。緒方らが指摘しているように、文解析と比較して文生成に関する研究は極端に少ないのが現状である。しかし、今村 [4] が指摘しているように、「は」と「が」の使い分けに関する研究は言語学者を中心として多くなされており、その研究が参考になると考えられる。

原沢 [15] は、日本語能力試験のレベルで N4 から N3 である日本語の初中級学習者の作文 33 枚を分析し、学習者に共通する誤用について調査した。作文 1 の分量はおおよそ 300 字から 600 字である。その結果、文の構造に関する誤用 89 件中、述語と格助詞

との組み合わせに関するものが 33 件と非常に多いことを報告している。同時に、主題化に関する誤用が 16 件と格助詞に関する誤用に比べて少ないことを報告しているが、これは初中級レベルの日本語文においては主格が主題化されるものばかりであり、主語に機械的に「は」を付与すれば問題がないからであることを指摘している。

3

分析対象とするデータ

本実験の実施にあたっては、京都大学テキストコーパス^{*}(京大コーパス)と NAIST テキストコーパス[†](NAIST)を用いた。これらは毎日新聞95 1 1 から 17 日までの全記事約 2 万文、1 月から 12 月までの社説記事約 2 万文、計 4 万文に対して係り受け情報や述語の格関係などをアノテートしたコーパスである。

3.1 を用いて分析対象とするデータの説明を行う。

3.1. 戦争はすべて残虐であり、悲惨である。

3.1 には、京大コーパスによって図 3.1 のように係り受け情報と品詞情報が付与されている。それぞれの文節は一つまたは複数の形態素によって構成されている。この場合、「戦争は」は「悲惨である。」に係っており、「すべて」は「残虐であり、」に、「残虐であり、」は「悲惨である。」に係っている。

3.1: 京大コーパスによる係り受け情報

文節 ID	係り先の文節 ID	形態素の表記	形態素の品詞
-------	-----------	--------	--------

0	3	戦争	名詞
---	---	----	----

			助詞
--	--	--	----

1	2	すべて	副詞
---	---	-----	----

2	3	残虐であり	形容詞
---	---	-------	-----

			特殊
--	--	--	----

3		悲惨である	形容詞
---	--	-------	-----

			特殊
--	--	--	----

また NAIST によって図 3.2 のように述語の格関係情報が付与されている. np

*<http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス>. Accessed: 2016-01-30

†<http://sites.google.com/site/naisttextcorpus/>. Accessed: 2016-01-30 6

は名詞を, pred は述語を示す. 「残虐」のガ格項は「戦争」であり, また同時に「悲惨」のガ格項も「戦争」である.

3.2: NAIST による述語の格関係情報

<np id="1">戦争</np>はすべて<pred ga="1">残虐</pred>であり、
<pred ga="1">悲惨</pred>である。

まず NAIST を用いてガ格項と述語のペアを抽出し, 京大コーパスの係り受け情報を元にガ格項が省略されているか否かの情報を付与した. 具体的には, NAIST によって述語とされる語を含む文節 A に着目し, その述語に係るガ格項を含む文節 B が文節 A に係っているかどうかを京大コーパスを用いて確認した. ガ格項が直接係っていない述語は, ペアとなるガ格項が省略されているものとみなした.

3.1 では「残虐」と「悲惨」が述語であるが, それぞれのガ格項は「戦争」である. 「戦争」を含む文節である「戦争は」は「悲惨である。」に係っているため, 述語「悲惨」には「は」でガ格項に係っており, 述語「残虐」に係るガ格項は省略されている. ここで, 省略されている要素を括弧付きで復元すると例 3.2 のようになるが, このとき複数出現する「戦争」は同一の対象を指しており, この要素の列を共参照連鎖と呼ぶ.

3.2. 戦争はすべて残虐であり、(戦争が)悲惨である。

共参照連鎖が省略されずに表記される場合もあり, 例 3.3 においては NAIST によって

「橋本大二郎知事」と「橋本知事」が同じ実体を指すことが示されている。「橋本大二郎知事」に相当する共参照連鎖の要素は省略から復元したものを含めて6つである。

3.3. 高知県の「橋本大二郎知事」は三十一日、都道府県で初めて一般事務職の採用資格から国籍条項を（橋本大二郎知事ガ）撤廃する方針を明らかにした。高知県の現在の一般事務職の採用は日本国籍が要件。国は「公権力の行使や公の意思形成に携わる公務員は日本国籍が必要」との姿勢だが、地方公務員法では日本国籍がない人の任用を禁じる規定はない。橋本知事は「地方行政の運営上、国籍条項の必要性は（橋本知事ガ）感じない。少なくとも日本に生まれ育った在日韓国・朝鮮人を地方公務員として（橋本知事ガ）排斥する理由はない。戦後五十年を契機に（橋本知事ガ）実現させたい」と話している。

