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—

村田紘基*1・趙艷*1・高瀬治彦*1・北英彦*1
Email: 421m242@m.mie-u.ac.jp

*1：三重大学大学院工学研究科電気電子工学専攻

◎Key Words　誤り検出，作文添削，形態素解析

1. はじめに

近年，日本語教師の増加対して日本語学習者の増加が大きい。日本語学習の支援が必要とされているといえ，特に負担の大きい作文授業に支援が必要である。このような支援としてさまざまな試みがされているが，本研究では，計算機による作文の自動添削に着目する。

これまでにも，日本語の文章を自動添削（校正）するシステムは提供されてきたが，日本語学習者の作文授業に対して十分に誤りを検出できない。例えば，リクルート社の提供している A3RT[1]では，「今度の休みは、国に帰ようと思います」という誤りを含んだ文章に対して，誤りを指摘しない。そこで我々は，授業中に発生した誤りをもとに決定木を機械学習することで作文中の誤りを自動で検出できるシステムを構築した。しかし，この手法では作文を形態素解析器によりある程度正しく形態素に分割できていないと効果を発揮しない。

そこで，形態素解析か破綻している文節を検出し，それを除外することで，誤りの検出能力の向上をめざす。本稿ではその第一歩として，形態素解析時のラティス値を分析することで，形態素解析が破綻している文節の検出を試みる。

2. 作文の誤り指摘

文献(2)の手法では，与えられた作文を形態素解析・係り受け解析し，その結果得られた文節・形態素の情報を特徴量として抽出し，それを収集し機械学習することにより，正誤の判定を行う。しかし，文中に誤りが存在すると正しく形態素に分割できず，機械学習は効果を発揮しない。形態素解析では，さまざまな文法情報をもとに，解析を行うため，誤った文に対して解析が破綻する可能性がある。実際，よく用いられる形態素解析器 Mecab では，多少の誤りが含まれた文でも解析結果が破綻することは少ないが，学習者の作文に含まれる誤りには，破綻した形態素解析結果をもたらすものも含まれる。例えば，「あなたは」という文節を形態素解析すると「あなた」「は」という 3 つの形態素に分けられる。しかし，「あたなは」という学習者による誤りを含んだ文節を解析した場合，「あ」「た」「な」「は」と 4 形態素に分かれてしまう。
機械学習による誤文節の検出性能の向上のためには，このような文節を事前に取り除くのが良いと考える。そこで次の章で，このような形態素解析の破綻が生じた文節を検出する方法について提案する。

3. 提案法

形態素解析が破綻しているかどうかを判断するために，形態素解析の際に計算されているラティス値に着目する。ラティス値とは単語自体の表れにくさである生起コストと単語同士の連続しにくさである連結コストの累積の値である。形態素解析では一般に，この値を考えうるすべての分割パターンに対して計算し，最も値が小さいものを解析結果として出力する。このとき，形態素解析が破綻している場合には，どの分割候補も不自然な分割があるため，分割候補間のラティス値の差は小さくなるだろう。

そこで，形態素解析時のラティス値を分析することで形態素解析が破綻している文節を事前に検出する。例えば，「あなたは」と「あたなは」について，Mecab を使って形態素解析した際の，ラティス値が最も小さい 3 件分の結果は，表 1 のようになる。正しい結果が得られた「あなたは」の場合は，第 1 位と第 2 位の差が，第 2 位と第 3 位の差に比して大きい。それに対し，破綻した結果になっている「あたなは」の場合は，これらの差は「あなたは」の場合比べて顕著ではない。

表 1「あなたは」と「あたなは」のラティス値

候補	ラティス値
あなた(代名詞)+は(助詞)	6, 789
あな(名詞)+た(助動詞)+は(助詞)	17, 967
あなた(代名詞)+は(動詞)	18, 169
あ(フィラー)+な(助動詞)+た(助詞)+は(助詞)	18, 988
あ(感動詞)+な(助動詞)+た(助詞)+は(助詞)	20, 070
あた(動詞)+な(助詞)+は(助詞)	22, 349

この考えに基づいた形態素解析に破綻した文節の具体的な検出手順は以下のとおりである。

(1) 与えられた文を文節単位に分割する。
(2) 各文節に対して，形態素解析を行う。このとき，ラティス値の小さい順に n 通りの分析結果を得る。
(3) 各文節の n 番目に小さいラティス値と一番小さいラティス値の差を求める。
(4) しきい値を決め，ラティス値の差がしきい値以下の文節を，形態素解析が破綻した文節と判定する。

