前後の発話を文脈として考慮するニューラル音声認識誤り訂正

中村 朝陽[1,a)]　李 聖民[1,b)]　田村 鴻希[1,c)]　吉永 直樹[2,d)]

概要：Youtube などの動画共有プラットフォームの発達や，コロナ禍におけるオンライン授業や会議の増加に伴い，音声言語情報を含む膨大なマルチメディアデータが集積されるようになっている．このようなマルチメディアデータへのアクセスを容易にするためには，音声認識で字幕を付与することが有用であるが，既存の音声認識技術は認識のリアルタイム性に焦点を当てて局所的な発話文脈のみを考慮して認識を行っており，集積された音声言語データを書き起こす上で妥当な設定で研究されていない．そこで本研究では，前後の発話を考慮して音声認識の誤り訂正を行うタスクを設定し，テキスト生成技術を用いてこれを解く手法を提案する．具体的には，事前学習済みモデル T5 を用いて，前後発話の音声認識結果を追加で入力して発話の音声認識誤り訂正を行う．実験では，汎用のオープンソース音声認識モデル NVIDIA STT Conformer-CTC Large による音声認識結果に提案手法を適用し，前発話，後発話，またそれらの組み合わせについて，発話数を変化させたときの効果について検証する．

キーワード：音声認識誤り訂正，発話間情報，事前学習済みモデル，T5，Transformer

1. はじめに

Youtube などの動画共有プラットフォームの発達や，コロナ禍におけるオンラインコミュニケーションの増加に伴い，価値のある音声言語情報を含む膨大なマルチメディアデータが集積されるようになっている．このような音声言語データには，必ずしも人手で十分な注釈が与えられておらず，そのままでは十分に利活用することが難しい．

この問題に対し，音声認識技術を用いて音声言語データに含まれる音声を字幕に書き起こすがことできれば，聴覚障害者の視聴の補助のみならず自然言語を用いたデータの検索が可能になるなど，データへのアクセシビリティを改善することに繋がる．しかしながら，既存の音声認識モデルは，認識の即時性を強く指向した問題設定で解かれており，蓄積された音声言語データの書き起こしを行う上では必ずしも最適な設定で研究されていない．具体的には，多くの研究は発話単位で音声認識を行っており，認識対象の発話周辺の文脈を考慮して音声認識を行う試みは限定的である．また，蓄積される音声言語データの内容は多岐に渡るが，学習データが存在するドメインは少なく，学習データ外ドメインにおける音声認識性能には課題がある．

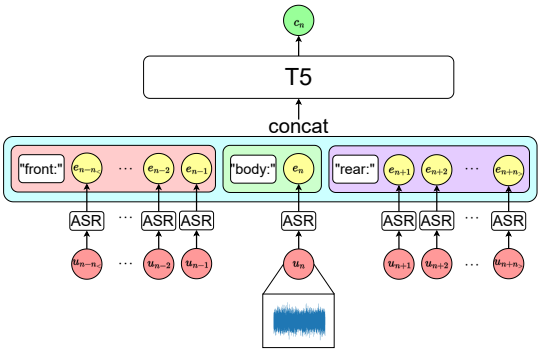

図 1　前後の発話文脈を考慮した音声認識誤り訂正．

そこで本研究では，汎用の音声認識モデルを用いて多様なドメインの音声言語データを書き起こすことを目的とし，前後の発話を文脈として考慮したニューラル音声認識誤り訂正手法 (図 1) を提案する．多様な語彙に対応しながら周囲の発話を文脈として捉えるために，我々は Transformer [1] に基づく学習済みテキスト生成モデル T5 [2] を基盤とし，修正対象の発話の音声認識結果に前後の発話の音声認識結果を連結して音声認識結果の修正を行うタスクでこれを微調整する手法を提案する．音声認識モデルと別に誤り訂正モデルを学習することで，単に学習コストが下がるだけでなく，出力を修正する自由度が上がり，後方文脈を含めた