以上のように、述語単位でガ格項の表出形式データを作成し、分析対象データとした⁷。対象となる述語数は49,281件であり、約3割が省略されている。その⁸外界照応及び連体修飾は除く

7

の詳細を表3.1に示す。

3.1: 分析対象とするガ格項と述語のペア

「が」に係るもの	「は」に係るもの	省略されるもの	合計
19055	15517	14709	49281

8

4

SVMを用いた決定手法

4.1 手法の概要

ガ格項が述語に対して「が」に係るか、「は」に係るか、ガ格項自体が省略されるかという3値分類を行う提案手法について紹介する。

与えられた記事中の述語にかかるガ格項について、表出形式の決定を行う。表出形式の決定は記事の最初の述語から最後の述語の順に行う。文章を1文ずつ生成する状況を想定しているため、既出のガ格項の表出形式に関してはすべて正しく決定できているものとする。

例えば、例 4.1 において、「要件」を対象にガ格項の表出形式を決定するとき、「撤廃する」と「明らかにした」のガ格項の情報をういても良い。また、「明らかにした」のガ格項の表出形式を決定するとき、「撤廃する」のガ格項が省略されているという情報をういても良いとした。今回このような設定としたが、同文中のガ格項の表出形式情報は未知であるとする設定も考えられる。

4.1. 高知県の橋本大二郎知事は三十一日、都道府県で初めて一般事務職の採用資格から国籍条項を「撤廃する」方針を「明らかにした」。高知県の現在の一般事務職の採用は日本国籍が「要件」。

述語の特徴をもとに Support Vector Machine[1] (SVM) を用いて決定する。

4.2 分類に使用する素性

本研究では、分類に使用する素性として、大きくわけて形態素と述語の前後関係、共参照連鎖、述語の位置、表出形式の既出情報の4つの素性を使用した。使用した素性の一覧を表 4.1) に示す。以下では、各素性について詳細を説明する。

4.1: 利用する素性

素性グループ説明

G1.1 同文中における

述語より前に出現した形態素

.2 同文中における

述語より後に出現した形態素

.3 読点で区切った節中における

述語より前に出現した形態素

.4 読点で区切った節中における

述語より後に出現した形態素

G2.1 共参照連鎖内で直前までに何個省略されたか .2 共参照連鎖内で何番目か

.3 同文中の共参照連鎖内で最初か否か

.4 同文中の共参照連鎖内で最後か否か

.5 直前の共参照連鎖要素との間の他の述語数

.6 共参照連鎖内で最初かどうか

G3.1 同文中で最初の述語かどうか

.2 同文中で最後の述語かどうか

G4.1 同一の実体を示すガ格項が

直前に「は」として表出されているかどうか

.2 同一の実体を示すガ格項が

直前に「が」として表出されているかどうか

4.2.1 前後に出現した形態素(G1

形態素の表記情報と品詞情報、表記上述語より前に出現したか否かの情報の 組み合わせを用いる。複文の場合、従属節の従属度の強さによってガ格項の表 出形式が決まることが多く、従属節の強さに関係のある形態素である「なが ら」や「ので」などが分類に効果がある可能性がある。

4.2 では、／によって形態素を分けており、『』がガ格項、「」が述語を示している。「高い」は分析対象となる述語であるが、[“小沢”，述語が後] や、[“ 端に”，述語が前] という特徴を持つ。このとき、“極端に”(表記)の他に、“きょくたんに”(読み)や“極端だ”(原形)も特徴として用いる。また、品詞情報を 併せ持つものと持たないものの 2 種を使用する。

4.2. 小沢//新進党///『好感/は』/相当/高い///から/『好 感/は』/極端に/低い/という際立った『傾向』/

10

「出た」/

読点は文において大きな句切れ目を示す役割を持っており、読点で区切った ときに同じ節にある述語と形態素には密接な関係があると考えられる。そこ で、読点で区切ったときに同じ節にあるもののみを対象とした形態素の出現情 報も特徴にいった。

