

古典文の係り受け解析への経験的優先規則の適用

4 J-10

齋藤 大輔[†] 石隈 博史^{*} 小川 泰幸[†] 江畑 秀規[§] 上原 徹三[†] 石川 知雄[†]武蔵工業大学大学院工学研究科電気工学専攻[†] 日立ソフトウェアエンジニアリング(株)^{*}
日本情報通信(株)[†] 日本光電工業(株)[§]

1 はじめに

平安時代の作品である伊勢物語を処理の対象として、意味解析に入りこまない範囲で、制約規則と経験的優先規則を適用して古典文の係り受け解析を行った。[1]

2 係り受け解析

係り受け解析は、前もって伊勢物語の解釈文[2]を形態素解析したものを入力情報として与えて、代表品詞と下接特性の設定処理、係り受け候補の決定処理、優先度の設定処理の3つの処理を行うことから成り立っている。前回の発表[1]では係り受け候補の決定処理での制約規則を主に述べたので、今回は優先規則を中心として述べる。

2.1 入力情報

入力情報は、まず伊勢物語の解釈文を文節ごとに区切り、文献[2]を参考に文節の単語ごとにその単語の読みと表記と品詞と活用形を示したものである。入力情報で用いている品詞は学校文法で用いられている品詞を中心とした11種類である。但し、名詞、副詞、助詞については複数に分類している。

2.2 代表品詞と下接特性の設定処理

ここでは、与えられた入力情報を基にして、文節ごとに、どのような受け特性を持つかを表す代表品詞と、どのような受け特性を持つ文節に係り受け可能かを表す下接特性を設定する[3]。

Dependency Analysis of Japanese Classics Sentence based on the Heuristic Priority Rules

[†]D.Saitou, T.Uehara, T.Ishikawa

[†]Y.Ogawa

[§]H.Ebata

^{*}H.Ishiguma

[†]Electrical Engineering, Graduate School of Research Division in Engineering, Musashi Institute of Technology

2.2.1 代表品詞の設定

ある文節の代表品詞の設定は、自立語の品詞により設定する。代表品詞の種類は、それぞれ、「体言」「用言」「副詞」「連体詞」「無し」(無しは、感動詞・接続詞の場合で受け文節とはならない)の5種類である。

2.2.2 下接特性の設定

ある文節の下接特性の設定は、その文節の最後の単語の品詞と活用形により設定する。単語の品詞によっては複数の下接特性をもつ場合もある。下接特性の種類は「体言」「用言」「終止」の3種類である。

2.3 係り受け候補の決定処理

ここでは、文節ごとに定めた代表品詞と下接特性を用いて係り受け候補の決定を行う。すなわち、係る文節の下接特性と受ける文節の代表品詞が適合した場合は、両者の係り受け関係が成立すると判断される。しかし、それだけでは候補数が膨大なものとなるため、係り受けの交差はない、文節の係り先は一つ、後ろの文節が前の文節に係ることはないという3つを絶対条件として、また、経験的に定めた係り受けの制約規則(9種類)を用意して係り受けの関係に制約を加えた。

2.4 優先度の設定処理

係り受けの決定処理を終えた段階でも、文候補は複数存在する。それで、選出されたすべての文候補の間で、係り受けの妥当な文を判定するために、優先度を設定する。設定方法は、各文節の係り先すべてに対して、ある規則(11種類)に従って評点を与える。そして、各文候補の評価点を求め、文候補の間で優先度を設定する。以下その主な規則について述べる。

1. 直結文節の評価

係り受け候補の決定処理の規則では、ある条件を満たす場合に限り、ある文節が構文上すぐ次の文節に係り得るのならば、その係り受け関係を採用するという制約規則を用いた。この直結文節が構成された場合の優先度を高く設定する。

2. 文節末が接続助詞の文節の評価

「接続助詞」は、それまでの叙述を下の叙述に接続するはたらきをして、述語になる場合が多い。よって、「接続助詞」が文節末にある文節では、係り先の文節の文節末に接続助詞がある場合と下接特性が「終」の文節に係る場合に、前の述語が、後ろの述語に係るということで優先度を高く設定する。

3. 読点「、」で終わる文節の評価（読点文節）

解釈されにくい曖昧性のある文では、書き手自身が読点を入れる傾向がある。また、読点文節は独立性が強く、それだけ遠くに係る可能性も高い。そこで、次のような優先規則を設ける。

- 係り先がいくつかの候補の中で直近の読点文節の場合
前の事柄を説明するような文節の場合には、その係り先は近くの文節にくる場合が多い。そこで、直近の読点文節にかかる場合に優先度を高く設定する。
- 係り先がいくつかの候補の中で直近で、文節末が接続助詞の読点文節の場合
この場合には、係り先がそこまでの文の述語になる可能性が高い。そのような場合には優先度を高く設定する。
- 係り先の文節の下接特性が「終」の場合
この場合には、そこで意味が切れ、係り先が述語になる。そこで、加点することで優先度を高く設定する。

4. 距離による評価

経験的に、単純な文において係り先が複数存在する場合には、なるべく近くに係るといわれている。よって、読点が無い文節の場合には、近くの文節に係るほど優先度を高く設定する。

3 解析結果

伊勢物語の1段から66段までの歌を除いた176文について解析を行った。解析の結果と予め作成

した正解を比較した結果は表1~2のようになった。但し、予め作成した正解では、人手によっても、どの文節に係るかを限定できない場合にはどちらの文節に係っても正解としてある。

表1 係り受け候補の設定処理(176文)の結果

文節数	解析対象文数	解析に成功した文数	文候補数の平均	解析成功率
1~9	129文	114文	3.7文候補	88%
10~14	25文	17文	37.3文候補	68%
15~	22文	18文	902.9文候補	82%
全体	176文	149文	116.2文候補	85%

表2 優先度の設定を行った149文の結果

文節数	解析対象文数	正解が1位の文数	正解が1位の確率
1~9	114文	100文	88%
10~14	17文	10文	59%
15~	18文	7文	39%
全体	149文	117文	79%

結果を見ると、文全体での、係り受け候補の決定処理の解析率は85%、優先度の設定をして正解順位が1位となる確率は79%と両方とも高い率になっている。しかし、文節数毎の結果を見ると、表1では文候補数の平均は文節数が増えると極端に平均値が大きくなっている。また解析率も文節数が増えると、文節数が1~9のときに比べて悪くなる傾向にある。表2でも文節数が増えると結果は悪くなっており、15文節以上の場合の優先度の順位が1位となる確率は、5割を下回っている。

このように、文節数が増えると(長文になると)結果が悪くなる傾向にあるのは、長文では文構造が複雑になり、そのような長文ではあまり成り立たないような規則を、係り受け候補の決定処理、優先度の設定処理で用いているためだと思われる。

今後の課題としては、長文での解析結果の問題点を検討して、長文での解析結果の向上を計って行きたい。

なお、本研究は一部、平成7年度科学研究費補助金(一般研究C)「日本語文語文のコンピュータによる処理」によって進めたものである。

参考文献

[1] 齋藤, 他: “経験的優先規則による古典文の係り受け解析”, 情報処理学会第54回全国大会論文集, 1997.
 [2] 木村雅則: 伊勢物語翻刻・解釈ファイル, 勉誠データセンター, 1993
 [3] 三ツ木 徳彦: 伊勢物語精釈, 中道館, 1994
 [4] 兵藤 安昭 池田 尚志: “表層的情報とN近傍ブロック化手法による日本語長文の骨格構造解析”, 情報処理学会論文誌, Vol36, No9, pp2091-2101 (1995).