Vol.2022-SLP-143 No.4
2022/10/22

BARTを用いた音声認識誤り訂正のオンライン適応手法の検討

芦川 平[1,a)]　岩田 憲治[1]

概要：テレビ番組等の字幕制作において，音声認識技術を活用して，自動で字幕を生成する技術がある．近年，深層学習技術の発展により音声認識の精度は向上しているが，専門用語を含む発話や表記ゆれなどが原因で，音声認識の結果を人手で訂正する必要がある．しかし，認識結果を訂正しても，次の発話でも同じように訂正する必要があるため，修正コストが高い．そこで，今回，音声認識結果を訂正した場合に，即時に訂正結果が反映され，次の発話以降では，音声認識が同じような誤りをした場合に自動で訂正するオンライン適応手法を検討した．とくに，文訂正タスク等で近年採用されている事前学習モデル BART（Bidirectional and Auto-Regressive Transformers）を訂正モデルとして利用することを試みた．実験の結果，14 のニュース番組での字幕制作において，訂正モデルを用いないベースラインと比較した場合に，通常サイズの訂正モデルではエラー削減率が平均 2.28%，モデルサイズが大きい BART Large を用いた場合では，エラー削減率が平均 7.84%となり，訂正モデルの効果が確認できた．また，オンライン適応においても，類似の内容を分けて放送するニュース番組等では，本手法の有効性が確認できた．

1. はじめに

講演やテレビ番組の字幕制作において，音声認識技術を活用して，自動で字幕を生成する技術が開発されている [1]．近年，深層学習技術の発展により，音声認識の精度は，急速に向上しているが，背景音，言い間違い，専門用語を含む発話等が原因で，音声認識が誤る可能性がある．実際の字幕制作においても，音声認識だけで完全な字幕を制作するまでには至っておらず，音声認識をした後に，人手で確認し，誤っている箇所を訂正して，最終的な字幕を作成することが多い．

また，音声認識結果の訂正結果を利用し，言語モデル適応技術により，音声認識の精度を向上させることが知られている．ただし，適応するには，適応モデルを作成するためのデータと作成時間がかかるため，オンライン適応する（例えば，次の発話の結果に訂正結果を即時反映する）ことはできない．そのため，前の音声認識結果を修正しても，次の発話で同じように修正する必要があり，修正コストが高い．

そこで，本論文では，音声認識結果のエラーを訂正した場合に，即時に修正を学習し，次の発話以降では，音声認識が同じような誤りをした場合には自動で訂正する，オンライン適応手法を検討する．特に，近年，自然言語処理の文訂正などのタスクで良好な結果を出している事前学習モデル BART（Bidirectional and Auto-Regressive Transformers）を，音声認識のオンライン適応に応用することを検討する．

以下，本論文では，2 章に関連研究，3 章に提案手法，4 章に実験について記述する．

2. 関連研究

音声認識結果のエラー訂正に関するレビュー文献 [2] によれば，音声認識のエラーの要因は，以下の３つに依存するとされている．

- 話し手の音声の多様さ（経年劣化など．学習データにはすべてはない）
- 話し言葉の多様さ（アクセント，方言，ボキャブラリーなど）
- 学習データと評価データのミスマッチ

これらを解消するために，様々な研究が行われているが，音声認識のエラー訂正は主に，

(1) 音声認識結果のエラーを自動で検出する
(2) 検出したエラーを訂正する

という２段階に分けて，研究されている．

エラー検出手法としては，音声認識システム（ASR）が生成する特徴（信頼度や，コンフュージョンネットワーク），または，生成された音声認識結果（仮説単語列）の

[1] 株式会社東芝　研究開発センター　知能化システム研究所　メディアＡＩラボラトリー
[a)] taira.ashikawa@toshiba.co.jp

N-gram や品詞情報などの追加情報を利用し，その認識結果が正しいか誤りかを分類する手法がとられている [3]．一方，エラー訂正手法に関しては，エラー検出と比較して，既存研究は少ないが，複数の候補をユーザに提示する，または，訂正結果を利用して，レキシコンを適応するようにユーザに提案するといった，ユーザへのサポートツールが検討されている [4]．また，ユーザの介入が不要なものとして，文献 [5] において，音声認識結果に対して，N-Gram を用いて，スペル誤りを検出した後，訂正候補を生成し，コンテクストから最大スコアの候補を自動で選択して，訂正する手法が提案されている．