[1] 東京大学．共同第一著者
[2] 東京大学生産技術研究所
[a)] nakamu-t@tkl.iis.u-tokyo.ac.jp, 現在は企業に所属．
[b)] lee-s@tkl.iis.u-tokyo.ac.jp
[c)] tamura-k@tkl.iis.u-tokyo.ac.jp
[d)] ynaga@iis.u-tokyo.ac.jp

より長い文脈を考慮することも可能となる．

提案手法の有効性を確認するため，CORAAL 英語音声対話コーパス [3] を用いて，汎用音声認識モデルである NVIDIA STT Conformer-CTC Large [4] の出力を修正する実験を実施した．実験では，修正モデルが参照する前後の発話を，参照する数を変えながら入力発話に追加して，広範囲の文脈を考慮する有効性を検証した．その結果，過去発話，未来発話のいずれも音声認識誤りを訂正するのに有効であり，かつ，両方の文脈を組み合わせた場合も相対的に WER の改善がみられることを確認した．

本研究の研究の貢献は以下の通りである．

- 過去と未来の両方の発話を文脈として用いる音声認識誤り訂正モデルを提案した．
- 音声認識誤り訂正のために T5 を微調整することが有効であることを確認した．
- 過去の発話に加えて未来の発話を参照することが音声認識誤り訂正において有効であることを示した．
- 特に固有名詞の音声認識誤りの訂正に提案手法が有効であることを確認した．

2. 関連研究

本節では，まず音声認識の枠内で言語モデルを用いて発話文脈を考慮する手法 (§ 2.1) について説明し，次に音声認識結果を修正する音声認識誤り訂正手法 (§ 2.2) について述べる．最後に，本研究でも用いる文脈を考慮したテキスト生成手法について説明する (§ 2.3)．

2.1 言語モデルを用いた音声認識

音声認識タスクは，発話単位で音声の書き起こしを行うタスクとしてデータセットが整備され，研究が行われている．処理のリアルタイム性が重視されていることや，音素や文字に対応する音響特徴量の系列を入力として扱うことから，入力（音声）のレベルで複数の発話を入力することは実際的に難しい．

そのため，出力である言語の流暢さを捉えるのに用いられる言語モデルを介して，直前の発話 [5]，また直前・直後の発話 [6] の音声認識結果を考慮して音声認識を行うことが試みられている．また，音声認識モデルの出力のリランキングに過去の発話の音声認識結果を考慮した言語モデルを用いる手法 [7] や，発話時のドメイン（ユーザ情報，使用位置）を考慮した n-gram を文脈として用いる手法 [8] が提案されている．

これらの音声認識モデルでは，音声認識結果を補助的に調整するために言語モデルが使われることから，言語モデルの強い生成能力を十分に発揮できないという問題がある．また，言語モデルを含め，音声認識モデル全体を訓練する必要があることも多く，多様なドメインに特化したモデルを得るのはコストが大きいという問題もある．

2.2 音声認識誤り訂正

音声認識誤り訂正は，音声認識モデルによって書き起こされた発話を入力とし，音声認識の誤りを訂正した発話を出力とするテキスト生成タスクとして定式化される [9], [10]．古典的には，単語周辺の共起語を参照してスペル訂正を行う手法 [11], [12] が提案されているが，近年では，機械翻訳や自動要約などのテキスト生成タスクにおける深層学習ベースのモデルの成功を受けて，Long-Short Term Memory や Transformer に基づくニューラルテキスト生成モデルを用いたモデルが広く用いられるようになっている [13], [14], [15], [16], [17], [18]．

本研究と同様に周囲の発話を文脈として考慮する誤り訂正モデルとしては，過去の発話から以降の発話に出現しやすい人物名のリストを抽出して，発話中の疑わしい語列をこのリストの中の人物名に置き換える手法 [19] が研究されているが，誤り訂正において過去や未来の発話を直接参照する研究は行われていない．また，本研究と同様に，事前学習済み生成モデル BART [20] を用いて音声認識の誤り訂正を行う手法 [17] が提案されているが，利用する情報は発話内に留まっており，長い文脈を用いて学習された事前学習済みモデルの真価は発揮できていない．

