日本語作文推敲支援システム「ナツメグ」における誤用検出手法の評価

八木 豊（株式会社ピコラボ）
ホドシチェク・ボル（国立国語研究所）
阿辺川 武（国立情報学研究所）
仁科 喜久子（東京工業大学）

Evaluation of Error Detection in Japanese Composition Support System “Nutmeg”

Yutaka YAGI (Picolab Co., Ltd.)
Bor HODOŠČEK (National Institute for Japanese Language and Linguistics)
Takeshi ABEKAWA (National Institute of Informatics)
Kikuko NISHINA (Tokyo Institute of Technology)

１．はじめに

国立国語研究所による「現代日本語書き言葉均衡コーパス」（以後 BCCWJ）に加えて、近年、日本語学習者の作文に含まれている誤用に対してタグ付けを施した学習者作文コーパスが公開されてきた。それに伴い、日本語教育の分野では、それらのコーパスおよび自然言語処理の技術を利用して作文に含まれる誤用を自動的に検出し、学習者の作文を支援するための研究が行われている（今村（2012）、水本（2013）、八木（2012））。

我々も、レジスターの誤り検出を中心に、日本語作文推敲支援システム「ナツメグ」の開発を進めてきた。レジスターとは、「社会的な拘束力をもつ言語学上の規範」における言語使用域の変異のことであり、Halliday（2004）はレジスター機能として次の 3 項目を挙げている。

(1) コミュニケーションの目的と主題に関わる「フィールド」（Field of Discourse）
(2) コミュニケーションを行うための手段に関わる「モード」（Mode of Discourse）
(3) コミュニケーションパートナー同士の関係に関わる「テナー」（Tenor of Discourse）

つまり、書き手と読み手がどのような関係で、どのようなコンテキストのもとで言語表現を使用するかによって、異なる語彙・文法項目で記述されることを示すものである。学習者作文においては、論文や授業で提出するレポートの中で話し言葉を使用しているなど、場にそぐわない表現がレジスターの誤りに該当する。

本稿では、日本語作文推敲支援システム「ナツメグ」におけるレジスターの誤り検出について示し、学習者作文コーパスに基づいてその精度を評価する。また、日本語学習者に実際にレポート形式の作文を書いてもらい、その過程で学習者がシステムを利用して推敲する評価実験の経過を報告する。

２．レジスターの誤り検出

ホドシチェク（2011）では、書籍、雑誌、新聞など様々なレジスターの書き言葉を収録している BCCWJ と、科学技術論文のコーパスを使用してレジスターの誤り判定を行う手法を

提案している。学習者が作文の目的とするレジスターを想定し、BCCWJ の各レジスターと科学技術論文の中から目的のレジスターに近いデータを準正用データ、目的のレジスターから遠いデータを準誤用データとして設定する。そのうえで、「名詞＋格助詞＋動詞」、「名詞＋格助詞＋形容詞」、「形容詞＋名詞」の 3 つのパターンに一致する共起表現の出現分布をもとにカイ二乗検定を行い、準誤用データにおける出現が有意に多い場合に、その共起表現は学習者が作文の目的とするレジスターの下ではふさわしくない表現、即ち、誤用であると判定するものである。八木（2012）では、学習者の作文の目的を論文やレポートを書くことと想定し、科学技術論文、白書など比較的硬い文章のレジスターを準正用データ、Yahoo!知恵袋、Yahoo!ブログ、国会会議録など比較的くだけた文章のレジスターを準誤用データとすることで、「ことがある」、「問題が起きる」、「一緒にする」、「いい経験」のように日本語として誤っているわけではないが論文やレポートの場合には別の表現に書き換えたほうがよい表現を、レジスターの誤りとして自動的に検出できることを述べた。

本稿では、レジスターの誤り検出対象が上記の特定の 3 つのパターンに制限されていた点を改善し、任意の句の 3-gram、2-gram、1-gram および、単独の形態素について同様の手法でレジスターの誤りを検出できるように拡張したものを使用する。作文の目的として想定するのは同様に論文やレポートで、科学技術論文、白書、法律を準正用データ、Yahoo!知恵袋、Yahoo!ブログ、国会会議録を準誤用データとした。ただし、科学技術論文には言語処理学会年次大会予稿集の本文が新たに追加されている。

３．学習者作文コーパスに基づく評価

レジスターの誤り検出の精度を確認するために、学習者作文コーパス「なたね」のデータを利用して評価を行った。

「なたね」は、我々が独自に収集した学習者作文に対して日本語教師による添削を行った誤用タグ付きデータで、2014 年 2 月現在、大学院や大学あるいは語学学校に在籍する 192 人の日本語学習者による 285 作文（総文字数 205,520 字）に含まれる約 6,500 箇所の誤用に対しておよそ 9,000 件の誤用タグを付与して公開中である[1]。誤用タグは大きく「誤用の対象」、「誤用の内容」、「誤用の要因・背景」という 3 つの視点から成り、さらにそれぞれを 3 階層に細分類している。レジスターに関する誤用は「誤用の要因・背景」の一つで、483 件が登録されている。そこから 100 件を含む文を取り出してシステムによるレジスターの誤り検出を行い、日本語教師が添削で指摘したレジスターの誤用との一致度合いを人手により確認した。結果は、日本語教師が指摘した箇所の再現率が 78.0%、システムの検出結果全体の精度が 77.6%であった。内訳を表 1 に示す。

まず、日本語教師が添削で指摘した誤用 100 件のうち、78 件をシステムが自動的に検出した。この中には、以下に挙げる誤用例 1、2 のような副詞に関するものや、誤用例 3、4