この素性は品詞を問わずすべての形態素を対象としている。しかし、名詞や 形容詞、動詞などは、ガ格項の表出形式への影響が少ないことと、出現数が少 なく分類として有用でないことが考えられる。そこで、すべての形態素を使用 するのではなく、機能語と呼ばれる以下の品詞に該当する形態素のみを使用した実験も行った。ただし、機能語の中でも接続 詞は、述語と同文節になりにくいため除いてある。

接尾辞

対象とする述語と同じ文節中の接尾辞を用いる。特徴として用いられる接尾 辞には、「れる」「られる」「いる」などが含まれる。

判定詞

人や事物がある属性をもつことを述べる文は主題をもちやすく、判定詞である「だ」「である」等が使われやすい。対象とする述語と同じ文節中の判定詞 を素性として用いる。「だ」

が対象となる。「である」「でした」「です」等に活用する。

助詞と助動詞

対象とする述語と同じ文節中の助詞と助動詞を用いる。助詞としては「や」「など」「ものの」「ばかり」等が、助動詞としては「のだ」「べきだ」「らしい」等が含まれる。

4.2.2 ガ格項の共参照連鎖(G2)

同じ実体を指す要素の連続(共参照連鎖)の情報を素性として利用する。提案素性を例 4.3 を用いて説明する。

4.3. 『村山富市首相は』二十八日「会見し」、見通しを「示した」。また、衆院解散・総選挙の可能性に否定的な見解を「表明」、内閣改造を明確に「否定した」。

11

出現する述語に共通するガ格項である「村山富市首相」は、「示した」に直接係っている。「会見し」「表明」「否定した」のガ格項はいずれも省略されている。ここで「否定した」に注目すると、直前までに「表明」に係るガ格項の 1 個が省略されている。そのため G2.1 素性に相当する特徴は 1 となる。この省略要素数が多いと、参照要素が談話の中心になっており、以降の文においてもより省略されやすいと考えられる。

4.2.3 文中の述語の位置(G3)

文中において、最初の述語であるか否かという素性と、最後の述語であるか否かという素性を用いた。

4.4 において、「示した」は最後の述語であり、『村山富市首相』が係っている。主語が「は」で示されるとき、文の最後の述語に係りやすく、同文中で最後でない述語に係るガ格項は省略されやすい。

4.4. 『村山富市首相は』二十八日「会見し」、見通しを「示した」

また、主語が「が」で示されるとき、直近の述語に係りやすい。例 4.5 において、『男性』は「取り出して」に係っており、「笑っていた」のガ格項は省略されている。

4.5. 初老の『男性が』仮設住宅申し込み証明書を財布から「取り出して」見せながら寂しそうに「笑っていた」

この素性はそれらの特徴を捉えるために導入した。

4.2.4 表出形式の既出情報(G4)

既出のガ格項の表出形式を用いて、ガ格項が示す実体の顕現性の特徴を利用した。直前に「は」ないし「が」で係るガ格項の実体が、表出形式を決定する対象と同一のものであるか否かという素性を用いる。

4.6 において、「姿勢だ」のガ格項は『国』であるが、直前に「は」で係ったガ格項は『橋本大二郎知事』であり、ガ格項の実体は同一ではない。しかし、「禁じる」に注目すると、直前に「は」で係ったガ格項は『国』であり、ガ格項の実体は同一である。ある述語に注目したとき、直前に同一のガ格項の実体が「は」で係っている場合、その述語のガ格項は省略されやすく、その特徴を捉えるためにこの素性を導入した。

4.6. 高知県の 橋本大二郎知事は 三十一日、都道府県で初めて一般事務職の採用資格から国籍条項を (橋本大二郎知事が) 撤廃する方針を明らかにした。

12

高知県の現在の一般事務職の採用は日本国籍が要件。国は「公権力の行使や公 の意思形成に携わる公務員は日本国籍が必要」との「姿勢だ」が、地方公務員 法では日本国籍がない人の任用を (国が)「禁じる」規定は ない。

実験と考察

5.1 実験設定

実験では、京都大学テキストコーパス及び NAIST テキストコーパスを利用し、第 3 章で述べた方法によって対象となるガ格項と述語のペアを抽出した。抽出したデータを SVM の学習に用いる訓練用、SVM のパラメータ調整に用いる開発用、素性の評価に用いる評価用の三種に分けた(表 5.1)。