2.3 文脈を考慮したテキスト生成

翻訳や文法誤り訂正などのテキスト生成タスクにおいて，入力の（主に前）文脈を考慮して，出力テキストを生成する手法が研究されている [21], [22]．これらの手法では，前文脈と入力を連結して通常の encoder-decoder 型の生成モデルに入力して出力を得る所謂 2-to-1 モデル [21], [23], [24] と，前文脈と入力を別々のエンコーダで処理するマルチエンコーダ型のモデル [23], [25], [26], [27], [28], [29], [30], [31] が主に利用されている．

文脈を考慮したモデルで最も問題となるのは，学習となる文脈を含む学習データの量である．機械翻訳は文を単位としてデータセット（対訳データ）が整備されており，対訳文はアラインメントによって対訳文書から収集したものが多く，多くの場合，文脈が失われている．この点を考慮して，近年では，文脈を含む出力言語文書から事前に学習した文脈考慮型言語モデルを用いて文脈翻訳を行う手法 [32], [33] が提案されている．

3. 提案手法

本節では，前後の発話の書き起こしを考慮しながら音声認識モデルの出力を修正するモデルを提案する．提案手法は，入力を原言語，出力を目的言語とみなして，入力の前後の文脈を入力に追加して出力に写像するニューラル文脈翻訳の枠組みで，音声認識結果の訂正を行う．具体的に，モデルは誤り訂正対象の書き起こし発話 e_n に加えて直前の書き起こし文 $e_{n-n_<:n-1}$ と直後の書き起こし文 $e_{n+1:n+n_>}$

を入力し，e_n に対応する文の訂正文 c_n を出力とする．

音声認識誤り訂正モデルの学習データは，音声認識の学習データに訂正対象の音声認識モデルして書き起こし発話を得ることで構築することができるが，元々の音声認識データセットの規模が小さいため，生成モデルを学習する上で十分なサイズが得られることが少ない．目的ドメインのテキストに対して音声合成により付与した音声データを用いたデータ拡張により，大規模な擬似学習データを構築することもできるが，別途音声合成モデルを用意し，必要によっては目的ドメインで再学習する必要がある．さらに，擬似学習データは一般的に人手で用意された学習データよりも質が劣るため，高品質のモデルを学習するためには大規模な擬似学習データを用意する必要があり，学習コストという点で大きな課題が残る．

そこで，本研究では事前学習された生成モデルである T5 [2] を基盤モデルとして誤り訂正モデルを学習することで，評価タスクの学習データの規模の問題を回避する．大規模データからサブワードを語彙に用いて事前学習したモデルを用いることで，低頻語や固有名詞などの修正精度が改善すると期待できる．

T5 は，さまざまなタスクをテキスト同士の変換として定義している点が特徴であり，各タスクは各種の入力するテキストに付与される接頭辞によって表される．本研究では，訂正対象の書き起こし文に “body:” の接頭辞を，$n_<$ 文の過去の発話の書き起こし文を改行記号で分割した 1 つの文字列にしたうえで “front:” の接頭辞を，$n_>$ 文の未来の発話の書き起こし文を同じく改行記号で分割した 1 つの文字列にしたうえで “rear:” の接頭辞をつけ，それぞれの接頭辞とそれに結び付けられる文字列は改行記号を挟んだ 1 つの文字列にまとめて入力したものを，それぞれの発話の入力として扱う．また，出力文は訂正対象文を訂正した文字列である．T5 を誤り訂正モデルとして活用するため，我々は以上の形式に沿った入出力の文によって事前学習済みモデルを微調整した．

4. 実験設定

本節では，事前学習済み生成モデル T5 [2] を誤り訂正器として微調整することの効果，および周囲の発話を文脈として考慮することの効果を確かめるため，考慮する周囲の発話の数を変えながら CORAAL (Corpus of Regional African American Language) 音声認識データセット [34] の入力音声を NVIDIA STT Conformer-CTC Large で認識した書き起こし文に対して誤り訂正を行い，その精度と推論時間について報告する．