一方，最近の自然言語処理において，文訂正などのタスクで利用される BART が注目されている．BART は，文献 [6] で提案された事前学習モデルであり，BERT（Bidirectional Encoder Representations from Transformers）のエンコーダと，Autoregressive のデコーダで構成される．日本語の文訂正タスクにおいても，文献 [7] で評価されており，従来と比較して，良好な結果が得られている．また，文献 [8], [9] においても，BERT 等の深層学習モデルを用いた文訂正の研究が行われている．

音声認識結果に対するエラー訂正への BART の応用については，文献 [10] にて行われており，中国語の音声認識のエラー訂正に関して，従来手法より良好な結果が得られている．ただし，音声認識のエラー訂正に関するオンライン適応については，先述の N-gram を用いた文献 [5] があるが，BART を含む，最近のニューラルネットワークを用いた手法は存在しない．

一方，文献 [11] では，翻訳タスクに関して，人手の修正を即時フィードバックするために，DNN を用いた潜在表現から k-nearest を利用し置換するという，オンライン適応手法が提案されている．

3. 提案手法

前述の通り，エラー訂正のオンライン適応においては，N-gram ベースでの研究 [5] が行われている．しかし，N-gram ベースでは，長いコンテクストを考慮することが難しい．N=10 以上等の長いコンテクストを考慮した N-gram ベースでの訂正変換器も作成可能であるが，訂正変換器が過剰に複雑になってしまう．

一方，BART は，自動ノイズ化（挿入，削除，または置換）されたテキストからオリジナルテキストを再構築するように学習できるため，認識のエラー訂正にも応用可能である．また，BART で得られる潜在表現は，長いコンテキストを踏まえた特徴が得られる．そのため，訂正変換器に，BART の潜在表現を利用することによって，訂正変換器が過剰に複雑にならずに，長いコンテクストを考慮することができるため，精度が向上すると考えられる．

そこで，今回，音声認識結果のエラー訂正に対して，

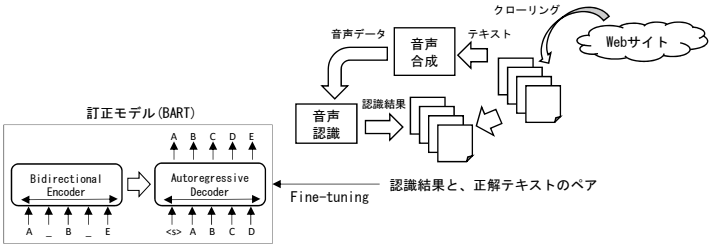

図 1 準備フェーズの流れ

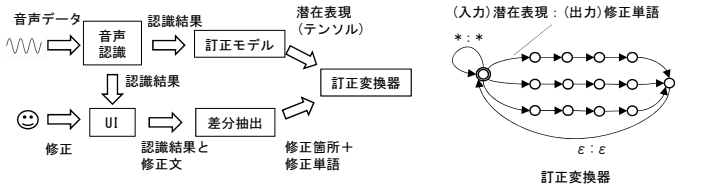

図 2 オンライン学習の流れ（左）と，訂正変換器（右）

BART を用いたオンライン適応手法を提案する．以下，準備フェーズ，オンライン学習フェーズ，及び，適応フェーズに分けて記載する．

まず，準備フェーズ（**図 1**）では，音声認識を行う前に以下を行う．

(1) 事前学習された BART モデルを準備する
(2) 音声認識結果と正解テキストのペアを収集する
(3) BART モデルに対して，2. のペアでファインチューニングを行い，音声認識結果訂正用の訂正モデル *MODEL* を作成する

オンライン学習フェーズ（**図 2**）では，以下を行う．

(1) 音声データの音声認識結果 *asr_text* と，訂正結果 *repaired_text* を取得する
(2) 訂正モデル *MODEL* に，音声認識結果 *asr_text* を入力し，潜在表現 *asr_rep* を取得する
(3) 音声認識結果 *asr_text* と，訂正結果 *repaired_text* の差分を求め，訂正位置と訂正単語を決定する
(4) 訂正位置の潜在表現 *asr_rep* を入力，訂正単語を出力とするアークを，訂正変換器 *fix_fst* に追加する

次に，適応フェーズでは，以下を行う．

(1) 音声データの音声認識結果 *asr_text* を取得する
(2) 訂正モデル *MODEL* に，音声認識結果 *asr_text* を入力し，潜在表現 *asr_rep* を取得する
(3) 音声認識結果の潜在表現 *asr_rep* と，訂正変換器 *fix_fst* のアークの入力の潜在表現 *fix_fst_input_rep* とのコサイン類似度を計算し，最短マッチする経路を探索する．
(4) 最短経路をたどり，アークの出力を訂正結果 *fix_result* として出力する

