検索エンジンを用いた主格省略文の自動判定

Automatic detection of subject ellipsis using a search engine

[*]中村慶太・[**]北村達也・[***]川村よし子
[*]NAKAMURA Keita・[**]KITAMURA Tatsuya・[***]KAWAMURA Yoshiko
[*&**]甲南大学・[***]東京国際大学
[*&**]Konan University・[***]Tokyo International University
〒658-8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail [*]keita264@gmail.com・[**]t-kitamu@konan-u.ac.jp・[***]kawamura@tiu.ac.jp

Abstract：In this study, we propose a system for detecting subject ellipsis. Subject ellipsis is known to make sentence comprehension difficult for non-native speakers. This system first extracts a noun which may be the subject and the predicate from an input sentence by using a dependency parser. Next, the system makes two types of sentences with the extracted noun and predicate: one with the particle *ga* and the other with the particle *wo*. Then it compares the frequency of each sentence in Web texts using a search engine to judge whether the extracted noun is the subject of the predicate. Based on the results, it can be determined whether the noun is the subject or the subject has been deleted.

キーワード：難易度，主格省略，構文解析，検索エンジン，用例検索

1　はじめに

我々は日本語非母語話者にとって難しい構文を検出し，彼らにとってわかりやすい文を書くことを支援する技術を開発している．日本語非母語話者が難しく感じる構文を抽出することは日本語教師にとっては比較的容易であるが，一般の日本語母語話者にとっては難しい．川村ら(2011)は文の難易度を高める傾向にある構文として，ゼロ格(特に主格省略)，ハ・ガ構文，中止法(テ形および連用形)，名詞修飾節などを挙げている．本研究では，このうち主格省略の自動検出システムの開発を行った．提案法では文中の述語に係る名詞が主格であるか否かに関して検索エンジンを利用して判定する．

2　主格省略の自動検出

2.1 処理の流れ

自動検出の流れは次のとおりである．まず，構文解析システム KNP(黒橋，2000)により入力文の係り受け構造を得る．次に，文中の述語に係る名詞とそれに続く助詞を列挙する．その助詞が主格を取り得るものだった場合，その名詞が当該の述語の主格になり得るかを検索エンジンを用いて判定する．本研究では，主格を取り得る助詞として「が、ぐらい（くらい）、こそ、さえ、しか、だけ、でも、は、ばかり、ほど、まで、も」の 12 種を選択した．

2.2 一般の動態述語の場合

動態述語については，一般のものと「できる」など可能を表すものに分けて処理を行う．

一般の動態述語の場合は，文中の当該の名詞と動詞をガ格でつないだ文「(名詞)が(述語)」とヲ格でつないだ文「(名詞)を(述語)」とを検索エンジンで検索する．述語は基本形とする．例えば，「私は読んだ。」という入力文の場合，主格を取り得る助詞「は」とつながっている名詞は「私」なので，「私が読む」と「私を読む」を検索し，それぞれの検索数を比較する．ガ格でつないだ検索文とヲ格でつないだ検索文の検索数（それぞれ N_{ga}，N_{wo} とする)を比較し，以下の規則を用いて判定する．

- 規則 1： $N_{ga} < \alpha$ かつ $N_{wo} < \alpha$ ならば判定不能
- 規則 2： $(1+\beta)N_{ga} < N_{wo}$ ならば，主格にならない
- 規則 3： $(1-\beta)N_{ga} < N_{wo}$ かつ $(1+\beta)N_{ga} \geqq N_{wo}$ ならば主格になる可能性がある
- 規則 4： $(1-\beta)N_{ga} > N_{wo}$ ならば，主格になる

ここで α と β は変数であり，これらを調節することによって判定が変わる．なお，β は 1 以下の値をとる．ただし，「私は本を読んだ」のように述語にヲ格の文節が係っている場合，上のような検索を行わなくても「私は」が主格であると判定できる．

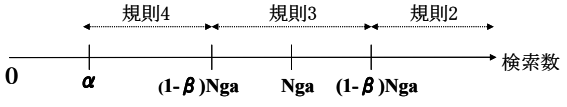

図 1．判定基準

規則1は検索数が少なすぎる場合結果が信頼できないためそれへの対応策である．規則2において，単純に $N_{ga}<N_{wo}$ としないのは N_{ga} と N_{wo} が拮抗している場合も信頼性が高いとは言えないためである．

