日本語学習における作文からの誤り検出

―機械学習による 1 文節内の誤り検出性能―

趙艷[*1]・高瀬治彦[*1]・北英彦[*1]
Email: 419de52@m.mie-u.ac.jp

*1: 三重大学大学院工学研究科

◎Key Words　　日本語学習，作文授業，誤り検出

1.　はじめに

近年，第 2 言語として日本語を学ぶ学習者が大幅に増加しているが，教師の増加が追い付いていない．海外日本語教育機関調査の結果報告書によると，2015 年時点では，日本語学習者の数は約 29 倍に増加しているのに対し，教師の数は約 16 倍の増加にとどまっている(1)．この結果，日本語教室の受講者人数は多くなりがちであり，教師の支援が必要とされている．

言語の学習において修得すべき技能には，「聞く」，「話す」，「読む」，「書く」の 4 つがある．本稿では，書く技能の学習に着目し，そのために有効な授業方法である作文授業に着目する．作文授業では，学習者の作文を添削することが必須だが，一人の教師が添削できる作文の量には限界がある．そのため多人数講義では，作文を利用した授業の効果を十分に発揮することが困難である．

本稿では，これらの困難を解消するために．計算機システムによる作文の誤りを自動的に指摘させることを試みる．これが成功すれば，添削に関する教師の負担が減り，教師に学生の状況に注意を払うことができる余裕が生まれ，授業の効果を高めることができるだろう．本稿では特に，作文授業に実際に犯した誤りを用いて機械学習するとこで，1 文節だけで判断できる誤りを検出するシステムを構築し，作文授業における有効性について議論する．ここで対象としている誤りは，作文授業の支援という観点からは十分ではないが，このようなシステムを講師の少ない労力で構築できる可能性を検討するために，まずはこのように限定した誤りの自動検出を通じて議論する．

2.　作文の誤りを指摘するシステム

2.1　作文授業支援に関する従来研究

これまでにも，作文の誤りの検出を計算機システムに行わせる研究が多く行われてきた．例えば，英語学習者の犯す誤り検出に関する研究では，Liu らは動詞の誤りの検出を行った(2)．Rozovskaya and Roth らは前置詞の誤りの検出を行った(3)．Dahlmeier and Ng らは前置詞と冠詞の誤りの検出を行った(4)．また，日本語学習者の誤り検出に関する研究では，谷之口らは日本語学習支援システムにおける数詞の誤り検出を行った(5)．大木ら，橋本ら，南保らは日本語助詞の誤りの検出を行った(6-8)．Oyama ら，笠原ら，今枝らは日本語格助詞の誤りの検出を行った(9-11)．いずれの研究でも，学習者の犯す誤りを 1 種類あるいは 2 種類に限定しており，特に助詞の誤りの検出を行っている．しかし，Zhang によると，作文授業では，学習者が犯しやすい誤りは自然さと自他動詞である．助詞は誤り種類全体の 17%を占める(12)．そのため，助詞以外の誤りの検出が必要であり，誤りの検出の支援には十分ではない．

また，作文授業の支援に限らず，人が書いた文章の誤りを検出したり，評価したりする研究が行われている．石井らは誤りに基づく日本語学習支援システムの一部として，登録文字列と一致する誤りを指摘する機能を提供した(13)．この機能は，適切な誤りを事前に登録できればその効果を発揮するが，そのような事前登録は教師にとって負担が大きい．また，文章の校正を行う Web サービスがいくつか提供されている．例えば，リクルート社による A3RT では機械学習技術を利用して文章の校閲を行う(14)．しかし，この既存のツールは日本語が母国語であるような人の文章を対象として，誤りを検出しかできない．また，学習者作文評価システムである jWriter は，学習者の作文のレベルを評価するが，添削をするわけではない(15)．現状ではこのように，学習者の作文の添削支援という意味では十分に支援できていない．

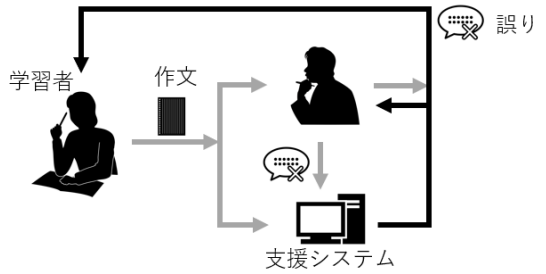

図 1 システムの構成

2.2 構築する支援システムの概要

本稿で作成する誤り検出システムの概要を図1に示す. システムの動作は大きく以下の 3 段階に分かれる.

