E-003

外国人を対象とした日本語の助詞の習得における個人の誤り傾向に応じた問題文の生成

Generating Exercise Sentences Based on Individual Error Trends in the Acquisition of Japanese Particles by Foreigners

蔡 宇鋒[1)]　望月 久稔[1)]
Yufeng CAI　Hisatoshi MOCHIZUKI

1　はじめに

日本に在留する外国人は年々増加し，令和 5 年 6 月の時点で日本の人口の約 2.6 %である 322 万人程度であり [1]，今後もさらに増加すると予想できる．しかし，日本語を支援できる教員は不足しており，十分な支援を受けられない外国人が多いため [2]，言語処理技術を用いた日本語学習用の問題文の生成は学習者にとって有用である．また，人により誤りやすい傾向は異なるため [3]，個人の誤りを考慮することで，学習効率の良い問題文を生成できると考える．本研究では日本語学習者が誤りやすい助詞 [4] を対象とし，誤用する確率を用いて個人に適した問題文を生成する．

2　個人の誤り傾向に応じた問題文の生成

全学習者と個人の誤り傾向から，助詞ごとに個人の練習する重要度を求め，その重要度に基づいて問題文を生成する．

2.1　助詞の重要度

個人が各助詞を練習する重要度を定義するために，個人が誤りやすい助詞に加えて，全学習者が誤りやすい助詞は個人にとっても潜在的に誤りやすいと予想でき，その助詞を練習することは重要であると考える．よって，助詞の誤りにおける誤用と正用のペアを p とし，問題文の生成に必要な助詞の重要度 $W(p)$ を式 (1) で定義する．$W(p)$ は全学習者の誤用した確率 $A(p)$ と個人の誤りやすさを数値化した $I(p)$ の和である．

$$W(p) = A(p) + I(p) \tag{1}$$

$A(p)$ は NAIST 誤用コーパス [4][5] を用いて求める．NAIST 誤用コーパスは，日本語を学ぶ外国人が書いた誤用を含む作文，教師が添削した正しい作文，誤用の種類の三つが含まれる．ある p に対して全学習者が誤る確率を $A(p)$ とする．

個人が誤った助詞の傾向を捉えるために，全学習者と比較した場合の個人の誤りやすさ $I(p)$ を式 (2) で定義する．$I(p)$ は個人の誤りやすさの方向 $D(p)$ と誤りやすさの程度 $E(p)$ の積で求める．

$$I(p) = D(p)E(p) \tag{2}$$

誤りやすさの方向 $D(p)$ を式 (3) で定義する．

$$D(p) = \frac{g(i,p) - g(a,p)}{|g(i,p) - g(a,p)|} \tag{3}$$

$g(i,p)$ と $g(a,p)$ はそれぞれ個人 i と全学習者 a が p を誤用した確率である．$D(p)$ は個人と全学習者の確率差の正負を表し，正の場合は全学習者と比較して，個人が誤りやすい，負の場合は誤りにくいことを示す．表 1 の確率の例で，p が（が，は）の場合，個人と全学習者の誤用確率 $g(i,p)$ と $g(a,p)$ はそれぞれ 0.5，0.4 である．式 (3) より，$D($(が，は)$)$ は 1 であり，全学習者より，個人が（が，は）を誤りやすいことを示す．

表 1　全学習者と個人の誤用した確率の例

助詞のペア p	確率	
	$g(i,p)$	$g(a,p)$
(が，は)	0.5	0.4
(で，に)	0.4	0.4

次に誤りやすさの程度 $E(p)$ を式 (4) で定義する．

$$E(p) = \frac{|g(i,p) - g(a,p)|}{\sum |g(i,k) - g(a,k)|} \tag{4}$$

$E(p)$ は個人の p の誤りやすさの程度を表す．式 (4) の分子は個人と全学習者の誤用した確率の差（以降は確率差と表す）の絶対値であり，分母は個人が誤った全ての助詞における確率差の総和である．$E(p)$ の範囲は 0〜1 であり，値が 0 に近ければ，個人にとって p を誤用する確率は全学習者と近く，値が 1 に近ければ，全学習者と大きく離れることを示す．