4.1 データセット

提案手法の学習・評価に用いる音声認識誤り訂正データセットについては，社会言語学的インタビュー 231 件を収録した CORAAL [34][*1] を元に構築した．原文の書き起こしに従い，各インタビューを発話単位に分割した．聞き取れないか不明瞭な音声，また非言語的な音声は削除した．前後の文脈の把握のため，複数の発話が重なっている場合は削除しなかった．各インタビューは 9:0.5:0.5 の割合で学習・検証・テストデータに分割した．データセットの詳細は表 1 に示す．

	train	validation	test
インタビュー数	207	12	12
発話数	201,032	10,846	12,455
平均発話数/インタビュー	971.17	903.83	1037.08
平均単語数/発話	6.20	6.09	5.91
時間	122.37	5.90	7.33
ASR による書き起こし			
平均単語数/発話	5.83	5.76	5.60
WER	27.70	27.30	26.50
CER	17.32	17.35	16.70

表 1 CORAAL を元に作った音声認識誤り訂正データセット．

4.2 音声認識モデル

ベースラインとなる音声認識器には，Nemo Toolkit に収録されている，オープンソースで公開されている音声認識モデルである NVIDIA STT Conformer-CTC Large[*2]を採用した．このモデルは Conformer [35] のエンコーダでの損失を CTC [36] のものに，デコーダを単層 LSTM から線形層に変更して非自己回帰モデルにした音声認識モデルであり，Nemo Toolkit 付属の英語音声複合データセットである NeMo ASRSET の音声を用いて学習されている．なお，このデータセットには CORAAL の音声データは含まれていない．このモデルを用いて CORAAL データセットの各発話の音声を認識し，音声認識誤り訂正タスクのデータセットを生成した．音声認識誤り訂正の入力となる，得られたデータセットの Word Error Rate (WER), Character Error Rate (CER) を表 1 に示す．

4.3 音声認識誤り訂正モデルの訓練

事前学習済み T5-base モデル[*3]を元に，前述の音声認識誤り訂正データセットを用いて，以下で述べるパラメータ設定で誤り訂正モデルを学習した．基本的には T5-base の設定に沿い，埋め込み（d_{model}）は 768 次元，エンコーダ，デコーダはそれぞれ 12 層，各層の出力（d_{ff}）の次元は 3072 次元とし，各注意機構は 12 チャンネルの Multi-head Attention，ドロップアウト率は 0.1 とした． また，変更点

[*1] https://oraal.uoregon.edu/coraal
[*2] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_large
[*3] https://huggingface.co/t5-base

手法	$n_<$	$n_>$	WER	CER	WER (固有名詞)
(input)	n/a	n/a	26.50	16.70	74.71
T5	0	0	24.88	17.06	65.17
T5	5	0	24.46	16.92	63.09
T5	10	0	24.32	16.88	60.98
T5	15	0	24.20	16.78	62.10
T5	0	5	24.37	16.85	62.44
T5	0	10	24.36	16.80	61.62
T5	0	15	24.39	16.87	61.47
T5	5	5	23.90	16.73	60.49
T5	10	15	23.82	16.78	60.15
T5	15	15	**23.71**	**16.67**	**59.15**

表 2 音声認識誤り訂正結果.

としては, バッチサイズを 32, 学習率を 1e-4 とし, Adafactor optimizer [37] を用いて学習した．詳細なパラメータとしては，$\epsilon_1 = 10^{-30}, \epsilon_2 = 10^{-3}$, clipping threshold $= 1.0$, decay rate $= -0.8$, weight decay $= 0.0$ のように設定し，relative step と scale parameter，warmup init はそれぞれ無効化した.