4. 実験

提案手法の効果を確認するため，まず BART を用いた

表 1 エラー訂正モデルの文字誤り率

ID	NAME	音声長	文字誤り率			誤り削減率	
			(1) ASR	(2) BART	(3) BART Large	BART	BART Large
1	朝のニュース 5 月 14 日	0:23:11	13.43	13.80	12.55	-2.70	6.59
2	昼のニュース 5 月 14 日	0:26:17	10.32	10.21	9.65	1.03	6.46
3	夜のニュース 5 月 14 日	0:40:19	18.72	19.41	18.84	-3.69	-0.62
4	昼のニュース 5 月 16 日	0:12:44	9.65	8.95	7.64	7.30	20.90
5	昼のニュース 5 月 17 日	0:15:29	11.24	10.64	10.00	5.34	11.05
6	夜のニュース 5 月 17 日	0:35:58	18.95	19.20	18.81	-1.30	0.73
7	朝のニュース 5 月 18 日	0:19:59	12.23	10.69	10.03	12.62	17.98
8	昼のニュース 5 月 18 日	0:22:02	11.65	9.83	9.21	15.57	20.95
9	夜のニュース 5 月 18 日	0:25:56	13.77	14.13	13.75	-2.61	0.12
10	朝のニュース 5 月 19 日	0:16:51	12.14	11.60	11.08	4.41	8.73
11	昼のニュース 5 月 19 日	0:26:43	12.10	11.66	10.72	3.63	11.39
12	昼のニュース 5 月 20 日	0:36:17	14.08	14.60	13.61	-3.71	3.31
13	朝のニュース 5 月 21 日	0:31:38	10.18	10.71	9.80	-5.18	3.76
14	朝のニュース 5 月 24 日	0:07:46	9.01	8.00	7.87	11.24	12.62
	平均	0:24:22	12.68	12.39	11.68	2.28	7.84

場合の訂正モデルの有効性を確認した．次に，オンライン適応時の評価を行った．なお，音声認識のモデルは，弊社作成のモデル [12] を利用した．

4.1 BART を用いたエラー訂正モデル

BART を用いた場合のエラー訂正モデルの有効性を確認するため，以下の事前学習モデル，学習データ，及び，評価データを利用した．

4.1.1 事前学習モデル

BART の事前学習モデルは，ウェブ上に公開されている，京都大学の日本語 BART モデル ver2.0 [*1] を利用した．このモデルは，日本語 Wikipedia （約 1800 万文）を用いて作成されており，Base モデル（エンコーダ・デコーダ 6 層，隠れユニット数 768），Large モデル（エンコーダ・デコーダ 12 層，隠れユニット数 1024）の 2 つのモデルが提供されている．

4.1.2 学習データ

事前学習モデルのファインチューニングに必要な学習データ（音声認識結果と正解テキストのペア）は，以下の手順で準備した．

(1) ニュース番組サイトに対して，データ 5 日分のページをウェブクローリングして，記事のタイトルとコンテンツを抽出したテキストを収集する
(2) ウェブクローリングして収集したテキストを，文分割し，文テキストを作成する（以下，オリジナル文テキスト）
(3) オリジナル文テキストを音声合成により，音声データを作成する．なお，音声合成エンジンは，ESPNet-TTS[13] を利用した．
(4) 作成した音声データに対して，音声認識処理を行う
(5) 音声認識結果とオリジナル文テキストに対して，漢数字の統一化，全角化，句読点削除等のノーマライズを行う．
(6) 以下に該当するペアを，学習データから除外する
- オリジナル文テキストが，１０文字以下，または，数値，記号のみ
- 音声認識結果の文字誤り率が，0.4 以上

上記により，約 15,000 のペア文（音声認識結果とオリジナル文テキストのペア）を，ファインチューニング用の学習データとして準備した．

4.1.3 評価

評価データは，ニュース番組の動画ファイル計 14 個の音声データを用いた．なお，動画ファイルは，同じニュース番組であり，各ファイルで音声長が異なる．評価データに対する，エラー訂正モデルの文字誤り率を**表 1** に，訂正結果のサンプルを**表 2** に示す．表 1 中の文字誤り率は，それぞれ，(1) ASR が音声認識の結果，(2) BART がファインチューニング後の BART モデルで訂正した結果，(3) BART Large がファインチューニング後の BART Large モデルで訂正した結果の，文字誤り率である．また，誤り削減率は，ベースラインの ASR と比較した相対エラー削減率である．