表 1 日本語教師の指摘箇所とシステムの検出箇所の比較

	システムによる検出有り	システムによる検出無し
日本語教師による指摘有り	78 件	22 件
日本語教師による指摘無し	92 件 （うち 38 件は誤検出）	n/a

[1] 学習者作文コーパス「なたね」　http://hinoki.ryu.titech.ac.jp/natane

のような句に関するものも含まれている。これらは、レジスターの誤り検出の拡張前には検出対象に含まれておらず、今回、拡張した効果といえる。

【誤用例 1】なぜなら日本は今少子化により、労働不足の問題はどんどん大きくなる。
【誤用例 2】ちょっとロボットに興味をもっている私はわくわくした。
【誤用例 3】今中国ではお父さん一人の給料では家族全員を養うことが非常に難しい。
【誤用例 4】でも、外国人の先輩は大学院で研究しながら、会社に務めるそうです。
※例中の下線部は日本語教師が添削で指摘したレジスターの誤用箇所

一方、誤用例 5、6 はシステムでは検出できなかった誤用の一例である。

【誤用例 5】たぶん誇張だと思いました。
【誤用例 6】別に利益のためじゃない将来保証のためである。

データを確認したところ、準正用データとしている科学技術論文のなかに言語処理学会の会誌「自然言語処理」および年次大会予稿集の本文が登録されており、「たぶん」や「じゃない」のような表現が含まれていた。その結果、準正用、準誤用における出現頻度に有意差が出ず、誤用として検出できなかったことになる。ただし、「自然言語処理」および年次大会予稿集の当該箇所では例文の一部として使用されており、出現頻度を集計する際は、そういった例文や他から引用されている文を除外するなどの対応が必要であると考える。

また、システムは、同一文中に含まれる日本語教師が指摘しなかった 92 箇所の誤用を検出した。そのうちの 54 件は誤用として妥当なもので、残り 38 件は誤検出だった。さらに、54 件の半数以上は「です」、「ます」の使用を指摘したものである。これらについては、日本語教師が見落としたのではなく、日本語教師のほうで意図的にそこまでチェックせず、その他の添削を優先したものと思われる。

４．日本語学習者による評価実験

2014 年 1 月から、上記のレジスター誤り検出機能を備えた作文推敲支援システムの評価実験を行っている。学習者に依頼する内容の大まかな手順は以下の通りである。

1. J-CAT（Japanese Computernized Adaptive Test）を受験する
2. 背景調査のアンケートに回答する
3. 4 つの課題について、システムを使用して作文を書く
4. システムに関するアンケートに回答する

実験に先立ち、学習者の現在の日本語能力および言語的な背景を確認し、後日、分析に使用するため、J-CAT の受験と背景調査のアンケートを必須とした。実験では、あらかじめ決められた 4 つの課題について、順番に作文を書いていく。各課題は、最初にテーマとプロンプトを示し、学習者がそれに関する作文を一通り書き終えてから、1 回だけ、レジスターの誤り検出機能による添削を受けて書き直せるようにした。1 回だけに制限した理由は、添削前後の作文の比較を容易にするためである。また、1 つの課題を終えて次の課題に取り掛かるまでに最低 3 日間空けなければならないよう設定した。

図 1 は、レジスターの誤り検出機能による添削を受けた後、学習者に表示される添削結果画面である。画面左下に添削結果を表示しており、背景色を変えた箇所や下線の箇所が、システムが誤りの可能性を指摘している単語や表現である。学習者はそれぞれの単語や表現をクリックして指摘された内容を確認し、必要なら、辞書や「なつめ」で使い方を調べ

図　1　システムによる添削結果画面

たうえで、画面右下の入力欄で作文を書き直す。

日本語学習者による評価実験は開始したばかりであるが、システムの添削結果を受けて、以下のように適切な修正を施す学習者も散見される。

【修正例 1】私が日本人について理解できないことはギャルです→である。

【修正例 2】人をあっ→に会ったりするのは私が考えたより少なかった。

【修正例 3】性についての認識も違うだ→異なると思う。

※矢印の左側の下線部がシステムによる指摘箇所、矢印の右側が学習者による修正内容

システムによる指摘箇所と学習者による修正内容を始め、作文途中の文章（10 秒毎に記録）、作文完了までに要した時間、クリックして確認した箇所など、システム内における学習者の操作はバックエンドでデータベースに記録しており、今後は、学習者の日本語能力や言語的な背景と合わせて分析する予定である。

謝辞

本研究は、文部科学省科学研究費補助金基盤研究（C）「日本語作文支援システムで考慮すべき学習者属性情報と提示項目の分析研究」(研究代表者：阿辺川武、研究期間：2012 年 4 月～2015 年 3 月）による助成を得て実施しています。

参考文献

今村賢治、齋藤邦子、貞光九月、西川仁（2012）「小規模誤りデータからの日本語学習者作文の助詞誤り訂正」自然言語処理, vol.19, no.5, pp.381-400.

水本智也、小町守、永田昌明、松本裕治（2013）「日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得」人工知能学会論文誌, vol.28, no.5, pp.420-432.

八木豊、ホドシチェク・ボル、仁科喜久子（2012）「BCCWJ と学習者作文コーパスを利用した日本語作文支援－表記と共起に関する誤用添削プロトタイプ構築－」第 1 回コーパス日本語学ワークショップ予稿集, pp.315-320.

Halliday M.A.K. and C.M.I.M. Matthiessen (2004). An Introduction to Functional Grammar. 3rd ed. London: Arnold

ホドシチェク・ボル、仁科喜久子（2011）「作文支援システムにおけるレジスターの扱い」世界日本語教育研究大会 異文化コミュニケーションのための日本語教育 2, pp.522-523.