留学生の日本語助詞修正システム

三浦雅則†　横山晶一‡

山形大学大学院理工学研究科

1. はじめに

本研究では、留学生の日本語助詞誤用について修正を行うシステムを作成する。

現在、日本語学習者に対する指導は、日本語教育者が直接行っている。しかし、教育には相応の時間と労力を要するため教育者の負担が大きい。

そこで、日本語教育をシステム化することができれば、大量のデータを扱えるうえ、一定の基準をもって判定し、添削できる。これは学習者にとっても有用となる。

このような自動修正を目的として、留学生の助詞誤用を検出、修正するシステム[1,2]が提案されている。

石川らの手法[1]では、人手によりルールを構築し、前処理を行い、日本語語彙大系[3]を用いた格フレームによる照合処理を行う。このシステムでは、入力を単文に限り、誤用は一箇所だけとするという制約がある。

南保らの手法[2]では、文節の特徴から機能的学習を用いて検出および自動校正を行う。この手法では、格助詞以外の助詞を扱えるという特徴がある。

本システムでは、大規模格フレーム[4]を用いることによりこの問題の解決を試みた。

実験の結果、先行システムより検出・修正ともに性能の向上がみられた。

2. 研究内容

2.1 格フレームによる格助詞誤用修正システム

本システムの処理過程を以下に示す。

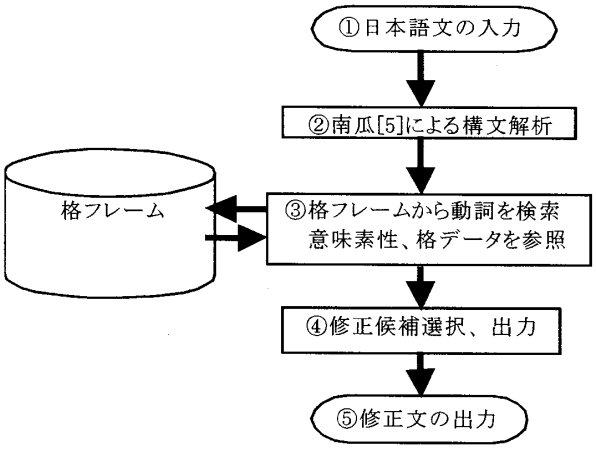

図1　システムの処理過程

2.2 大規模格フレーム[4]

河原らが、約4億Webページから約5億文からなるテキストコーパスを作成し、構築した約5万用言からなる格フレームである。格フレームとは、人間のもっている常識的な知識のうちもっとも基本的なものである。これにより、用言の取り得る助詞と、その助詞の取り得る名詞を判断することができる。

このデータの名詞から意味属性をとり修正を行う。

積む:動1
〈ガ格〉人:11, 者:6, 選手:4, …
※〈ヲ格〉経験:10150, 体験:414, 知識:39, …
〈デ格〉会社:22, 現場:22…

2.3 システム適用例

「家の居間で木をあります。」という入力文を例に、本システムの処理過程を説明する。

①日本語文の入力

入力文:家の居間で木をあります。

②南瓜による日本語解析

日本語係り受け解析器「南瓜」を使い入力文の係り受け関係と単語単位での品詞情報を取得する。すると次の結果が得られる。

```
＊ 0 1D(文節番号　係り先番号)
家　　イエ　家　　名詞-一般
の　　ノ　　の　　助詞-連体化
＊ 1 3D
居間　イマ　居間　名詞-一般
で　　デ　　で　　助詞-格助詞-一般
＊ 2 3D
木　　キ　　木　　名詞-一般
を　　ヲ　　を　　助詞-格助詞-一般
＊ 3 -1O(最終節)
あり　アリ　ある　動詞-自立五段・ラ行連用形
ます　マス　ます　助動詞特殊・マス基本形
。　　。　　。　　記号-句点
```

③格フレーム情報の検索

②の情報を元に用言と体言を検索する。この例では、「ある」という動詞に対して「居間」「木」という2つの名詞が検索対象になる。

格は取り得る格助詞の候補を示す。また、係り受け、頻度のデータが付与されている。

・「居間」の意味属性

[1 名詞/2 具体/533 具体物/706 無生物/760 人工物/863 建造物/864 家屋/866 家屋(部分)/867 家屋(部分〈場〉)/868 部屋]

・格:ガ closest, 位置:3, 頻度:5861, 自:1, 係先:3D
・格:ニ closest, 位置:3, 頻度:9199, 自:1, 係先:3D
・格:デ closest, 位置:3, 頻度:335, 自:1, 係先:3D

「居間」に対して、以上の格候補が挙げられる。

・「木」の意味属性

[1 名詞/2 具体/533 具体物/534 生物/671 植物/672 植物(固体)/673 樹木]

[1 名詞/2 具体/533 具体物/706 無生物/760 人工物/769 資材/771 木材/772 材木]

[1 名詞/1000 抽象/2422 抽象的関係/2670 時間/2671 暦日/2678 年月日/2681 週]

・格:ガ closest, 位置:5, 頻度:4820, 自:2, 係先:3D
・格:ニ closest, 位置:5, 頻度:867, 自:2, 係先:3D
・格:デ closest, 位置:5, 頻度:9, 自:2, 係先:3D

Japanese Particle Correction System for Foreign Students
†MIURA Masanori, Yamagata University
‡YOKOYAMA Shoichi, Yamagata University