4.4 評価設定

音声認識誤り訂正における T5 の有効性，事前学習の有効性，固有名詞の音声認識誤り訂正における T5 の有効性，前後発話の数が推論時間に与える影響，という側面において評価を行った．$n_<$ と $n_>$ の値をそれぞれ 0 から 15 まで変化させ，前後発話の数の影響を実験した．$n_<$ と $n_>$ が共に 0 の場合は，T5 を単なる seq2seq 言語モデルとして使用し，各入力発話の誤りを訂正する．なお，各評価時に参照する周囲の発話の数と同じ数の周囲の発話を用いて複数の音声認識誤り訂正モデルを訓練している．評価指標として，単語誤り率 (WER)，文字誤り率 (CER) を用いて，モデルの精度を評価した．各実験は 2 回ずつ実行し，その平均と標準偏差を報告する.

5. 実験結果

5.1 前後発話考慮の効果

入力発話に対して，過去の発話のみ・未来の発話のみ・両方の発話を同じ数だけ与えた場合の結果を，表 2 に示す．この表の $(n_<, n_>) = (0, 0)$ は，修正対象の発話のみを入力とする音声認識誤り訂正モデルであり，WER において元の ASR の結果を上回っている．これは，T5 を用いて音声認識誤り訂正が可能であることを示している．また，$n_< + n_> > 0$ のいずれも $(n_<, n_>) = (0, 0)$ の結果を上回っており，特に過去の発話については $n_<$ が大きいほど良好な結果が得られている．さらに，過去と未来発話の両方を使用することで，片方だけを使用するより精度を向上させることができた.

ASR transcription	T5 output
toug of (tougher) people	**a lot of** people
matter of fact she was *mis-teened* (ms teen)	matter of fact she was **miss**
start your *mode* (motor) and you drive	start your **manifest** and you drive

表 3 過剰修正の例．音声認識結果・正しい書き起こし・T5 による過剰修正の結果を，それぞれ斜体・格好内・太字で表した.

T5 input

`front:` even though i didn't do that with my tenth grade yet but i didn't do that with my tenth grade ...

`body:` i show that i can do it my *ninightware* yeaha i end this school with a three point two

T5 output

i showed that i can do it my **ninth grade** yes i ended this school with a three point two

T5 input

`body:` i had a nineteen thirty five *fod* at the time

`rear:` brand new car that was one of the first that was the first v eight but it was one of the first veight engines that the ford put out

T5 output

i had a nineteen thirty five **ford** at the time

T5 input

`front:` you can't remember when you pulled the trick on the teacher oh in the classroom ... we set the tash can on fire the teacher come in she took off her coat her coat goll burned up ...

`body:` i got *married* with this teacher

`rear:` and i asked one of the janitories for some water with some ware in the bucket ... it was a string on where the bucket was attached ... i asked him to pull the string down because it wouldn't come down and the water fell all over

T5 output

i got **mad** with this teacher

表 4 提案手法による訂正例

一方で，CER はほとんど改善されないか，わずかに悪化する傾向にあった．これは，T5 が発音を考慮せずに修正を行うため，過剰な修正が行われる場合があるからだと推測される．過剰な修正によって ASR の結果が捉える音響的情報が失われ，CER が悪化する原因になると思われる．その過剰修正例を表 3 に示す.

最後に，モデルの修正例を表 4 に示す．この 3 つの例は，それぞれ過去，未来，両方の発話を用いた修正の結果を示している．最初の例では，ASR の結果が悪いにもかかわ

手法	$n_<$	$n_>$	WER	CER
(input)	n/a	n/a	26.50	16.70
T5 (random)	0	0	27.74	18.91
T5 (random)	5	5	27.31	19.04
T5 (random)	10	10	26.91	18.69
T5 (random)	15	15	27.31	18.87

表 5 ランダム初期化した T5 を用いた音声認識誤り訂正結果.

