5W-09

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
Generating Particle Misuse Questions for Japanese Language Learners

蔡 宇鋒[1)]　　望月 久稔[1)]
CAI YUFENG　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人が多い[2]．そこで，言語処理技術を用いて日本語学習用の問題文を生成すること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．そこで，本研究では日本語学習者が最も誤りやすい助詞[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　個人と全学習者の誤用した助詞の分布による問題文の

まず，NAIST誤用コーパス[4][5]を用いて，全学習者と個人が誤用した助詞の分布を作成する．次に，分布から個人が誤った助詞の重要度を求め，重要度により問題文のリストを生成する．

2.1　個人と全学習者の誤用した助詞の分布の作成

個人と全学習者の誤用した助詞の傾向を捉えるために，それぞれが誤った助詞の誤用と正用のペアをNAIST誤用コーパスから抽出して，誤用する確率を求める．確率を用いて，誤用した助詞の分布を作成する．

誤用した助詞の分布の例を表1に示す．全学習者とある個人が誤用した(が，は)の確率はそれぞれ0.098と0.050である．全学習者と比較すると，その個人の誤用する確率は0.048低いため，(が，は)を誤用しにくい傾向がある．また，(で，に)において，個人の誤用する確率は全学習者より0.152高いため，誤用しやすい傾向がある．

2.2　個人が誤った助詞の重要度の定義

生成する問題文に用いる助詞のペアを抽出するために，全学習者と個人それぞれの誤用する確率を用いて，助詞の重要度を定義する．個人の誤りやすい助詞に加えて，全学習者の誤りやすい助詞を練習することは重要であると考え，助詞のペア p の重要度 $W(p)$ を，個人の誤りやすさと全学習者の誤用する確率を考慮して式(1)で定義する．

$$W(p) = \frac{1}{2}(A(p) + I(p)) \tag{1}$$

全学習者の誤用する確率 $A(p)$ と個人の誤りやすさ $I(p)$ を足し合わせて助詞の重要度 $W(p)$ を求める．$I(p)$ は式(2)で求める．

$$I(p) = \frac{g(i,p) - g(a,p)}{\sum |g(i,k) - g(a,k)|} \tag{2}$$

1)　大阪教育大学

表1　誤用した助詞の分布

誤用した助詞のペア	確率	
	全学習者	個人
(が，は)	0.098	0.050
(で，に)	0.048	0.200

$I(p)$ は，全学習者と比較して個人の誤りやすさを表す．個人 i と全学習者 a それぞれにおける助詞のペア p の誤用した確率を $g(i,p)$ と $g(a,p)$ と表す．つまり，式(2)の分子は個人と全学習者の誤用した確率の差(以降，確率差と表す)である．また，分母は確率分布間の絶対差の総和である．よって，$I(p)$ の値が -1 に近ければ，全学習者と比較して誤用の確率が低いことを表し，1に近ければ誤用の確率が高いことを表す．

上記で定義した式(1)(2)の例を示す．表1より，$g(i,(\text{で，に}))$ と $g(i,(\text{が，は}))$ がそれぞれ0.200と0.050であり，$g(a,(\text{で，に}))$ と $g(a,(\text{が，は}))$ がそれぞれ0.048と0.098である場合，$I(p)$ は以下のように求められる．

$$I((\text{で，に})) = \frac{0.200 - 0.048}{|0.050 - 0.098| + |0.200 - 0.048|} = 0.76$$

$$I((\text{が，は})) = \frac{0.050 - 0.098}{|0.050 - 0.098| + |0.200 - 0.048|} = -0.24$$

次に，$I(p)$ を用いて $W(p)$ を求める．

$$W((\text{で，に})) = \frac{1}{2}(0.048 + 0.76) = 0.404$$

$$W((\text{が，は})) = \frac{1}{2}(0.098 + (-0.24)) = -0.071$$

重要度W(p)はそれぞれ0.404，−0.071である．よって，その学習者にとって(が，は)よりも(で，に)が重要であることを表す．

2.3　問題文の生成

2.2節で定義した重要度を用いて問題文のリストを生成する．まず，助詞のペアの重要度を用いて，個人が誤ったペアの中で最も重要度が高いデータをNAIST誤用コーパスから抽出し，誤用した助詞を省略した文を質問文として，誤用・正用の助詞を選択肢とした選択問題を生成する．例を(3)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。　　(3)

重要度が高い助詞のみを用いると，同じ助詞の問題文のみを生成する．そこで，一度生成した助詞の重要度を r 倍で更新し，同じ助詞の問題文の生成を抑制する．r を式(4)に示す．分母を指定した問題数，分子を残りの問題数とすることで，残りの問題数が少ないほど，同じ問題文の生成を抑制する．例えば，学習者が問題数を5と指定し，4問目まで進んだ場合を考える．式(4)より