1. データの収集
   実際の作文授業において, 学生が提出した作文と, その添削結果を蓄積することで, 正しい文・誤った文をデータベース内に蓄積する.
2. 機械学習による誤り検出ルールの自動獲得
   蓄積された正･誤の文例を用いて機械学習を行う.
3. 獲得した誤り検出ルールを使用した誤り検出
   教師による添削と並行して, システムも作文の誤りをチェックする.

作文授業では, 学習者は, 作文を作成し講師に提出し, 教師は提出された作文を添削し学習者に返却する. 問題点を修正しながら提出･添削を繰り返すことで, 作文能力を高めてゆく. ここで, 誤り検出システムに学生の作文をチェックさせることで, 学習者は教師からの指摘だけでなく, システムからの指摘も受け取るようになる. その結果, 学生は教師が作文をチェックし終わるまで必ずしも待つ必要がなくなり, 見直しを早い段階でできるようになる. また, 教師はすべての誤りを指摘する必要がなくなり, 添削の手間が軽減するだろう.

このシステムは, 機械学習を用いているため, 最初は正しい指摘ができない. 授業が進むことで, システムは教師が何度も指摘する誤りをシステムが把握するため, 教師の指摘内容と適合する指摘ができるようになる. また, 機械学習の方法を工夫することで, 同一の誤りだけでなく, 類似した誤りも自動検出できるようになるだろう.

3. 作文の誤りの検出

この節では, 本稿で対象としている 1 文節だけで判断できる誤りを検出する誤り検出規則を機械学習により獲得する方法を示す. 形態素と呼ばれる意味を持つ最小単位で, それぞれの文法情報を獲得し, これと正誤の情報を対にして, 正誤を判定する識別器を機械学習する. 具体的な手順は以下のとおりである(16).

1. 正しい文節・誤った文節を多数収集する.
2. 収集した各文節を形態素単位に分割し, 各形態素の文法情報を得る.
3. 手順 1,2 で得た各文節の文法情報から, その文節の特徴量ベクトルを作成する.
4. 作成した特徴量ベクトルとその文節の正誤を対にして, 多数の学習用データを作成し, これを用いて機械学習を行う.

手順 1 では, 正文と誤文から自動で, 文節を抽出すべきである. 本稿では, 機械学習の可能性を議論することに注力するため, 正文と誤文から切り出した多数の文節について, 日本人の筆者が正誤の情報を付与したデータを使用した.

手順 2 では, 形態素解析器と呼ばれる自然言語処理ツールを利用しこれを行う. ここで抽出する情報は, 各形態素の品詞（名詞・動詞など）・品詞の詳細情報（固有名詞など）・活用形（連用形・基本形など）とした.

手順 3 では, 各文節に対して, 固定長のベクトルを特徴量として作成する. 各形態素の特徴量は, 抽出した文法情報の組み合わせにより番号付けし, 一つの数値で表現する. 文節内の形態素数は一定でないため, 先頭から 4 個の形態素を選び, その文法情報のみを用いて 4 次元のベクトルを作成する. なお, 4 形態素に満たない文節では, 不足分を末尾に -1 を付加する. 例えば,「システムが誤りを自動で指摘する」という文を実験でも用いる CaboCha を用いて分析すると,「システム/が//誤り/を//自動/で//指摘/する」のように「/」の位置に形態素区切り,「//」の位置に文節区切りがあることが分かる. 最終文節である「指摘/する」については,「指摘」がサ変接続の名詞,「する」が自立する動詞の基本形であることが分かる. これらをそれぞれ 6 と 20 という数値に置き換え, 不足する 2 個の要素を-1 で埋めることで, (6, 20, -1, -1) という 4 次元のベクトルに変換できる.