ベースラインの音声認識結果と比較して，BART モデルを用いた場合に，エラー削減率は平均で 2.28%，BART Large モデルを用いた場合に，平均で 7.84%であった．ただし，BART モデルに関しては，14 番組中 5 番組，BART Large モデルに関しては，14 番組中 1 番組で，音声認識よ

[*1] http://lotus.kuee.kyoto-u.ac.jp/nl-resource/JapaneseBARTPretrainedModel/{japanese_bart_base_2.0.tar.gz, japanese_bart_large_2.0.tar.gz}

表 2 訂正結果のサンプル

1	ASR	新型**このウイルス**の新規感染者が
	BART	新型**コロナウイルス**の新規感染者が
2	ASR	大規模接種では**モデルの性能ワクチン**が使用され
	BART	大規模接種では**モデルナ製のワクチン**が使用され
3	ASR	スリランカ国籍の**３島さんだまり**さんの
	BART	スリランカ国籍の**ウィシュマサンダマリ**さんの
4	ASR	防衛省のホームページや**ライン**を経由した
	BART	防衛省のホームページや**ＬＩＮＥ**を経由した
5	ASR	来月末に制限の**全面会場**を目指しています
	BART	来月末に制限の**全面解除**を目指しています
6	ASR	イギリスで**え**パブの屋内営業などが
	BART	イギリスでパブの屋内営業などが
7	ASR	インド**編い**株の感染が拡大していることから
	BART	インド**由来**株の感染が拡大していることから
	正解文	インド変異株の感染が拡大していることから
8	ASR	説明を受け**この中**での活動を気遣い，大変な状況ですが
	BART	説明を受け**コロナ禍**での活動を気遣い**ました**，大変な状況ですが
	正解文	説明を受けコロナ禍での活動を気遣い，大変な状況ですが

り性能劣化するところがあった．

表 2 のサンプルの通り，新語や固有名詞などが正しく訂正できている（1〜5）他，不要語の削除（6）もできていることがわかる．一方で，意味が近い単語に訂正されたり（7），不要な挿入（8）も見られた．

4.2 オンライン適応の評価

次に，オンライン適応手法の評価を行った．具体的には，番組毎に，各発話に対して，発話時間順に，以下を実施し，訂正結果後の文字誤り率を算出した．

(1) オンライン適応フェーズに記載の方法で，音声認識結果を訂正
(2) オンライン学習フェーズに記載の方法で，訂正変換器を更新

訂正モデル，学習データ，及び，評価データに利用したデータは，前節 4.1 と同じである．BART Large モデルを用いた場合の各音声データの文字誤り率の結果を，**表 3** に示す．表 3 には，各番組の発話毎に対する，音声認識結果の文字誤り率の平均（ASR），訂正変換器で訂正した後の文字誤り率の平均（COR1），音声データの全ての発話に対して，訂正変換器の更新を実施した後に，同じ音声データの各発話に対して，訂正変換器で訂正した後の文字誤り率の平均（COR2），を示している．

発話順に適応した場合（COR1）には，平均で 0.16%の削減となり効果が薄かったが，同じ音声データの場合（COR2）には，平均で 6.24%の削減となり，有効性が確認できた．これは，ニュース番組の１番組内においては，あまり効果がないが，番組が別だが，類似のニュースを分けて放送するニュース番組など（例えば，朝，昼，夜で共通のニュース内容を放送する場合）では，本手法が有効である可能性があることがわかった．

表 3 オンライン適応の文字誤り率

ID	NAME	ASR	COR1	COR2
1	朝のニュース 5 月 14 日	14.02	13.85	8.52
2	昼のニュース 5 月 14 日	11.33	11.11	4.86
3	夜のニュース 5 月 14 日	23.97	24.08	12.43
4	昼のニュース 5 月 16 日	10.30	10.37	3.91
5	昼のニュース 5 月 17 日	12.25	12.15	13.45
6	夜のニュース 5 月 17 日	16.22	15.89	6.97
7	朝のニュース 5 月 18 日	10.30	9.87	5.04
8	昼のニュース 5 月 18 日	13.25	12.82	6.19
9	夜のニュース 5 月 18 日	11.13	11.16	5.68
10	朝のニュース 5 月 19 日	12.34	12.29	3.32
11	昼のニュース 5 月 19 日	11.96	12.00	5.29
12	昼のニュース 5 月 20 日	11.98	11.95	7.63
13	朝のニュース 5 月 21 日	28.88	28.35	21.68
14	朝のニュース 5 月 24 日	6.22	6.00	1.90
	平均	13.87	13.71	7.63

