K-033

外国人の初級日本語学習支援システムにおける数詞誤りの訂正方式
Error Correction in Kana Expressions of Numerals for Foreigner's Basic Japanese Language Learning

谷之口 優人† 杉野 勝也† 佐藤 俊也† 絹川 博之†
Yuto Taninokuchi Katsuya Sugino Toshinari Sato Hiroshi Kinukawa

1．はじめに

初級日本語を学習している外国人は,仮名表記において初歩的な誤りをすることが多い.その中でも数詞の正しい読みは数字表現と助数詞との組み合わせによって音韻変化があるので学習者にとって誤りやすい.例えば,数字の六本を「ろくほん」と読むような誤りである. 本稿では,数字表現と助数詞を組み合わせたもの数詞と呼ぶ.

このような誤りを含む日本語文を解析する方式はあまり発表されておらず,これらの検出,訂正は日本語教師などの人手に頼っているのが現状である. そのため,学習者が独学で文章作成を学習することは困難である.

そこで我々は外国人学習者が独学で文章作成を学習できることを目標として日本語学習支援システムを開発している.現段階では,対象を初級日本語にしぼり,学習者の作成した文の誤りのうち振り仮名の誤りを検出,訂正する方法を研究している.この方法の一つとして,数字表現と助数詞の組み合わせにおいて,数字表現の促音変化や,助数詞の濁音,半濁音変化に規則性があることに着目した数詞の訂正方式について報告する.

2. 初級日本語支援システム

2.1 対象とする日本語

本研究では外国人のための初級日本語を研究対象にしているが,ここでの初級日本語とは,財団法人日本国際教育支援協会と独立行政法人国際交流基金が行っている日本語能力試験の N3 レベルに相当しており,漢字は 300 字程度,語彙は 1,500 語程度が必要とされている.

2.2 文章作成支援

本システムは学習者が文章を入力すると,システムが誤り検出,訂正を行い,学習者に誤りの指摘と正解を提示する.なお入力される文章は,アラビア数字や漢数字などは使わず,全て平仮名表記したものを想定している.

3. 数詞の訂正

数詞の検出の流れは次のようになっている(図 1 参照).まず最初に文節区切り処理を行い,助数詞候補の検出,数字表現候補の検出,そして最後に数詞の訂正を行う.

3.1 数詞誤り

初級日本語を学習している外国人は,数詞誤りにおいて規則的な間違い方をすることが多い.日本語の数詞は,数字表現と助数詞の組み合わせによって規則的に変化する

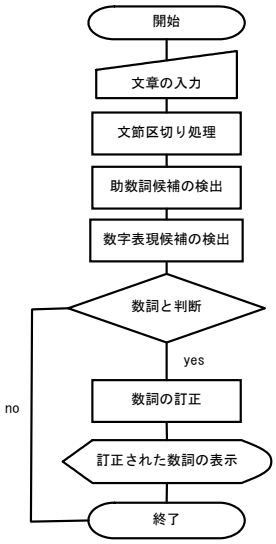

図 1 .数詞の訂正の流れ

が,学習者はこの正しい組み合わせを誤ることが多い.

例えば数字の六本の正しい読みは「ろっぽん」であるが,学習者は「ろくほん」のように誤る.この場合数字表現,助数詞,共に誤りといえる.

その他の誤りとして数字表現においては,促音の抜けがある.「むっつ」を「むつ」と書いてしまうような誤りである.

助数詞に関しては,濁音,半濁音の変化,漢字の別の読みをしてしまう誤りがある.漢字の別の読みというのは,「4 月」を「よんげつ」と読むような誤りである.

3.2 文節区切り処理

数詞の訂正を行う前処理として,入力された文章に対して,文頭処理,文末処理,そして文節に区切る処理を行う.この処理で,明らかに数詞ではない部分を検出しておく.

初級日本語の教育において「私は,～です.」「これは～です.」のような決まった文型で教えることが多い.学習者もこのような文型に沿って文章の作成を行うので,「私は」「これは」のような文頭からは,数詞の検出処理を行わないようにする.文末処理も同じように「です.」「しました.」のような文末を検出する.

文節に区切る処理では,助詞や句読点などを検出し,そこで区切る処理を行う.この処理を行うことで,文節の繋がりによって,数詞と判定されてしまうような文字の並びが現れることを防ぐ.

†東京電機大学大学院　未来科学研究科

Graduate School of Science and Technology for Future Life, Tokyo Denki University

3.3 助数詞候補の検出

文節区切り処理を行った文章から,助数詞を検出する.助数詞の文字列データはプログラム上に格納しておく.このデータには学習者が間違えやすい誤りも含まれている.「いちがつ」の「月」の読み間違いである「げつ」などである.助数詞候補を検出したら,その前にある文字列を数字表現候補として格納する.