手順 4 では, 識別器・機械学習手法としてランダムフォレストを用いる.

このような手法を採ることで, 正誤の判定が抽出された文法情報によって行われることになる. これは, 作文の課題が変わり使用する語が変化しても, 獲得した誤り検出ルールを利用できることにつながる. そのため, 学習し

た文節そのもの以外にも，文法的に同じ誤りの文節も検出できるだろう．

4. 実験

この章では，提案法によって，実際に作文授業で得られた正誤情報から，どのような未知の誤りを検出（1 文節内での誤り検出）（文法的な誤り検出）できるのか検討する．

4.1 実験条件

実験には，作文データとして，日本語学習者 19 人による， 2 回の作文授業における作文を使用した．これを CaboCha と呼ばれる係り受け解析器(17)を用いて，文節・形態素単位に分割した．得られた結果から，無作為に 200 個の文節を選択し，日本人の筆者により正誤（当該文節のみで誤りと判断できるかどうか）のラベルを付与した．これらの文節を特徴量ベクトルに変換し，表記が異なる 142 文節（うち 63 文節が誤り）を学習用データとして用いた．

ここでは，プログラミング言語 python およびその機械学習ライブラリである scikit-learn を用いて機械学習部分を実装した．

4.2 実験結果と考察

まず，第 1 の実験結果として，142 文節分の学習用データを用いたリーブワンアウト交差検証の結果を表 1 に示す．正誤の正解率は 80%であり，誤りと判定すべきもののうち 75%を検出し，誤りと検出したものの 20%が誤検出であった．この結果は，学習していない文節に対する判定結果である点を考えれば，不十分ながらも機械学習の可能性を示す結果であると考える．

表 1：重複しない文節を用いた学習結果

		判定結果 正	誤
付与された	正	67	12
ラベル	誤	16	47

ここで，実際の作文からの間違いの検出について，詳細に分析する．

3 章で述べたとおり，この手法では特徴ベクトルが同一であれば，学習していない誤りも検出できる．例えば，表 2 に示す文節は文字列としては異なるが，特徴量ベクトルを生成するときに使用する文法情報は同一である．そのため，「言いました」を学習して正しく指摘できるようになれば，その他の「出ました」についても正しく指摘できるようになる．また，「渡すの」と「渡るの」についても同様である．

この反面，同一の特徴量ベクトルのデータに対して矛盾する正誤情報が付与されていた場合，一方しか正しく判断できない．例えば表 3 に示す 4 個の文節はいずれも同じ特徴量ベクトルに変換される．しかし，正しい文節と誤った文節が混在している．このような場合，作成したシステムでどちらか一方しか正解しない．これについては今後の検討が必要だが，この理由としては特徴量ベクトルを生成する際の情報不足，形態素の文法情報の抽出ミス，正誤のラベルを付与する際のミスなどが考えられる．

表 2：特徴量ベクトルが一致するデータ

文節	特徴量に用いた文法情報	正誤
言いました	動詞,自立,連用形, 助動詞,＊,連用形, 助動詞,＊,基本形	正
出ました	動詞,自立,連用形, 助動詞,＊,連用形, 助動詞,＊,基本形	正
渡すの	動詞, 自立, 基本形, 助詞, 終助詞,＊	誤
渡るの	動詞, 自立, 基本形, 助詞, 終助詞,＊	誤

表 3：正誤情報が矛盾していたデータ

文節	特徴量	正誤
子馬に	名詞,一般,＊, 名詞,接尾,＊, 助詞,格助詞,＊	正
子馬が	名詞,一般,＊, 名詞,接尾,＊, 助詞,格助詞,＊	正
一つ川が	名詞,一般,＊, 名詞,接尾,＊, 助詞,格助詞,＊	誤
栗鼠声が	名詞,一般,＊, 名詞,接尾,＊, 助詞,格助詞,＊	誤

次に，第 2 の実験結果として，最初に用意した 200 文節を用いてリーブワンアウト交差検証の結果を表 4 に示す． この実験は，実際の授業で収集したデータをほぼそ