$D(p)$ と $E(p)$ より，$I(p)$ の値が −1 に近ければ，個人が全学習者と比較して p を誤用しにくく，1 に近ければ，誤用しやすいことを示す．また，$I(p)$ が 0 の時，全学習者と個人が p において特徴が同じであることを示す．

以上の定義をまとめ，$W(p)$ の構成を以下の図 1 で示す．$W(p)$ は全学習者の誤用確率 $A(p)$ と個人の誤りやすさ $I(p)$ からなり，$I(p)$ は個人と全学習者の誤用確率の差を利用して個人の誤りやすさの方向 $D(p)$ と程度 $E(p)$ から定義した．p に対して，全学習者が誤りやすく，さらに個人が全学習者より誤りやすければ，$W(p)$ が高くなる．

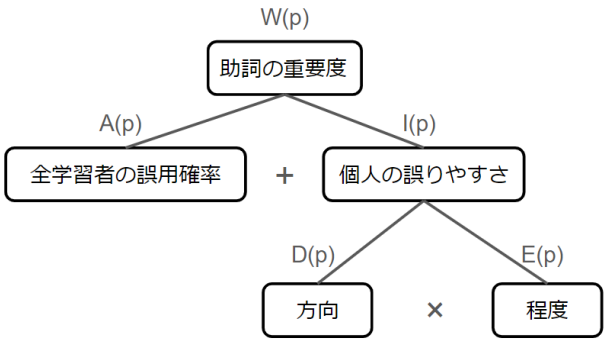

図 1　助詞重要度 $W(p)$ の構成図

1）　大阪教育大学　Osaka Kyoiku University

例えば，表 1 より，全学習者の誤用確率 $g(a,(\text{が},\text{は}))$ と $g(a,(\text{で},\text{に}))$ がともに 0.4 であり，個人の誤用確率 $g(i,(\text{が},\text{は}))$ と $g(i,(\text{で},\text{に}))$ がそれぞれ 0.5 と 0.4 である場合，以上の定義より，(で，に) において $A(p)$，$I(p)$，$W(p)$ はそれぞれ 0.4，0，0.4 である．個人の誤りやすさ $I(p)$ は 0 であるため，$W(p)$ は $A(p)$ と同じ値である．(が，は) において，$A(p)$，$I(p)$，$W(p)$ はそれぞれ 0.4，1，1.4 である．$I(p)$ が 1 であるため，個人が全学習者より p を誤りやすいことを示し，$I(p)$ により $W(p)$ は $A(p)$ より高くなる．よって，個人と全学習者の誤用確率の差があれば，$I(p)$ により $W(p)$ は $A(p)$ より増減する．

2.2 重要度による問題文の生成

重要度 $W(p)$ に応じた問題文を生成するために，$W(p)$ を確率に変換する．$W(p)$ に負の値が存在するため，まず重要度を正規化し，値を 0〜1 の範囲にする．次に，正規化した重要度を問題文を生成するための確率に変換する．例えば，助詞のペア $p1, p2, p3$ の重要度を式 (1) の $W(p)$ で計算した結果が $(0.3, 0.1, -0.2)$ である場合，正規化すると $(1, 0.6, 0)$ であり，確率に変換すると $(0.625, 0.375, 0)$ である．このとき，$p1, p2, p3$ に関する問題文はそれぞれ 0.625，0.375，0 の確率で生成する．

問題文の生成は，NAIST 誤用コーパスから助詞の誤用に関する文章を抽出し，誤用した部分を省略し，誤用と正用を選択肢とした選択問題を生成する．例を (5) に示す．[’ で’，’ に’] の 2 つの助詞から，正解を選ぶ問題である．

公園 [’ で’，’ に’] 散歩する。 (5)

3 生成した問題文の評価

NAIST 誤用コーパスの 313 文の作文から無作為に 30 文を抽出し，助詞の誤用に限定して生成した問題文を評価する．

式 (1) の $A(p)$, $I(p)$, $W(p)$ のそれぞれにより生成した問題文が個人の誤り傾向を捉えられたものであるかを評価するために，生成した問題文に含まれる助詞の割合と個人の誤用する確率の相関係数 R と P 値を求める．R が 0.2 以上であれば，生成した助詞の割合と個人の誤用する確率に正相関があると言える．正相関の中に P 値が 0.05 未満の場合，R は偶然に得られた結果ではないと言える．実験の結果を表 2 に示す．