5. おわりに

本論文では，音声認識結果のエラーを訂正した場合に，即時に修正を学習し，次の発話以降では，音声認識が同じような誤りをした場合には自動で訂正する，オンライン適応手法を検討した．特に，近年，自然言語処理の文訂正などのタスクで良好な結果を出している事前学習モデル BART を，オンライン適応へ応用した．実験の結果，14 のニュース番組での字幕制作において，訂正モデルを用いないベースラインと比較した場合に，通常サイズの訂正モデルではエラー削減率が平均 2.28%，サイズが大きい BART Large モデルを用いた場合では，エラー削減率が平均 7.84%となり，訂正モデルの効果が確認できた．また，オンライン適応においても，類似の内容を分けて放送するニュース番組等では，本手法の有効性が確認できた．

今後は，蒸留を用いたモデルサイズの削減や，継続学習を用いた訂正モデルの更新等に取り組み，オンライン学習における即時適応の高速化と精度向上にむけて検討していく．

参考文献

[1] 小森智康：生放送番組における自動字幕制作の最新動向，NHK 技研 R&D2020 年夏号，Vol. 182, No. 3 (2020).
[2] Errattahi, R., El Hannani, A. and Ouahmane, H.: Automatic Speech Recognition Errors Detection and Correction: A Review, *Procedia Computer Science*, Vol. 128, pp. 32–37 (online), DOI: https://doi.org/10.1016/j.procs.2018.03.005 (2018).
[3] Chen, W., Ananthakrishnan, S., Kumar, R., Prasad, R. and Natarajan, P.: ASR error detection in a conversational spoken language translation system, *2013 IEEE International Conference on Acoustics, Speech*

and Signal Processing, pp. 7418–7422 (online), DOI: 10.1109/ICASSP.2013.6639104 (2013).

[4] Hwang, M.-Y., Yu, D., Acero, A. and Deng, L.: Unsupervised Learning from Users' Error Correction in Speech Dictation, *Proc. Int. Conf. on Spoken Language Processing*, International Speech Communication Association (2004).

[5] Bassil, Y. and Semaan, P.: ASR Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset, (online), DOI: 10.48550/ARXIV.1203.5262 (2012).

[6] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 7871–7880 (online), DOI: 10.18653/v1/2020.acl-main.703 (2020).

[7] 田中　佑，村脇有吾，河原大輔，黒橋禎夫：日本語 Wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの構築，自然言語処理，Vol. 28, No. 4, pp. 995–1033（オンライン），DOI: 10.5715/jnlp.28.995 (2021).

[8] 小椋忠志, 藤本雅清, 沈　鵬, Xugang, L., 河井　恒：BERT を用いた音声翻訳のための音声認識結果訂正の検討，研究報告音声言語情報処理，No. 20, pp. 1–4 (2022).

[9] Yang, J., Li, R. and Peng, W.: ASR Error Correction with Constrained Decoding on Operation Prediction, (online), DOI: 10.48550/ARXIV.2208.04641 (2022).

[10] Zhao, Y., Yang, X., Wang, J., Gao, Y., Yan, C. and Zhou, Y.: BART Based Semantic Correction for Mandarin Automatic Speech Recognition System, *Proc. Interspeech 2021*, pp. 2017–2021 (online), DOI: 10.21437/Interspeech.2021-739 (2021).

[11] Wang, D., Wei, H., Zhang, Z., Huang, S., Xie, J. and Chen, J.: Non-parametric Online Learning from Human Feedback for Neural Machine Translation, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 10, pp. 11431–11439 (online), DOI: 10.1609/aaai.v36i10.21395 (2022).

[12] Fujimura, H., Nagao, M. and Masuko, T.: Simultaneous Speech Recognition and Acoustic Event Detection Using an LSTM-CTC Acoustic Model and a WFST Decoder, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5834–5838 (online), DOI: 10.1109/ICASSP.2018.8461916 (2018).

[13] Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., Takeda, K., Zhang, Y. and Tan, X.: Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7654–7658 (online), DOI: 10.1109/ICASSP40776.2020.9053512 (2020).