K-016

Joining-in-type RALL システムにおける学習者の繰り返し回答に対する音声認識結果識別器の精度評価
Evaluation of a classifier of automatic speech recognition results for learners' repetitive responses in Joining-in-type RALL system

萱原 健太郎† 山本 誠一† 加藤 恒夫†
Kentaro Kayahara Seiichi Yamamoto Tsuneo Kato

1. はじめに

第二言語によるコミュニケーション能力の向上を支援するため，当研究室ではロボット 2 体を用いて学習者と英語による対話訓練を行う Joining-in-type RALL(JIT-RALL)システムを提案している[1]．最近の取り組みで，質問への回答を学習者に繰り返させることで，英語表現の習得が促進されるとともに，音声認識の精度劣化要因である発話中のフィルドポーズも削減されることが確認された[2]．誤認識を多く含む音声認識結果を用いて回答のレベルに応じたフィードバックを返すために開発した識別器について，回答の繰り返しによる精度改善の効果を検証する．

2. Joining-in-type RALL

JIT-RALL システムは学習者に模範的な対話を提示し，適切な発話を促すシステムである．システムは図 1 に示すように，学習者に対して教師役ロボット(R1)と生徒役ロボット(R2)を配置し，会話シナリオに沿って対話を進め，ロボット間の対話により質問とその回答例を学習者に提示する．続いてロボットが学習者に類似した質問をすることで，回答例に類似した適切な表現の使用を学習者に促す．

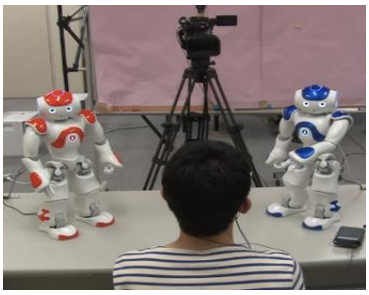

図 1　Joining-in-type RALL システム

3. 学習者のデータ収集

システムを Wizard-of-Oz 法を用いて制御し，学習者回答のデータ収集を 6 日間行う．2 日目と 3 日目に R2 の回答をリピートするタスクを設け，第二言語の聞き取り能力を測定することで，学習者を高難易度グループと低難易度グループに振り分ける．学習者は表 1 に示す対話シナリオのもと，質問に対して回答を行う．学習効果を評価するため，1 日目と 6 日目はそれぞれプレテスト，ポストテストとして同一の対話シナリオを実行する．

回答の繰り返しでは，システムは学習者に質問を 2 回行い，学習者は同一の内容を回答する．データ収集では，回答の繰り返しを要求する学習者群と回答の繰り返しを要求しない学習者群に分け，両群の回答を収集し比較する．以降，回答の繰り返しを要求しない学習者群のデータ収集をデータ収集 1，回答の繰り返しを要求する学習者群のデータ収集をデータ収集 2 とする．データ収集 1，データ収集 2 はいずれも日本語を母語とし，英語を第二言語とする日本人大学生を対象として行う．データ収集 1 では 21 歳から 25 歳の 11 名を対象に，高難易度グループには 6 名，低難易度グループには 5 名が振り分けられた．データ収集 2 では 21 歳から 24 歳の 12 名を対象とし，高難易度グループには 8 名，低難易度グループには 4 名が振り分けられた．グループ振り分け後の学習者の回答に対して表 2 に示す基準で主観評価を行った結果，高難易度グループで回答の繰り返しによる英語表現の習得が促進された[2]．本研究では JIT-RALL システムの実現に向けて発話の自動識別を行う識別器に対する回答の繰り返しによる影響をリスニング能力の高い高難易度グループの回答を用いて検証する．

表 1　対話シナリオの種類

種類	構成内容
高難易度シナリオ	無生物主語 5 問で構成 例) What do you think makes him go abroad?
低難易度シナリオ	時制 5 問で構成 例) If you have a child, when would you give your child a smartphone?

表 2　回答の主観評価基準

Category		評価概要
A	A-1	適切な表現かつ文法としても正しい
	A-2	適切な表現ながら文法誤りが含まれる
B	B	文法上正しいが R2 の回答に類似しない
C	C-1	見当違いな回答
	C-2	無回答または”I don’t know”等の回答

4. 回答の繰り返しによる学習者回答への影響

“I don’t know”等の定型表現の回答が分類される評価 C の回答を除いた主観評価 A もしくは B の回答における平均フィルドポーズ数は，データ収集 1 においてプレテストの 0.81 回からポストテストの 1.03 回へ 0.22 回増加に対し，データ収集 2 においてプレテストの 0.18 回からポストテストの 0.07 回へ 0.11 回減少した．これはデータ収集 2 で英語表現の習得が促進されたことにより，フィルドポーズ数が減少したと思われる．一方，平均単語数は，同じく主観評価 A もしくは B の表現で，データ収集 1 においてプレテ

† 同志社大学　Doshisha University

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

ストの 5.94 単語からポストテストの 6.26 単語へ 0.32 単語増加に対し，データ収集 2 においてプレテストの 6.42 単語からポストテストの 6.14 単語へ 0.28 単語減少した．