5.1: 対象となる述語数

	総数	「が」	「は」	省略
訓練用データ	2912	1129	9235	859
開発用データ	3	3	1657	5
評価用データ	5339	2112	4625	157
	1481	5650		0
	9			454
				4

5.1.1 比較対象

また、提案手法の有用性を調べるために機械学習を用いない決定も行った。この決定結果をベースラインとする。ベースラインは二つの戦略を用いてガ格 項の表出形式を決定する。

共参照連鎖を用いた省略判定

共参照連鎖の要素のうち、最初に出現する文の要素以外は省略。最初に出現する文においては、最後の述語にかかる要素以外を省略する。まず、共参照連鎖を用いて省略されるか否かを判断し、省略されなかったものに対して「は」と「が」のどちらで係るかを判断する。

5.1 において、共参照連鎖が最初に出現したのは 1 文目であるから、2 文目以降の述語、「表明」と「否定した」に係るガ格項は省略と判断。1 文目では「会見し」に係るガ格項

を省略と判断する.

5.1.『村山富市首相は』二十八日「会見し」, 見通しを「示した」. また、衆院解散・総選挙の可能性に否定的な見解を「表明」, 内閣改造を明確に「否 定した」.

出現位置を用いた「は」「が」決定

同一文中では, 最初に出現する述語には「は」, その他の述語には「が」が 係ると判断する.

5.2 において, 最初の述語「攻撃」にはガ格項は「は」で係り, 以降の述語「炎上し」には「が」で係ると判断する.

5.2.『ロシア軍は』装甲車などで「攻撃」, 大統領官邸など『数力所が』「炎 上し」た.

5.2 決定実験とその結果

SVM ライブラリとして LIBLINEAR* を使用した. 表 5.2 にベースラインと提 案手法の正解率を示す. 正解率は, 表出形式が正しく決定されたガ格項数を, 評価データ総数で割ったものの百分率である.

参考として, 旧情報を示す「は／には」, 新情報を示す「が／を／に」, 及び「省略」の3値分類を手で行った正解率を併記する. これは飯田ら[3]によって 調査された, ガ格項, ヲ格項, 二格項の三つを合わせた場合の分類正解率の算 術平均である. NAIST において, ガ格項は約 11 万件であるのに対し, ヲ格項 は約 4 万件, 二格項は約 2 万件であり, ガ格項が 6 割を占める [6]. また, ガ格項 のうち省略されているものは約 58%であるが, ヲ格項のうち省略は約 16%, 二 格項のうち省略は約 9%であるから, ガ格項に比べてヲ格項, 二格項では省略するか否かという問題正解率は高くなり, 3 値分類正解率も高いと考えられる. 以上のことから, この正解率は, ガ格項のみを扱ったときの正解率に非常に近 いか, もしくは高いものと考えられ, 手法を改善したとしてもガ格項のみの分 類正解率はこの値を超えないものと考えられる.

*<http://www.csie.ntu.edu.tw/~cjlin/liblinear/> Accessed: 2016-01-28 15

5.2: 結果

	正解率
--	-----

すべて「が」と判断	38.1%
ベースライン	50.8%
提案手法	70.8%
* 人手で判断	82.9%

5.2.1 提案手法による実験

提案手法と、それぞれの素性グループを抜いたものの比較を表 5.3 に載せる。McNemar 検定によって提案手法に対する p 値を算出している。有意水準 0.05 元で、G1+G2+G3 以外はすべて提案手法に対して有意に精度が低下しており、特に対象となるガ格項以外の形態素を用いた G1 と、ガ格項の共参照連鎖を用いた G2 が素性として有用であるとわかった。

ただし、G4 のみを用いた決定であってもランダム決定の正解率(33.3%)よりも 13% 以上高い正解率を出しており、この素性が全く有用でないとは言えない。G4 の特徴が、同時に素性にいれた他の素性グループの特徴に含まれているために、提案手法から G4 を除いても有意差が出ないと推測される。

5.3: 素性の組み合わせを変えたものの比較

	正解率	提案手法に対する p
G1+G2+G3+G4 (提案手法)	70.8%	
G2+G3+G4	65.7%	0.000
G1+ G3+G4	60.8%	0.000
G1+G2+ G4	70.5%	0.014
G1+G2+G3	70.8%	1
G4	46.6%	

ここで、特に素性として有用であった G1 G2 の特徴の差を比較するために、G2 のみを素性として用いた手法をベースラインとし、G1+G2 を素性として用いた手法と比較した(表 5.4)。マトリクス表において、例えば縦の「が」の項目は、ガ格項が「が」で係る述語の数を示している。

詳細のマトリクス表や、すべて「が」で係ると判断したときの正解率を見ると、G2 は省略と非省略の分類に比べて「が」と「は」の分類は苦手であることが分かる。また、G2 G1+G2 では省略するか否かという分類では 0.2%の差で

あるが、「が」と「は」の分類では 8.7%の差がでており、G1 が「が」と「は」の 分類に有用であることが分かる.