「木」に対して、以上の格候補が挙げられる。

④修正候補助詞の出力

格候補の選択を頻度情報から行い、下線で示された格がそれぞれ選択される。それを最終的な出力とする。

出力文:

家の居間で{ガ:5861}**{ニ:9199}**{デ:335}{無:49}{ト:30}{マデ:41}{ヲ:8}木を**{ガ:4820}**{ニ:867}{無:175}{ト:41}{マデ:41}{ヨリ:10}{デ:13}{ヲ:12}{カラ:5}あります。

⑤修正文の出力

出力文:家の居間に木があります。

3. システムの実験・評価

3.1 実験データと実験条件

実験データは、国立国語研究所の日本語作文データベース[6]から、出身地域、学習期間などを考慮せずに、正文、誤文を含んだ単文のみ 200 文を用いた。この中には、複数箇所誤りを含む文も含まれる。

3.2 評価方法

本システムを評価するために次式で定める再現率、精度、F 値を用いた。

(a)検出を評価する尺度

$$再現率 = \frac{正しく検出できた誤用数}{助詞誤用数} \quad (1)$$

$$精度 = \frac{正しく検出できた誤用数}{システムが誤用と判断した数} \quad (2)$$

(b)修正を評価する尺度

$$再現率 = \frac{正しく修正できた誤用\ 数}{助詞誤用数} \quad (3)$$

$$精度 = \frac{正しく修正できた誤用\ 数}{システムが誤用と判断\ した数} \quad (4)$$

(c)性能評価の尺度

$$F値 = \frac{2\cdot精度\cdot再現率}{精度+再現率} \quad (5)$$

3.4 実験結果

以下のような結果が得られた。先行研究の結果を表 2 に示すが、本実験とは実験データ、評価方法が異なるため単純な比較はできない。先行研究[1]では、文中の助詞誤用を一箇所としている。本研究では、助詞誤用の数に制約がないため文中に複数の助詞誤用が複数存在する場合でも修正を行うことができる。

表 1　実験結果

	再現率(1),(3)	精度(2),(4)	F 値(5)
検出	0.71	0.78	0.73
修正	0.65	0.74	0.71

表 2 先行研究結果[1]

	再現率(1),(3)	精度(2),(4)	F 値(5)
検出	0.59	0.89	0.71
修正	0.49	0.33	0.39

4. 複文における助詞誤用修正システム

複文とは、動詞が 2 つ以上含まれる文のことである。そのため、複文は長くなる傾向があり、さらに助詞誤用があった場合に係り受けに間違いが生じる。特に、「は」は、節を越えて係るため、修正が必要である。

例)夕食**は(×⇒が)**終わったあとケーキを食べる。

この文の係り受け解析を行うと次のようになる。

①夕食は<係り先=4>

⇒②「終わった」に係るのが正しい

②終わった<係り先=3>

③あとケーキを<係り先=4>

④食べる。

この状態で格フレームを適応すると、「夕食」と「食べる」という関係から助詞「を」が適当だと判断される。

しかし、「夕食を終わったあと食べる。」は間違いである。そこで以下の手順で係り受け関係を変更する。

①用言から一番近い体言と係り受けを結ぶ

1. 夕食⇒終わる
2. 夕食⇒食べる

②それぞれの関係から格フレーム情報を元に主格、目的格をとるか判断

1.「が」⇒主格
2.「を」⇒目的格

③主格をとる場合、優先的に関係を結び助詞を補完

夕食**が**終わる

また、「は」、「が」は以下の場合分けができる。

1)目的格が主格よりも前にある文は、主題をもたない

⇒「が」を補完

2)動詞文では、語順が前の格成分が主題

⇒「は」を補完

これらの条件によって複文における助詞誤用の修正を行うことができる。

5. まとめ

本研究では、日本語学習者の書いた作文を対象として、格助詞の検出、修正を行った。実験結果の示す通り、大規模格フレームは助詞修正に有効であることがわかった。

今回、格フレームの名詞部分を意味属性に置き換えることで、格フレームを広く利用できるようにした。しかし、意味グループが広くなり、正確な格フレームとの対応がとれず、出力を行えない例も存在する。そのため、名詞を意味属性に置き換える作業では、格フレームごとに調査する必要がある。

今後の課題として、複文への対応が挙げられる。これに関して、現在 4 節で説明したシステムを作成し、実験を行っている。

謝辞

今回データを提供していただいた、国立国語研究所の宇佐美洋氏に深くお礼を申し上げます。

【参考文献】

[1]石川裕司、河合敦夫、多田直人、永田亮、桝井文人：日本語学習者の作文における格助詞の誤り検出と訂正、情報処理学会第 66 回全国大会 講演論文集 pp.2-323 - 2-324 (2004)

[2]南保亮太、乙武北斗、荒木健治:文節内の特徴を用いた日本語助詞誤りの自動検出・校正、情報処理学会研究報告自然言語処理研究会報告 NL-181pp.107 - 112(2007)

[3]池原悟、宮崎正弘、白井諭、他：日本語語彙大系、岩波書店(1997)

[4]河原大輔、黒橋禎夫:高性能計算環境を用いた Web からの大規模格フレーム構築、情報処理学会研究報告自然言語処理研究会報告 NL-171pp.67 - 73(2006)

[5]奈良先端科学技術大学院大学：日本語係り受け解析器南瓜

[6]国立国語研究所：日本語学習者による日本語作文と、その母国語との対訳データベース　ver.2(2001)