規則3では N_{ga} と N_{wo} が拮抗している場合，曖昧な判定をさせている．このような場合，利用者に判断をゆだねることになる．ただし，インターネット上のテキストデータは日々増加しているので，時間の経過とともに判定精度の向上が期待できる．規則4は N_{wo} が N_{ga} の $(1-\beta)$ 倍より小さい場合に，当該の名詞が当該の述語に対し主格であると判定するものである．

以上の処理を上記の「主格を取りうる助詞」が含まれるすべての文に対して実行する．規則1または規則3と判定された場合にはガ格と述語およびヲ格と述語の間に単語が入る可能性を考慮して改めて検索を行う．この場合，ガ格の検索文は「(名詞)が*(述語)」であり，ヲ格の検索文は「(名詞)を*(述語)」である．「*」はワイルドカードと呼ばれる記号で，検索において任意の形態素を意味する．この場合も各検索文の検索数に対して上記の規則を適用して判定する．本研究では追加する形態素の上限を2個とした．

「主格を取り得る助詞」が含まれる文節のいずれかが主格になると判定されれば，当該述語の主格が文中に存在すると判定する．

2.3 可能を表す動態述語の場合

可能を表す動態述語の場合，2.2節で示した方法では主格の有無を判定できない．そこで「(名詞)は*が(述語)」という文を検索する．検索数が閾値 γ を超えた名詞はその動態述語の主格であると判定する．閾値 γ の値が低いと，インターネット上の誤用による判定誤りを生じる可能性が増えるため，比較的大きい数値に設定する必要がある．

2.4 状態述語の場合

一般の状態述語の場合，前述の主格を取り得る助詞が含まれる文節が係っている場合には主格有り，係っていない場合には主格なしと判定する．ただし，「好き，嫌い，上手，下手，欲しい，得意，苦手，専門」を含む文においては，上記の規則では不都合が生じる．このような場合には，2.3節と同様に検索エンジンを用いて当該の文節の名詞が主格となるか否かを判定する．検索文は「(名詞)は*が(述語)」とした．そして，その検索数が閾値 γ を超えた名詞はその状態述語の主格であると判定する．

また，「私は嫌い。」のような文の下線部が主格をとるか否かを判定するのは非常に難しい．そこで，このような状態述語に係っている文節が1つの場合には，「主格または目的格が省略されている」と判定することにした．こうした文に対する主格省略判定は将来の課題とする．

3 評価実験

提案法を評価するため実験を行った．

3.1 方法

朝日新聞社の2011年9月26日分の社説「新しい公共の世紀へ―市民の力で社会を変える」を対象にした．

人による判定では，日本語を母語とする大学生2名に対し，対象となる述語39個に印をつけた文章を与え，主格の有無を判定させた．2名の判定が異なる場合は著者らが議論して決定した．提案法による主格省略では，変数を $\alpha=100$，$\beta=0.25$，$\gamma=10000$ と設定した．構文解析の誤解析により正しい係り受け構造が得られない場合には，評価対象から除外した．

3.2 結果

評価対象となった述語は39個中33個である．人と提案法とで判定が一致したのは28個(85%)であった．残りは判定不能と判定されたものが2個(6%)，判定誤りが3個(9%)であった．

4 おわりに

本研究では検索エンジンを用いた主格省略判定システムの開発を行い，実験を通して精度を検討した．今後，より多くの文章を対象にした評価実験を継続し，誤判定の分析によって精度向上のための改良を行っていく予定である．

謝辞 本研究の一部は平成23年度科学研究費課題番号21320095及び平成23年度私立大学等経常費補助金の支援を得て行われた．

参考文献

川村よし子・前田ジョイス・保原麗・川村ヒサオ（2011）「文章の難易度判定システム構築のための基礎調査」『ヨーロッパ日本語教育』15号，171-178.

黒橋禎夫(2000)「結構やるな，KNP」『情報処理』41(11)，1215-1220.