5.4: G1 G1+G2 の比較

	G2 65.5%	「が」	「は」	省略	
	「が」と予想 「は」と予想 省略と予想	4711 331 608	2995 954 676	292 210 404 2	
	G1 + G2 70.4%	「が」	「は」	省略	
	「が」と予想 「は」と予想 省略と予想	4049 1113 488	1511 2508 606	375 289 388 0	
		省略と非省略		「が」と「は」	
すべて「が」と判断		69.3%		55.0%	
G2 を用いた分類器		87.9%		55.1%	
G1 + G2 を用いた分類器		88.1%		63.8%	

5.2.2 形態素を用いた素性の詳細な実験

G1の素性の代わりに、対象となる述語と同じ文節の接尾辞、判定詞、助詞や 助動詞を特徴として用いた実験を行った。これらの品詞の語は機能語と呼ばれ る。提案手法と、G2+G3+G4+それぞれの素性、さらにG2+G3+G4の比較を表 5.5 に示す。提案手法に対する p 値はすべて 0.001 以下であったため、省略した。

5.5: 一部の機能語を用いた素性による実験

	正解率	特徴数
G1 + G2 + G3 + G4 (提案手法)	70.8%	38549 9

G2 + G3 + G4 + 接尾辞 + 判定詞 + 助詞 + 助動詞	69.4%	386
G2 + G3 + G4 + 接尾辞	67.9%	278
G2 + G3 + G4 + 判定詞	66.8%	118
G2 + G3 + G4 + 助詞	66.8%	170
G2 + G3 + G4 + 助動詞	66.1%	126
G2 + G3 + G4	65.7%	117

17

この結果を見ると、G1 の代わりに機能語を用いた手法の正解率は、提案手法にわずかに劣るものの、素性として有用であることがわかる。提案手法の素性に用いられた特徴数は385499 個であり、G2+G3+G4+接尾辞+判定詞+助詞+ 助動詞に用いられた特徴数は 386 個であった。

この手法は提案手法に比べて、精度をあまり落とさずにはるかに特徴数を少なくすることができ、なにが格項の表出形式を決めてるか分かりやすくなったと言える。

G2+G3+G4+接尾辞+判定詞+助詞+助動詞を素性として用いたときの SVM によって決定された重みのうち、G2 G3 G4 によるものを除いた上で代表的なものを表 5.6 に示す。

5.6: 同文節中の形態素の重み

品詞	原形		が」	省略
判定詞		0.40 4	-0.45 3	0.0
助動詞	まい	0.57	-0.63	0.03
	ばかりだ	3	8	6
	ことだ	0.50	-0.40	-0.27
	わけだ	8	2	6
		-0.38	-0.18	0.54
		2	6	2
		-0.42	0.386	-0.00
		1		7
助詞	なら(副助詞) なら(接続助詞)	-0.41	-0.59	0.87
		-0.25	5	2
		8	0.656	-0.61
		-0.62	0.689	7
	から	6	-0.10	-0.30
		0.38	5	3
				-0.42

		0.58 1	-0.55 5	9 -0.03 6
		-0.62 4	0.547 -0.35	0.03 9
		0.45 5	6	-0.17 6
接尾辞	ある	0.31 6	-0.50 1	0.11 9
	さん	0.41 6	-0.45 5	0.04 1
	られる	0.75 5	-0.62 5	-0.21 9
	させる	-0.41 3	-0.12 1	0.45 7
	おく	-0.02 2	-0.44 2	0.39 0.02
	れる	0.42 4	-0.41 2	0.50 6
		-0.62 5	-0.05 2	

この重みリストを見ればどの形態素が分類に有効かがわかりやすい。例えば、例 5.3 のように、判定詞「だ」を含む文節中の述語のガ格項は「は」で係りやすい。一般に人や事物がある属性をもつことを述べる文は主題をもちやす