らず， *ninightware* を **ninth grade** に正しく修正している．2 番目の例では，未来発話からの情報を活用して *fod* を自動車メーカーの **ford** に修正していることがわかる．最後の例では，前後発話から，不自然な単語 *married* を検出し，**mad** に修正している．

5.2 事前学習の効果

T5 の事前学習の効果を確認するため，T5 モデルのパラメータをランダムに初期化したモデルである T5 (random) を学習した．その結果を表 2 に示す．T5 (random) は入力文と比べて WER, CER 共に悪化しており，誤り訂正に失敗している．これは，入力データである音声認識結果のノイズの多さに起因すると考えられる．また，前後発話数を増やしても誤り訂正に失敗しており，これは事前学習の有効性を示唆する結果である．

5.3 固有表現の誤り訂正

大規模データにより事前学習された T5 により，固有名詞の誤り訂正性能の改善が期待されるため，固有名詞に焦点を当てた評価を行った．学習済み BERT [38][*4]により固有名詞を検出し，これらの表現を T5 が正しく訂正しているかの評価を行った．正解文と予測文のそれぞれから検出された固有名詞を 1 つの文字列とし，得られた 2 つの文字列から WER を計算した．その結果を表 2 に示す．5.1 節の結果と同様に，前後発話を使用することで固有名詞の訂正精度が改善しており，その改善幅は WER で 15.564 (74.71 → 59.15) と非常に大きいことが分かる．

5.4 前後発話数の推論時間への影響

表 6 は，バッチサイズが 32 の場合の平均推論時間を示す．表から，考慮する前後の発話数が増えるにつれ，推論時間は長くなることが分かるが，同時にその増加はある程度抑えられていることが分かる．この理由としては，T5 のアーキテクチャである Transformer が自己回帰型のモデルとなっており，エンコーダ側の処理は並列化できるがデコーダ側の処理は並列化しづらく，デコーダ側の処理が処理時間の大半を占めるためだと考えられる．

[*4] https://huggingface.co/vblagoje/bert-english-uncased-finetuned-pos

手法	$n_<$	$n_>$	バッチあたりの推論時間 (秒)
T5	0	0	276.502 (±89.461)
T5	5	0	301.130 (±104.892)
T5	10	0	323.655 (±103.556)
T5	15	0	348.090 (±115.295)
T5	0	5	303.635 (±103.558)
T5	0	10	324.930 (±107.161)
T5	0	15	334.583 (±101.220)
T5	5	5	323.457 (±106.181)
T5	10	10	363.729 (±111.727)
T5	15	15	399.280 (±116.274)

表 6 発話数による推論時間の比較 (括弧内は標準偏差).

6. おわりに

本研究では，事前学習済みテキスト生成モデルである T5 を用いて，前後の発話の音声認識結果を考慮した音声認識誤り訂正手法を提案した．実験では，社会言語学的インタビューを記録した音声認識用データセット CORAAL の入力を汎用の学習済み音声認識モデル NVIDIA STT Conformer-CTC Large で音声認識することで音声認識誤り訂正用のデータセットを構築し，提案手法の有効性を評価した．その結果，WER が最大で 2.79 (26.50 → 23.71) 改善することを確認した．今後の研究では，発音を考慮した損失を加えて制約付きの生成を行い，過剰修正を防止して CER を改善させることが求められる．

謝辞 この研究は国立情報学研究所 (NII) CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成を受けています．

参考文献

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Vol. 30, Curran Associates, Inc. (2017).

[2] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. et al.: Exploring the limits of transfer learning with a unified text-to-text transformer., *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67 (2020).

[3] Kendall, T. and Farrington, C.: The Corpus of Regional African American Language. Version 2021.07. Eugene, OR: The Online Resources for African American Language Project (2021).

[4] Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J. et al.: Nemo: a toolkit for building ai applications using neural modules, *arXiv preprint arXiv:1909.09577* (2019).

[5] Masumura, R., Ihori, M., Tanaka, T., Saito, I., Nishida, K. and Oba, T.: Generalized Large-Context

Language Models Based on Forward-Backward Hierarchical Recurrent Encoder-Decoder Models, *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, (online), DOI: 10.1109/ASRU46091.2019.9003857 (2019).