5. 学習者回答の音声認識結果識別器

JIT-RALL システムでは英語表現の習得を促進するため，学習者の回答のレベルに応じたフィードバックを返すことを想定している．その質問/回答のフローを図 2 に示す．システムは学習者の回答を音声認識し，認識結果をもとに学習者回答中の英語表現が R2 の回答中の英語表現と同一であるか否かを識別することで，学習者に回答のフィードバックを行う．R2 の回答中の英語表現と同一の場合は主観評価の A カテゴリ， R2 の回答中の英語表現と異なる場合は主観評価の B,C カテゴリに分類される．

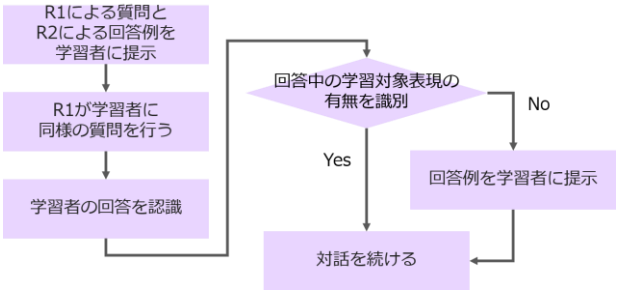

図 2　システムの質問/回答フロー

しかし学習者の第二言語音声の認識結果は，誤認識が多く含まれる．そこで当研究室では，学習者音声に対する認識結果が誤認識を多く含む場合でも正しく識別できるように，3 種類の識別手法の利用を検討した．3 種類の識別器に若干の性能差はあるが，誤認識と比較して識別率は比較的高く保たれていた（WER = 55.5%に対して，識別器は F1= 0.75）．ここでは単語単位の編集距離による識別器を用いた場合について検討する．

$$D(U_{R2}, U_L) = \frac{\#mismatched\ words}{max(\#words_{R2}, \#words_L)}$$

この編集距離による識別器では，閾値を用いて学習者の回答中における R2 回答中の英語表現の有無を識別する．編集距離が閾値より大きい場合，識別器は学習者の回答中に R2 回答中の英語表現が含まれない回答として識別する．

6. 回答の繰り返しによる識別器の精度改善評価

表 3，4 に，データ収集 1 とデータ収集 2 の 2 回目の回答に対する識別器の混同行列をそれぞれ示す．データ収集 1 における評価 A に対する F1 スコアはプレテストの 0.31 からポストテストの 0.90 へ 0.59 増加に対し，データ収集 2 における 2 回目の回答の評価 A に対する F1 スコアは，プレテストの 0.57 からポストテストの 0.68 へ 0.11 増加し，回答を繰り返すことで識別精度の向上が減少した．

また，音声認識結果における平均 WER(Word Error Rate)は，データ収集 1 においてプレテストの 58.9%からポストテストの 62.1%へ 3.2%上昇に対し，データ収集 2 においてプレテストの 51.4%からポストテストの 46.9%へ 4.5%低下し，回答の繰り返しにより平均的に音声認識の精度は向上した．

表 3　データ収集 1 における識別精度

		編集距離による識別評価			
		プレテスト		ポストテスト	
		A	B, C	A	B, C
主観評価	A	0.03	0.02	0.23	0.01
	B,C	0.13	0.81	0.03	0.72

表 4　データ収集 2 における識別精度

		編集距離による識別評価			
		プレテスト		ポストテスト	
		A	B, C	A	B, C
主観評価	A	0.15	0.10	0.29	0.25
	B,C	0.13	0.63	0.01	0.45

表 5　主観評価 A の回答の Word Error Rate[%]

	編集距離による識別評価			
	プレテスト		ポストテスト	
	A	B, C	A	B, C
データ収集 1	30.0	100.0	29.0	83.3
データ収集 2	45.9	54.7	45.2	79.8

一方，表 3,4 より主観評価で A と評価された発話につき，編集距離による識別評価で B，C と評価された回答の割合がデータ収集 1 では増加せず，データ収集 2 では大きく増加した．主観評価で A と評価された回答に対し，編集距離による識別の各評価で平均 WER を算出した結果を表 5 に示す． 編集距離による識別評価が B,C の回答は A の回答と比較して平均 WER が高く，主観評価 A の回答の増加に伴い音響的な特徴のばらつきが拡大し，音響モデルからのずれの大きな発話の WER が増加し，識別性能の劣化を招いていると考えられる．

7. おわりに

回答の繰り返しによる英語表現の習得の促進が確認された高難易度グループの回答において，音声認識の精度劣化の原因であるフィルドポーズ数が減少し，音声認識の精度が向上したが，編集距離による識別器の性能改善の大きさは低下した．また，正しい英語表現であるが認識誤りの多い回答は正しく識別できない割合が高く，これにより編集距離による識別器の性能改善の大きさが低下したと考えられる．この発話を正しく識別するために，より高性能な認識器とそれに基づく識別器が必要だと考えられる．今後は，認識器の高性能化よって識別性能の向上を図る予定である．

謝辞

本研究は JSPS 科研費 19K00927 から支援を受けた．

参考文献

[1] A.Khalifa, T.Kato, S.Yamamoto, “Joining-in-type Humanoid Robot Assisted Language Learning System”,Proc.LREC, pp.245-249 (2016)
[2] 山本一貴, 加藤恒夫, 山本誠一, “Joining-in-type RALL への回答の繰り返しの導入による第二言語学習効果の評価”, 電子情報通信学会総合大会,D-15-31 (2020)

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.