18

い。しかし、新聞特有の表現が含まれている可能性があるため、一般化して考えるのは危険である。

5.3. 田中さんは親切【】

結論

本稿では、新聞記事を対象としたガ格項の表出形式の自動決定課題を解く手法を提案した。特に、ガ格項が述語に係るときに主題化して「が」に係るか、主題化せず「は」に係るか、それともガ格項自体が省略されるかという 3 値分類の問題を扱った。この手法の性能を調査した結果、70.8%の正解率を得た。

個々の素性を比較した結果、共参照連鎖を用いた素性は省略するか否かという課題に向いており、形態素を用いた素性は主題化するか否かという課題に向いていることが分かった。

また、提案手法の素性を見直し、品詞によって使用する形態素を絞った実験をあらためて行い、正解率で 69.4%という結果を得た。この実験によって、述語と同節中の接尾辞、判定詞、助詞や助動詞といった機能語が課題を解く上で重要な特徴であることが分かった。

今回の実験ではガ格項のみを扱ったが、自然な文生成にあたってはヲ格項やニ格項にまで拡張した研究が重要である。また、対話システムへの応用を考えると、書き言葉ではなく話し言葉を対象とした研究が重要になる。対話の文体は新聞の文体と大きく異なり、有用である素性も異なってくると考えられ、今後の課題となる。

謝辞

本研究にあたっては笹野遼平助教より多くのご助言と丁寧なご指導を頂戴した。また、高村

大也准教授, および奥村学教授より様々なご助言を頂戴した. 記して感謝する.

参考文献

- [1] C. Cortes, V. Vapnik, “Support-vector networks.” *Machine learning*, 20, 3, pp. 273–297, 1995
- [2] S. Nariyama, “Grammar for Ellipsis Resolution in Japanese.” *Proceedings of the 9th International Conference on Theoretical and Methodological Issues*

in Machine Translation, pp. 135–145, 2002

- [3] 飯田 龍, 徳永 健伸, “日本語書き言葉を対象とした参照表現の自動省略 - 人間と機械処理の省略傾向の比較-”, 情報処理学会研究報告自然言語処理 NL , 2012-NL-206, 15, pp. 1–10, 2012
 - [4] 今村一子, “「は」VS「が」と日本語教育 ”, 信州大学留学生センター紀要, 3, pp. 45–59, 2002
 - [5] 緒方健人, 佐藤理史, 松崎拓也, “文節木の段階的実体化による日本語文生成 器の作成”, 人工知能学会全国大会論文集, 29, pp. 1–4, 2015
 - [6] 河原大輔, “言語処理基盤としての言語資源 ータグ付きコーパス, 生コーパ, そして新聞記事からウェブへー”, 言語処理学会 第 13 回年次大会 チュートリアル, pp. 23–40, 2007
 - [7] 佐藤理史, “「文生成器を作る」とはどういうことか”, 言語処理学会 第 21 回年次大会 発表論文集, pp. 1080–1083, 2015
 - [8] 新山 祐介: 形態素解析ツールの品詞体系, <http://www.unixuser.org/~euske/doc/postag/>, 2003, Accessed: 2016-01-30.
 - [9] 高村大也: 言語処理のための機械学習入門, 奥村学監修, コロナ社, 2010 [10] 田中穂積: 自然言語処理 ー基礎と応用ー, 電子情報通信学会, 1999
 - [11] 中岩浩巳, 池原悟, “語用論的・意味論的制約を用いた日本語ゼロ代名詞の 文内照応解析”, 自然言語処理, 3, 4, pp. 49–65, 1996
 - [12] 滑川裕樹, 乾健太郎, 徳永健伸, 田中穂積, “照応・省略を含む日本語論説文 生成”, 言語処理学会第5回年次大会, 3, pp. 153–156, 1999
- 22
- [13] 日本語記述文法研究会: 現代日本語文法 5, くろしお出版, 2009
 - [14] 橋本さち恵, 乾健太郎, 白井清昭, 徳永健伸, 田中穂積, “日本語文生成に おける照応表現の選択”, 情報処理学会研究報告自然言語処理(NL , 2001-NL-143, 54, pp. 33–40, 2001
 - [15] 原沢伊都夫, “日本語初中級学習者の作文指導 : 学習者の誤用分析をもと ”, 静岡大学国際交流センター紀要, 6, pp. 79–92, 2012

