日本語学習者の作文からの誤り検出
—決定木からの誤り種類の検出に関する検討—

外崎海斗・村田紘基・高瀬治彦・北英彦
Email: 422m232@m.mie-u.ac.jp

三重大学大学院工学研究科電気電子工学専攻

◎Key Words　自然言語処理，日本語作文授業，誤り検出

1. はじめに

近年，第二言語として日本語を学ぶ外国人が増加している。それに対し，日本語教師の増加はそれほどではなく，日本語学習の支援が必要とされている[(1)]。言語学習では読む力・書く力・話す力・聞く力の涵養が必要である。本稿では，書く力の涵養に対して，効果が高いが，教師への負担が大きい作文授業に注目する。作文授業の中で講師への負担が大きいのは，提出された作文の添削である。添削が不十分であったり不正確であったりすると，学習効果が上がらないが，多人数講義で正確な添削を十分に行うのは，教師一人では困難である。

このような状況を改善するために，計算機を用いたさまざまな支援が試みられている。宇佐美らは，作文指導を行う教師を支援することにより，間接的に学習者の学習を支援することを試みた[(2)]．この間接的に学習者の学習を支援する手法では，作文とそれに付ける添削結果をデータベースに蓄積し，教師の誤用分析を支援するが，学習者を直接支援するまでに至っていない。また，近年は，大規模言語モデルを利用した，校正が高い性能を示している。その反面，あらゆる誤りを訂正するため，学習範囲を超えた指摘がされたり，訂正理由が学習した内容に沿っていなかったりして，作文授業に用いるのには難がある。

そこで本稿では，学習内容に則した添削を行うシステムの構築を目標に，誤りの検出およびその理由付けを試みる。

2. 作文の誤り指摘

前節で述べたように，大規模言語モデルを使った作文校正は，精度は高いもののその理由付けにおいて難がある。そこで，決定木の学習に基づき誤り判定を行う手法[(3)]に着目し，得られた決定木の解釈を試みる。

この手法では，与えられた作文を解析し，作文を形態素・文節に分割する。文節ごとに，そこに含まれる形態素の情報を特徴量として抽出し，それに正誤のラベルを付けながら収集したものを機械学習することにより，正誤の判定をする正誤判定器を，決定木を用いて構築する。

しかし，作文を形態素・文節に分割する際に用いるツールは，文法的に正しい文を対象にしており，誤りを含む作文を解析する場合，不具合が発生する可能性がある。そのようにして得られた特徴量を用い機械学習を行った場合，学習した正誤判定器にも不具合が波及するだろう。

実際，学習の結果得られた正誤判定器によりある程度の正誤判定の精度を得ることができたが，得られた決定木を解析した結果，学習者に示す誤りの理由としては不適切な理由(文節内の形態素数など)による正誤判定が行われている事例が散見した。

3. 提案法

前節での問題点をふまえ，不適切な理由による正誤判定を軽減する手法について，本節で検討する。

まず，不適切な理由による正誤判定が，不適切な形態素・文節情報を学習した結果と考え，学習用のデータからそのようなデータを除く。形態素・文節情報は主に形態素解析により得ている。ここで文献(4)では，形態素解析の際に使用しているコストに着目することで，形態素解析結果に不具合があるかどうかを判定できる可能性が示されている。これに基づき，機械学習に用いるデータから不適切なデータ除くことは可能だろう。

また，文献(3)では，学習器としてランダムフォレストが好ましいとしている。また，文献(5)では，同じく決定木に基づく手法である XGboost を用いることで，判定精度の向上と，判定理由の解釈の容易化ができると報告している。

これらをふまえて，本稿では不適切な学習データを除き XGboost で学習した際の，判定精度・判定理由について，実験をつうじ確認する。

4. 実験

本節では，不適切な形態素解析結果を除くことの，判定精度・判定理由への影響を，簡単な実験をつうじて確認する。

4.1 実験方法と条件

実験には，作文対訳データベースと，文献(3)で使用していた実際の日本語授業の作文から，正しい文節 49 個，誤りの文節 39 個を含むデータを用いた。不適切な形態素解析結果が含まれている文節は，文献(4)による判定，または，日本人の著者二人による判定の二種類の手法により除いた。