[6] Sun, G., Zhang, C. and Woodland, P. C.: Cross-Utterance Language Models with Acoustic Error Sampling (2020).

[7] Xiong, W., Wu, L., Zhang, J. and Stolcke, A.: Session-level Language Modeling for Conversational Speech, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 2764–2768 (online), DOI: 10.18653/v1/D18-1296 (2018).

[8] Williams, I., Kannan, A., Aleksic, P., Rybach, D. and Sainath, T.: Contextual Speech Recognition in End-to-end Neural Network Systems Using Beam Search, *Interspeech 2018*, ISCA, (online), DOI: 10.21437/Interspeech.2018-2416 (2018).

[9] Tanaka, T., Masumura, R., Masataki, H. and Aono, Y.: Neural Error Corrective Language Models for Automatic Speech Recognition, *Interspeech 2018*, ISCA, (online), DOI: 10.21437/Interspeech.2018-1430 (2018).

[10] Guo, J., Sainath, T. N. and Weiss, R. J.: A Spelling Correction Model for End-to-end Speech Recognition, *ICASSP 2019*, Brighton, United Kingdom, IEEE, (online), DOI: 10.1109/ICASSP.2019.8683745 (2019).

[11] Sarma, A. and Palmer, D. D.: Context-based Speech Recognition Error Detection and Correction, *Proceedings of HLT-NAACL 2004: Short Papers*, Boston, Massachusetts, USA, Association for Computational Linguistics, pp. 85–88 (online), available from ⟨https://aclanthology.org/N04-4022⟩ (2004).

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

[13] Hrinchuk, O., Popova, M. and Ginsburg, B.: Correction of Automatic Speech Recognition with Transformer Sequence-To-Sequence Model, *ICASSP 2020*, Barcelona, Spain, (online), DOI: 10.1109/ICASSP40776.2020.9053051 (2020).

[14] Wang, H., Dong, S., Liu, Y., Logan, J., Agrawal, A. K. and Liu, Y.: ASR Error Correction with Augmented Transformer for Entity Retrieval, *Interspeech 2020*, ISCA, pp. 1550–1554 (online), DOI: 10.21437/Interspeech.2020-1753 (2020).

[15] Mani, A., Palaskar, S., Meripo, N. V., Konam, S. and Metze, F.: ASR Error Correction and Domain Adaptation Using Machine Translation, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, (online), DOI: 10.1109/ICASSP40776.2020.9053126 (2020).

[16] Weng, Y., Miryala, S. S., Khatri, C., Wang, R., Zheng, H., Molino, P., Namazifar, M., Papangelis, A., Williams, H., Bell, F. and Tur, G.: Joint Contextual Modeling for ASR Correction and Language Understanding, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, IEEE, pp. 6349–6353 (online), DOI: 10.1109/ICASSP40776.2020.9053213 (2020).

[17] Zhao, Y., Yang, X., Wang, J., Gao, Y., Yan, C. and Zhou, Y.: BART based semantic correction for Mandarin automatic speech recognition system, *Interspeech 2021*, (online), DOI: 10.21437/Interspeech.2021-739 (2021).

[18] Zhang, F., Tu, M., Liu, S. and Yan, J.: ASR Error Correction with Dual-Channel Self-Supervised Learning, *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (online), DOI: 10.1109/ICASSP43922.2022.9746763 (2022).

[19] Wang, X., Liu, Y., Zhao, S. and Li, J.: A Light-Weight Contextual Spelling Correction Model for Customizing Transducer-Based Speech Recognition Systems, *Proc. Interspeech 2021*, pp. 1982–1986 (online), DOI: 10.21437/Interspeech.2021-379 (2021).

[20] 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*, (online), DOI: 10.18653/v1/2020.acl-main.703 (2020).

[21] Tiedemann, J. and Scherrer, Y.: Neural Machine Translation with Extended Context, *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 82–92 (online), DOI: 10.18653/v1/W17-4811 (2017).