次の章では，この手順により形態素解析が破綻している文節を正しく検出できるかを確かめる。

4. 実験

4.1 概要

ここでは, 3 章で提案した手法によりの形態素解析が破綻している文節を検出できるのかを，簡単な実験を通じて確認する。

4.2 実験方法と実験条件

実験には，文献(2)と同様の，正しい文節 38 個，誤りの文節 60 個を含むデータを用いた。形態素解析結果が破綻している文節は，日本人の著者二人が共通で破綻していると判断したものを選んだ。共通で破綻していると判断した文節は 15 個であった。

実験の方法は以下のとおりである。3 章で示した手順における n を 5 とし，ラティス値の差を①2 番目−1 番目，②3 番目−1 番目，③4 番目−1 番目，④5 番目−1 番目の 4 種類求めた。

4.3 結果

図 1，2 に結果を示す。

まず，ラティス値の値と破綻の有無の関係を調べる。図 1 は, ラティス値の①2 番目−1 番目が最も小さい 7 件分の文節の形態素解析の成否とラティス値の差を示したグラフある。図 1 より，破綻している 15 文節のうち 5 文節がこの 7 件に入っており，破綻した文節では形態素解析において候補間のラティス値の差が小さくなるという予想はおおむね正しいといえる。しかし, 形態素解析が破綻している文節の中に正しい文節や誤りだが形態素解析は破綻していない文節が含まれているため，しきい値により判定することは難しいと考えられる。

次に，ラティス値の値そのものではなく，順位と破綻の有無の関係を調べる。図 2 は，前節①から④それぞれの場合について，差の昇順に文節を並べ，上位から 7 文節ずつの 14 グループに分割し，それぞれのグループ内で日本人が破綻していると判定した文節の数を数えたものである。凡例の①〜④は前節の①〜④を表している。横軸は文節を 7 個含むグループを表しており，左に行くほどそれぞれのラティス値の差が小さいグループで，右に行くほどラティス値の差が大きいグループである。縦軸はそれぞれのグループに含まれる形態素解析が破綻している文節の個数を表している。例えば，横軸 1 の①のグラフは，最も小さいラティス値と 2 番目に小さいラティス値の差が小さい 7 個の文節の中に，形態素解析が破綻している文節が 5 個あることを示している。図 2 より，ラティス値の差が大きくなるにつれて形態素解析が破綻している文節数は少なくなっているため，ラティス値による検出自体は有効であると考えられる。

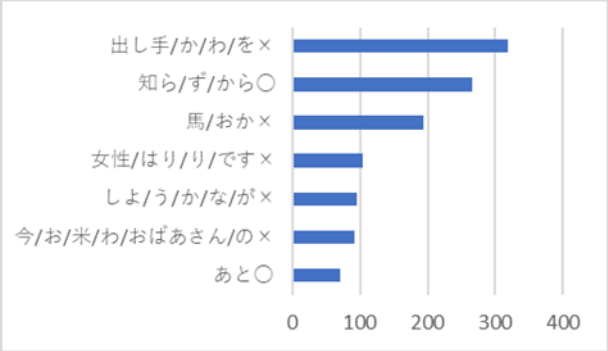

図 1　2 番目−1 番目のラティス値の差

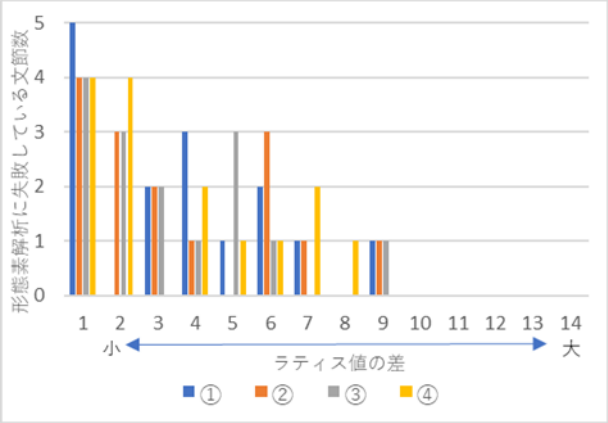

図 2　ラティス値の差に対する形態素解析が破綻している文節数

5. おわりに

本研究では，形態素解析時のラティス値を用いて形態素解析が破綻している文節の検出を試みた。

今回の実験ではあるしきい値を持って形態素解析が破綻している文節を完全に検出することは難しいが，ラティス値の差と誤り文節の数に負の相関があり，ラティス値を用いることの有効性は確認できた。

今後は，今回試した方法以外でのラティス値の活用や，いくつかの形態素解析が破綻した文節を除いた状態での作文の誤り検出を行う。

参考文献

(1) Proofreading API：“https://a3rt.recruit-tech.co.jp/product/proofreadingAPI/”, (accessed July 15, 2021).

(2) 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）