4.2 誤り指摘の正答率

本節では①すべてのデータを用いた場合，②人手で不適切なデータを除いた場合，③文献(4)の手法により不適切なデータを除いた場合について， XGboost で学習し行った正誤判断の結果を表 1,2,3 にそれぞれ示す。なお，正答率は① 60%，② 68%，③ 59%となった。人手で不適切なデータを除くことで正答率は向上したが，文献(4)の手法を用いた場合は，わずかに悪化した。この結果は，不適切なデータを用いずに学習することの効果を示しているが，不適切なデータの判定方法については問題があるこ

とを示している。

表１①の混合行列

		予測結果	
		正	誤
文節の正誤	正	32	17
	誤	18	21

表２②の混合行列

	予測結果		
		正	誤
文節の正誤	正	38	11
	誤	15	18

表３③の混合行列

	予測結果		
		正	誤
文節の正誤	正	28	17
	誤	13	16

4.3 指摘内容の解釈

本節では，正答率が改善した②の場合について，得られた決定木を可視化し，誤りの指摘するルールの解釈を試みた。図１，２にその一例を示す。図中，f0～f2 は判定対象の文節内の形態素(先頭から順に 0, 1, 2)から抽出した特徴量を意味する。特徴量の数値化の規則[3]と合わせ解釈すると，決定木から抽出できた判断条件は 3 つであった。

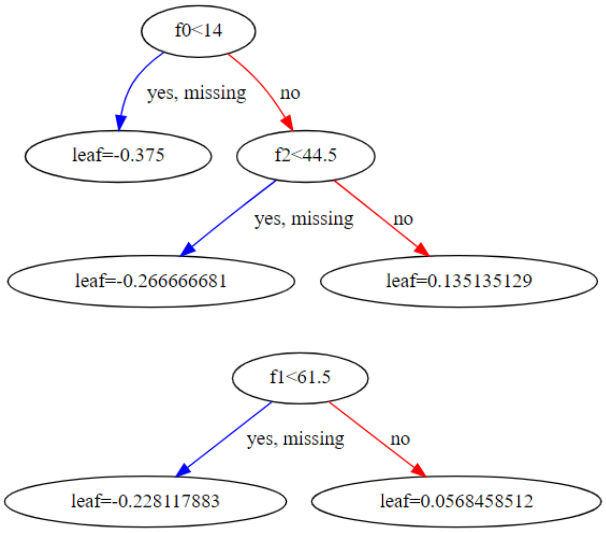

図１　決定木

(1) 文節の 1 つ目の形態素が名詞か接頭語であるか
(2) 文節の 2 つ目の形態素が名詞，接頭語，動詞，形容詞，副詞，接続助詞のいずれかであるか
(3) 文節中の 3 つ目の形態素が名詞，接頭語，動詞，形容詞のいずれかであるか

この判断条件から得られるルールはいずれも依然として，学習者に示すのには不適切な条件であった。しかし，形態素数のような文法的に意味のなさそうな情報ではなく，品詞の種類に基づいた判定をしており改善の兆しが見られる。

5. おわりに

本研究では，学習内容に則した添削を行うシステムの構築を目標に，誤りの検出およびその理由付けを試みた。その結果，不適切なデータを用いずに学習することの効果はあったが，不適切なデータの判定方法については問題があった。

また，今回決定木から抽出できたルールは，いずれも，学習者に示すのには不適切な条件ではあるが，改善の兆しが見られた。

参考文献

(1) 国際交流基金：“海外の日本語教育の現状”, pp.9-11 (2020).
(2) Usami Y., Yarimizu K.: “Design of XECS (XML-based Essay Correction System): Effects and implications”, In Proceedings of the CASTEL-J in Hawaii 2007, pp. 182-184 (2007)
(3) 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，32 巻，5 号，pp. 887-890（2020）
(4) 村田紘基，趙艶，高瀬治彦，北英彦：“日本語学習者の作文からの誤り検出 —形態素解析時のラティス値による誤り検出—”, PC カンファレンス 2022，pp. 233-234 (2022)
(5) 村田紘基，趙艶，高瀬治彦，北英彦：“日本語学習者の作文からの誤り検出 —決定木に基づく検出法に関する検討—”, PC カンファレンス 2021，pp. 244-245 (2021)