のまま使用した場合を意図しており，データ内に同一の文字列からなる文節が複数組存在している．この場合，正誤の正解率は 91%であった．この実験の場合，同じ文節が複数回現れるため，重複がない場合と異なる結果となる．結果としては，同一の文節が混ざっていない実験 1 よりは 11 ポイント高い結果が出た．実際にこのシステムを使用する場合は，同じ間違いは繰り返されるので，このような状況は容易に発生する．これをふまえると，91%という正解率は，このシステムを単独で使用する場合には十分ではないが，教師の助けにはなると考える．

表 4： 全文節を用いた学習結果

		判定結果	
		正	誤
付与された	正	104	10
ラベル	誤	9	77

5. おわりに

本研究では，日本語学習者を対象として，作文授業での誤りを自動的に検出することをめざし，これまでの授業の作文情報(作文および添削結果)を蓄積し，自然言語処理・機械学習することで，計算機に誤り検出ルール自動獲得させ，これを用いて誤りを指摘することを試みた．実験により，実際に収集したデータ 200 件に対して，約 91%正しく判定することを示した．

今後は，正解率の向上および検出できる誤りの種類を増やす方法について検討する．

参考文献

(1) 国際交流基金，海外の日本語教育の現状―日本語教育機関調査より―，独立行政法人国際交流基金（2015）

(2) Liu, X., Han, B. and Zhou, M.: Correcting Verb Selection Errors for ESL with the Perceptron, Proceedings of CICLing, pp.411-423 (2011)

(3) Rozovskaya, A. and Roth, D.: Algorithm Selection and Model Adaptation ESL Correction Tasks, Proceedings of ACL,pp.924-933 (2011)

(4) Dahlmeier, D. and Ng, H. T.: Grammatical Error Correction with Alternating Structure Optimization, Proceedings of ACL-HLT, pp. 915-923 (2011)

(5) 谷之口優人，杉野勝也，佐藤俊也，絹川博之，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，第 10 回情報科学技術フォーラム（FIT2011）第 3 分冊，pp.789-790 (2011)

(6) 大木環美，大山浩美，北内啓，末永高志，松本裕治，非日本語母国語話者の作成するシステム開発文書を対象とした助詞の誤用判定，言語処理学会第 17 回年次大会発表論文集，pp. 1047-1050 (2011)

(7) 橋本利典，島田静雄，外国人の書いた文章の助詞使用誤りの抽出，情報処理学会　NL 研 117-2，pp. 9-14 (1997)

(8) 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報告，pp. 107-112，情報処理学会 (2007)

(9) Oyama, H. and Matsumoto, Y.: Automatic Error Detection Method for Japanese Case Particles in Japanese Language Learners, In Corpus , ICT, and Language Education , pp. 235-245 (2010)

(10) 笠原誠司，藤野拓也，小町守，永田昌明，松本裕治，日本語学習者の誤り傾向を反映した格助詞訂正．言語処理学会第 18 回年次大会，pp. 14-17 (2012)

(11) 今枝恒治，河合敦夫，石川裕司，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，情報処理学会研究報告　コンピュータと教育研究会報告 pp.39-46 (2003)

(12) Li ZHANG & Hidehiko KITA, A Collaborative Learning Method Using Learner-participation Database, International Workshop on Regional Innovation Studies, pp.34-37 (2017)

(13) 石井皓太，張莉，北英彦，誤りを用いた日本語学習システム，2018PC カンファレンス論文集，pp. 41-44 (2018)

(14) Proofreading API, https://a3rt.recruit-tech.co.jp/product/proofreadingAPI/ (2020 年 2 月閲覧)

(15) jWriter 学習者作文評価システム, https://jreadability.net/jwriter/ (2020 年 2 月閲覧)

(16) 趙艶，高瀬治彦，北英彦，機械学習による日本語学習者の作文からの誤り検出－1 文節内の文法誤りの検出－，知能と情報，投稿中

(17) CaboCha，https://taku910.github.io/cabocha/ (2020 年 2 月閲覧)