表 2 より，全学習者の誤用する確率 $A(p)$ を用いた場合，$0 \leqq P < 0.05$ かつ $0.2 < R \leqq 1$ の割合 (以降は正相関の割合と表す) は 0.23 である．正相関でない割合は $P \geq 0.05$ かつ $0.2 < R \leqq 1$ の割合と $-1 \leqq R \leqq 0.2$ の割合の合計 0.63 であった．よって，$A(p)$ を用いた生成は 23% の学習者の誤り傾向を捉えられた．全学習者が誤用しやすい助詞は個人も誤用する確率が高いため，$A(p)$ を用いて個人の潜在的な誤り傾向を予測できると考える．しかし，個人の間には差があるため，個人の誤り傾向を捉えられなかった割合は 0.63 であった．よって，個人の誤り傾向を捉えるために，$A(p)$ のみでは不十分である．また，全てのモデルにおいて，相関係数が未取得の割合は 0.13 であった．学習者が誤った助詞の数が少ない場合，すべての助詞を誤用する確率が同じであるため，相関係数を取得できなかった．

表 2 各モデルにより生成した問題文に含まれる助詞の割合と個人の誤用する確率の相関係数

モデル	相関係数 R			
	$(0.2 < r \leqq 1)$		$(-1 \leqq r \leqq 0.2)$	未取得
	$P \geq 0.05$	$0 \leqq P < 0.05$		
$A(p)$	0.33	0.23	0.30	0.13
$I(p)$	0.27	0.53	0.06	0.13
$W(p)$	0.00	0.87	0.00	0.13

次に，個人の誤りやすさ $I(p)$ を用いた場合，正相関の割合，正相関でない割合はそれぞれ 0.53 と 0.33 であった．$A(p)$ のみを用いた場合に比べて正相関は 0.30 増加した．よって，個人と全学習者の誤用した確率の差を利用する場合はさらに個人の誤り傾向を捉えることができた．正相関でない割合が 0 にならなかった理由は個人と全学習者の誤用した確率が同じ場合の $I(p)$ の値が 0 になり，個人の誤りの傾向を捉えられなかったからであると考える．

最後に $A(p)$ と $I(p)$ を結合した $W(p)$ を用いた場合，正相関の割合，正相関でない割合はそれぞれ 0.87 と 0.00 であった．正相関でない割合が 0 になり，正相関の割合は $A(p)$，$I(p)$ のみを用いた場合に比べて最も高かった．$A(p)$ と $I(p)$ を組み合わせることにより，個人と全学習者の誤用した確率に差がない場合，$I(p)$ の値は 0 であるが，$A(p)$ により問題文を生成できた．よって，個人の誤り傾向を捉えた問題文を生成できたと考える．

4 おわりに

個人の誤り傾向を考慮した問題文を生成する方法を提案し，個人と全学習者の確率差を重視することで学習者の誤りの特徴を捉えられることが分かった．今後の課題として，学習効率の評価が挙げられる．

参考文献

[1] 在留外国人最多 322 万人 23 年 6 月、特定技能が 4 万人増，入手先<https://www.nikkei.com/article/DGXZQOUA123GQ0S3A011C2000000/>, (参照 2023-12-12).

[2] 日本語教育関係 参考データ集，入手先<https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/nihongo/nihongo_117/pdf/93833701_08.pdf>, (参照 2024-1-7)

[3] 若井誠二，岩澤和宏，WAKAI Seiji and IWAZAWA Kazuhiro：ハンガリー人日本語学習者のビリーフス. 日本語国際センター紀要，Vol.14，pp.123〜140(2004).

[4] 大山浩美，小町守，松本裕治：日本語学習者の作文における誤用タイプの階層的アノテーションに基づく機械学習による自動分類，自然言語処理，Vol.23，No.2，pp.195-225，(2016).

[5] Hiromi Oyama, Mamoru Komachi and Yuji Matsumoto:Towards automatic error type classification of Japanese language learners’writings,In Proceedings of the 27th Pacific Asia Conference on Language Information and Computation (PACLIC 27), pp.163–172, (2013).

[6] 蔡宇鋒，望月久稔：日本語学習者を対象とした助詞に関する誤用問題文の生成，情報処理学会第 86 回全国大会講演論文集，pp923-924，2024.