[22] Chollampatt, S., Wang, W. and Ng, H. T.: Cross-Sentence Grammatical Error Correction, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 435–445 (online), DOI: 10.18653/v1/P19-1042 (2019).

[23] Bawden, R., Sennrich, R., Birch, A. and Haddow, B.: Evaluating Discourse Phenomena in Neural Machine Translation, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 1304–1313 (online), DOI: 10.18653/v1/N18-1118 (2018).

[24] Sugiyama, A. and Yoshinaga, N.: Data augmentation using back-translation for context-aware neural machine translation, *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, Hong Kong, China, Association for Computational Linguistics, pp. 35–44 (online), DOI: 10.18653/v1/D19-6504 (2019).

[25] Wang, L., Tu, Z., Way, A. and Liu, Q.: Exploiting Cross-Sentence Context for Neural Machine Translation, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2826–2831 (online), DOI: 10.18653/v1/D17-1301 (2017).

[26] Tu, Z., Liu, Y., Shi, S. and Zhang, T.: Learning to remember translation history with a continuous cache, *Transactions of the Association of Computational Linguistics (TACL)*, Vol. 6, pp. 407–420 (online), DOI: 10.1162/tacl_a_00029 (2018).

[27] Jean, S., Lauly, S., Firat, O. and Cho, K.: Does Neural Machine Translation Benefit from Larger Context?, *arXiv preprint arXiv:1704.05135*, (online), available from ⟨https://arxiv.org/pdf/1704.05135⟩ (2017).

[28] Miculicich, L., Ram, D., Pappas, N. and Henderson, J.: Document-Level Neural Machine Translation with Hierarchical Attention Networks, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2947–2954 (online), DOI: 10.18653/v1/D18-1325 (2018).

[29] Voita, E., Serdyukov, P., Sennrich, R. and Titov, I.: Context-Aware Neural Machine Translation Learns

Anaphora Resolution, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1264–1274 (online), DOI: 10.18653/v1/P18-1117 (2018).

[30] Maruf, S. and Haffari, G.: Document Context Neural Machine Translation with Memory Networks, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1275–1284 (online), DOI: 10.18653/v1/P18-1118 (2018).

[31] Maruf, S., Martins, A. F. T. and Haffari, G.: Selective Attention for Context-aware Neural Machine Translation, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT))*, pp. 3092–3102 (online), DOI: 10.18653/v1/N19-1313 (2019).

[32] Yu, L., Sartran, L., Stokowiec, W., Ling, W., Kong, L., Blunsom, P. and Dyer, C.: Better Document-Level Machine Translation with Bayes' Rule, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 346–360 (online), DOI: 10.1162/tacl_a_00319 (2020).

[33] Sugiyama, A. and Yoshinaga, N.: Context-aware Decoder for Neural Machine Translation using a Target-side Document-Level Language Model, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, Association for Computational Linguistics, pp. 5781–5791 (online), DOI: 10.18653/v1/2021.naacl-main.461 (2021).

[34] Gunter, K., Vaughn, C. and Kendall, T.: Contextualizing /s/ retraction: Sibilant variation and change in Washington D.C. African American Language, *Language Variation and Change*, Vol. 33, No. 3, p. 331–357 (online), DOI: 10.1017/S095439452100020X (2021).

[35] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition, *Proc. Interspeech 2020*, pp. 5036–5040 (online), DOI: 10.21437/Interspeech.2020-3015 (2020).

[36] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA, Association for Computing Machinery, p. 369–376 (online), DOI: 10.1145/1143844.1143891 (2006).

[37] Shazeer, N. and Stern, M.: Adafactor: Adaptive Learning Rates with Sublinear Memory Cost, *Proceedings of the 35th International Conference on Machine Learning* (Dy, J. and Krause, A., eds.), Proceedings of Machine Learning Research, Vol. 80, PMLR, pp. 4596–4604 (online), available from ⟨https://proceedings.mlr.press/v80/shazeer18a.html⟩ (2018).

[38] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL*, (online), DOI: 10.18653/v1/N19-1423 (2019).