K-010

外国人の初級日本語学習における仮名表記誤りの分類と訂正方式

Error Classification and Correction System of of Kana spelling in Foreigner's Japanese Language Learning

佐藤 俊也† 杉野 勝也† 絹川 博之†
Toshinari Sato Katsuya Sugino Hiroshi Kinukawa

1. はじめに

近年，コンピュータが教育分野で利用されるようになり外国人を対象とした日本語教育でも多く利用されるようになったが，外国人日本語学習者が作成した文章を添削するシステムはほとんど見られず，日本語教師等の人手によって添削され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．第１段階として，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．なお，ここでの振り仮名誤りとは，日本語の読みの誤りをいい，例えば「たべる（食べる）」を「たへる」などである．

2. 初級日本語文の誤り検出訂正システム概要

外国人の初級日本語文の誤り検出システム[1]の具体的な概要としては，まず，日本語形態素解析ソフトJUMAN[2]を使用して，形態素解析する．この後，解析により得られた品詞が，初級日本語の文型辞書と照合一致するかを調べる．ここで，文型辞書とは，初級日本語で扱う文の正解日本語文を形態素解析結果の形式で登録したもののことである．そして，外国人の作成した日本語文の形態素解析結果と文型辞書が，一致しない場合は，文に誤りがあると判断し，一致する場合は，文型としては確認できたが，誤り単語を含む場合があるので，さらに各単語を初級単語辞書と比較し，正誤判断する．その結果，各単語が，初級単語辞書に登録されている場合は，正しい文と判断して，そうでない場合は，誤りの文であるとする．この際に検出された誤り単語は，正しく訂正し正解文を示す．

3. 誤り訂正方式

単語の誤りの種類を大きく分類すると規則性のある誤りと規則性のない誤りに分けられる．規則性のない誤りは，不規則性誤り訂正辞書を用いた訂正方式を採用して訂正を実現することを試みる．

3.1 規則性のある誤り訂正方式

規則性のある誤りで扱う種類を以下に示す．

1. 非清音を清音にしている
2. 非濁音を濁音にしている
3. 長音（う）の間違い
4. 拗音を別の拗音にしている
5. 促音が抜けている
6. 濁音を半濁音にしている
7. 濁音を別の濁音にしている

以上についてそれぞれの規則性を利用して，誤り訂正を行う．

規則性のある誤り訂正方式とは，与えられたひらがなの誤り単語を以下にそれぞれ示す処理方針に沿って訂正し，正解単語候補を複数生成する．その複数の候補を初級単語辞書と比較し，絞込みを行い，訂正を実現する方式である．初級単語辞書とは，初級の日本語学習者が、使用すべき日本語単語を全て登録した辞書である．

3.1.1 「非清音を清音にしている」の訂正候補生成

濁音および半濁音を含む単語「ゆうびんきょく（郵便局）」を「ゆうひんきょく」に，「パン」を「ハン」にするなどの誤りで，ローマ字表記での濁音は，G,Z,D,B，半濁音はPである．濁音に対応する清音はそれぞれK,S,T,H,半濁音は，Hである．よって単語中のK,S,T,HをG,Z,D,B,Pに変え単語候補を生成する．

3.1.2 「非濁音を濁音にしている」の訂正候補生成

「おとこ（男）」を「おどこ」にするなどの誤りで，この場合は，3.1.1と反対の処理をする．単語の中のG,Z,D,Bをそれぞれ K,S,T,H,Pに変えて単語候補を作る．なおBは，HとPに変換して単語候補を作成する．

3.1.3 「拗音を別の拗音にしている」の訂正候補生成

「しゃちょう（社長）」を「しゃちゅう」にするなどの誤りで，拗音の母音を変化させ単語候補を生成させる．

3.1.4 「濁音を半濁音にしている」の訂正候補生成

「しんぶん（新聞）」を「しんぷん」にするなどの誤りで，ＰをＢに変え単語候補を生成する．

3.1.5 「濁音を別の濁音にしている」の訂正候補生成

「かぜ（風邪）」を「かぞ」，「ざっし（雑誌）」を「ぎっし」にするなどの誤りで，母音のみの誤りと母音と子音の誤りの２通りであり，濁音を他の１９の濁音に変化させて単語候補を生成する．

3.1.6 「長音（う）の間違い」の訂正候補生成

(1)「う」の欠如の訂正候補生成

「う,お段＋う」の単語「くうき（空気）」や「ひこうき（飛行機）」の「う」が欠如し，「くき」「ひこき」にしている誤りで，「う段」と「お段」を発見し，その後ろに「う」を挿入して単語候補を生成する．