3.4 数字表現候補の検出

助数詞候補の検出の後,数字表現候補文字列について、数字表現であるか判定する.正しい数字表現であった場合は,そのまま検出し訂正する.

しかし,助数詞候補を機械的に検出すると「れいぞうこ」のような単語でも,助数詞の「個」を検出し,数字表現候補を「れいぞう」として格納してしまう.こういった明らかに数字表現ではないものを判定する処理として,助数詞候補の直前 3 文字の文字列を 1 文字ずつ,数字表現の読み仮名候補の文字列に一致するか確認する.

(1) 助数詞候補文字列の直前の文字

助数詞直前の 1 文字目に来ると思われる文字列として,促音の抜けによって,数字表現の先頭の文字が現れることが考えられる.「むっつ」であれば「むつ」のような誤りである.促音が使われる数字表現は,1 であれば「いっ」,3 であれば「みっ」,4 であれば「よっ」などの表現があり,1 文字と促音を組み合わせた表現の先頭文字から一致するかを確認し数字表現であるか判定する.「い」であれば 1 のように判定する.

(2) 助数詞候補文字列の前方 2 文字目

2 文字目には,2 文字の数字表現の先頭文字が現れると考えられる.例えば 1 の数字表現は「いち」「いっ」などがあり,「い」が現れれば 1 であると判断する.「ふ」であれば,2 の数字表現である「ふた」と考え,2 であると判定する.

(3) 助数詞候補文字列の前方 3 文字目

3 文字の数字表現である「きゅう」「ここの」「じゅう」などから,「き」,「こ」,「じ」の文字で判定を行う.

数字表現であると判定したとき,数詞の訂正で利用するためにその数がいくつであるかという情報を格納しておく.以上の処理で,数字表現と判定されなかったものは数詞訂正の対象としない.

3.5 数詞の訂正

正しい数詞に訂正するには,数がいくつであるかという数値と,助数詞はどれであるかという情報を与える.プログラムには助数詞別に正しい読みが格納されており,正しい数詞の読みが文字列として返される仕組みになっている.

「6」という情報と「本」という情報を与えると「ろっぽん」という文字列が返される.数値で情報を扱っているため,「6 本」といったアラビア数字と漢字で表した助数詞とからなる表現に変換することもできる.

助数詞の「回」と「階」は一部に濁音の変化があるだけであり,判定が難しいため,結果には両方の候補を挙げる.

実際のプログラムでは下のような形で処理が行われる.

```
文章の入力：そのぺんをろくぼんください。
(文節区切り処理 - 文頭：その　文末：ください)
文節 1：ぺん　文節 2：ろくぼん
*****************************
文節 1：ぺん
- 助数詞候補の検出 – 無し
- 処理の終了
*****************************
文節 2：ろくぼん
- 助数詞候補の検出 – 「ぼん」
- 数詞候補：ろく ぼん

■ 数詞の訂正
数詞候補：ろく ぼん
助数詞：ぼん (本)
数字表現候補の検出：ろく　数値：6
正解：ろっぽん (6 本)
```

4. 考察

現段階として,日本語学習者向けの教科書内に出てくる例文を参考にし,数詞誤りを含んだものを入力として実験を行った.結果は,165 文字列中正しい候補のみを出したものが 108 文字列,正しい候補と誤った候補の両方を出したものが 14 文字列,誤った候補のみを出したものが 6 文字列であった.助数詞検出後に数字表現で無いと判断されたものは 54 個であった.

誤った候補として,「じゅういち」などの数字表現を「いち」で検出している,「かい」という助数詞を「かい」,「か」(日),の二つの助数詞を検出して,候補を出しているものがあった.

数字表現候補の検出では,明らかに数字表現でないような候補は除外されていたので,今回提案した検出方法が有益であることがわかった.

数字表現と助数詞の組み合わせの誤りについては正しい数字表現と助数詞を使っていれば,容易に検出ができるので,日本語学習者の初歩的なミスには対応ができると思われる.

5. おわりに

数詞の仮名表記の誤りを検出し,訂正する方式を提案した.現在,提案した方式に基づいてプログラムを開発している.開発したプログラムを用いて実験評価し,さらにプログラムを改良していく予定である.

謝辞

本研究を行うにあたり,学校法人 吉岡学園 千駄ヶ谷日本語学校に御協力を頂きました.この場を借りて御礼を申し上げます.

参考文献

[1] 杉野勝也, 佐藤俊也, 絹川博之:外国人の初級日本語学習における仮名表記と文法の初歩的誤りの検出方式, 第9回情報科学技術フォーラム(FIT2010)第3分冊 (2010).

[2] スリーエーネットワーク編著:みんなの日本語 初級I, 本冊, スリーエーネットワーク (1998).