表2　誤用数による確率差の変化

誤用数	確率		
	平均値	最大値	最小値
5	**0.22**	**0.44**	**0.14**
10	0.12	0.32	0.06
15	0.08	0.11	0.05
20	0.05	0.06	0.03
25	0.11	0.11	0.11
30	0.09	0.09	0.09
35	**0.02**	**0.02**	**0.02**

r の値は 5 分の 1 であるため，4 問目の助詞の重要度が 0.5 の場合，その助詞の重要度は 0.1 に更新され，生成が抑制される．

$$r = \frac{\text{残りの問題数}}{\text{指定した問題数}} \tag{4}$$

3　誤用数による確率差の変化と生成した問題文のリストの評価

NAIST 誤用コーパスから，学習者が誤った助詞の誤用と正用のペアを抽出して，提案した式 (1) による問題文リストの評価実験する．

3.1　誤用した助詞の回数による確率差の変化

NAIST 誤用コーパスを調査した結果より，誤用した回数 (以降，誤用数と表す) は 0 から 35 回までであったため，誤用数が 5 から 35 までである学習者らと全学習者の確率差の変化を考察する．まず，各誤用数における学習者の誤用した助詞の分布と全学習者の分布を比較し，各学習者が誤用した助詞の確率差の平均値を求める．そして，同じ誤用数における学習者らの中で求めた平均値の平均値，最大値，最小値を表 (2) に示す．結果より誤用数が増加するとともに，平均値，最大値，最小値は減少する傾向があり，個人が誤用する確率は小さくなる傾向がある．

3.2　生成した問題文のリストの評価

助詞を誤用した回数が 10，20，30 であった学習者の中からそれぞれランダムに 1 人 (以降，それぞれを **A，B，C** と表す) を抽出し評価する．まず，全学習者と個人の誤用した助詞の分布を用いて，2 章で提案した方法により問題文を 10 問生成する．正用と誤用のペアの生成した回数，全学習者と抽出した 3 人の確率分布を表 3 に示す．

表 3 の全学習者の確率より，誤用した確率が高い助詞のペアは (∗，の) と (が，は) の 2 つであり，ともに 0.098 であった (∗は助詞が必要ないことを表す)．学習者 A はこの 2 つを誤用したが，生成した回数は，ともに 0 であった．(∗，の) と (が，は) における B と全学習者の確率差を計算すると，ともに −0.048 で負であるため，全学習者より誤用が少なかった助詞を用いた問題文の生成を抑制できたと考える．

また，表 3 の確率より，A，B，C と全学習者の確率差が一番高い助詞のペアは (∗，と)，(∗，に)，(∗，の) である．それぞれの確率差は 0.193，0.189，0.232 であり，それぞれを生成した回数は 3，5，6 であった．よって，学習者 A，B，C と全学習者の確率差が高い助詞のペアは多く生成されることがわかった．全学習者の誤用した助詞の分布における確率の最大値は 0.098 であった．表 3 における (∗，の) の確率より，C の確率の最大値は 0.33 であり，全学習者の確率の 3.3 倍と高かった．よって，2.2 節で定義した式 (1) に基づいて生成した問題文は全学習者の傾向より個別の傾向を重視したと考えられる．

表3　生成した助詞のペアの回数と確率の分布

助詞のペア	生成した回数			確率			
	A	B	C	全学習者	A	B	C
∗，の	0	0	**6**	**0.098**	0	**0.05**	**0.33**
が，は	0	0	**1**	**0.098**	0	0.05	0.10
∗，は	0	0	**1**	0.062	0	0.05	0.10
∗，に	**2**	**5**	**1**	0.061	0.20	**0.25**	0.06
が，を	0	**1**	0	0.056	0	0.10	**0.03**
は，を	0	**1**	0	0.017	0	0.10	0
∗，も	**1**	0	0	0.014	**0.10**	0	0
に，は	**1**	0	0	0.013	0.10	0	0
は，も	**1**	0	0	0.013	0.10	0	0.03
が，も	**1**	0	0	0.012	0.10	0	0
も，を	**1**	0	0	0.008	0.10	0	0
∗，と	**3**	0	0	0.007	**0.20**	0	0
じゃ，では	0	0	**1**	0.004	0	0	0.06
で，には	0	**1**	0	0.003	0	0.05	0
から，で	0	**1**	0	0.003	0	0.05	0
から，の	0	**1**	0	0.001	0	0.05	0

さらに，表 3 の学習者 A，B，C の生成した回数より，それぞれの生成した助詞のペアは 7，6，5 種類であり誤用数が増えるとともにその種類が減った．3.1 節より，個人が誤用する回数が少ない場合，誤用する確率は高くなるため，2.2 節で定義した式 (1) より，生成する助詞の種類が増えると考えられる．

4　おわりに

本研究では全学習者と個人の誤用した助詞の分布を用いて，個別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．今後は数値の指標を用いて，生成する問題文による学習効率の変化を評価する．

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/>，(accessed 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>，(accessed 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).