なお，*U,*0の後に「う」が，すでに入っている場合には，「う」は，挿入しない．

(2)余計な「う」の訂正候補生成

「きょねん（去年）」に余計な「う」を追加して「きょうねん」にする誤りで，余計な「う」を「う段」と「お段」の後ろに検出して，削除して単語候補を生成する．

†東京電機大学 大学院 情報メディア学専攻
Tokyo Denki University, Graduate School

3.1.7「促音が抜けている」の訂正候補生成

「ざっし（雑誌）」を「ざし」にしている誤りで，この場合は，先頭の文字の後から文字と文字の間に「っ」を挿入して単語の候補を作成する．

3.2 不規則性誤り訂正辞書を用いた訂正方式

規則性のない誤り単語とその単語の正しい読み仮名と漢字と品詞をひとつのペアにして蓄積し，辞書を作成する．この辞書を不規則性誤り訂正辞書と呼ぶ．誤り単語を検出した際にこの辞書と比較し，その辞書情報から正しい単語に訂正する方式を「不規則性誤り訂正辞書を用いた訂正方式」と呼ぶ．

不規則性誤り訂正辞書に登録する誤り単語の種類を以下に示す．

(1)数詞の読み誤り単語

例として，「ようか」を「はちにち」の様な時間や日時などの数に関する読み間違いなどの誤り方である．辞書には，21 単語が登録されている．

(2)名前に関する誤り単語

例として，「エレン」を「ユッネン」にしてしまうような名前を間違えてしまう誤り方である．辞書には，2 単語が登録されている．

(3)「ざ,ず,ぞ」を「じゃ,じゅ,じょ」にしている誤り単語

例としては，「かざる」を「かじゃる」で，発音的類似性により取り違えを起こしている誤り方である．辞書には，55 単語が登録されている．

(4)カタカナ語単語末の長音記号を欠如している誤り単語

例としては，「メニュー」の長音記号を欠如して「メニュ」にしてしまった誤り方である．辞書には，17 単語が登録されている．

(5)文字の形の類似性に関する誤りを含む単語

例として，「わるい」を「ねるい」ように文字の形状類似性に起因して起こっている誤り方である．辞書には，18 単語が登録されている．

(6)その他の誤りを含む単語

例としては，「しょちょう」を「しゃきゅう」にしているなどの誤りで，規則性のある誤り訂正方式で訂正することの出来ない規則性を持った誤り単語や複数種類の誤りをひとつの単語の中に含んでいるような誤り単語である．辞書には，38 単語が登録されている．

4. 実験評価

実験は、規則性のある誤り訂正方式と不規則性誤り訂正辞書を用いた訂正方式の２回に分けて、それぞれの機能的な完成の程度を検証するために行った。

$$\text{精度} = \frac{\text{正しく訂正できた単語数}}{\text{訂正処理した単語数}}$$

$$\text{再現率} = \frac{\text{正しく訂正できた単語数}}{\text{訂正すべき誤り単語数}}$$

4.1 規則性のある誤り訂正方式

精度は、109/129(=85%)で，再現率は，109/129(=85%)となった．問題点は，単語候補に正解単語以外にも初級単語辞書に登録されている単語候補を生成してしまったことである．したがって，複数の正解候補から正しいひとつの候補だけに絞ることが，今後の課題となっているのである．

4.2 不規則性誤り訂正辞書を用いた訂正方式

精度は、122/122(=100%)で，再現率は，122/218(=56%)となった．問題点は，訂正すべき誤り単語全てを不規則性誤り訂正辞書に登録できていないことにあり，今後，いまだ登録されていない誤り単語を登録して再現率を向上させていきたいと考えている．

5. 考察

全体精度は，231/251(=(109+122)/(129+122)=92%)となり，全体再現率は，231/347(=(109+122)/(129+218)=67%)となる．全体再現率は，今後の改善が必要であるが，その主要な原因は，4.2 で述べた単語の登録数の問題である．

6. おわりに

本実験成果として，規則性のある訂正方式と不規則性誤り訂正辞書を用いた訂正方式を実現するプログラムを作成した．

今後は，現在のシステムをより改善していき訂正精度を高めて行きたいと考えている．

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,絹川博之, “外国人の初級日本語文の誤り検出方式”, 第 7 回情報科学技術フォーラム(FIT2008)第 3 分冊, pp563-564(2008).

[2] 杉野勝也, 絹川博之, “外国人の初級日本語文における振り仮名の誤り検出”, 情報処理学会 第 71 回全国大会分冊 4, pp.615-616(2009).

[3] 杉野勝也,佐藤俊也,絹川博之, “外国人の初級日本語文における振り仮名の誤り検出”,第 8 回情報科学技術フォーラム(FIT2009)第３分冊, (2009).

[4] 益岡隆志,田窪行則, “基礎日本語文法 –改訂版-”, くろしお出版(1992).

[5] 益岡隆志, “24 週日本語文法ツアー”, くろしお出版 (1993).

[6] 吉川 武時, “日本語文法入門”, アルク(1989).

[7] 千駄ヶ谷日本語教育研究所著, “コミュニケーション日本語 1”, 千駄ヶ谷日本語研究所 (1999) .

[8] 千駄ヶ谷日本語教育研究所著, “コミュニケーション日本語 2”, 千駄ヶ谷日本語研究所 (1999) .

[9] 千駄ヶ谷日本語教育研究所著, “コミュニケーション日本語 3”, 千駄ヶ谷日本語研究所 (1999) .

[10] 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.)

[11] スリーエーネットワーク編著, “みんなの日本語 初級Ⅱ 本冊” スリーエーネットワーク(1998.)