生成的 ASR 誤り訂正における Web 検索の活用と文脈処理の最適化

春日宥一郎 [1] 大野正樹 [1]
[1] 株式会社 RevComm
{yuichiro.kasuga,masaki.ono}@revcomm.co.jp

概要

大規模言語モデル（LLM）の強力な言語理解を自動音声認識の誤り訂正に用いる，生成的誤り訂正（generative error correction; GEC）が注目されている．本研究は，希少な語彙あるいは固有名詞に対する音声認識エラーを対象にした，Web 検索を用いた GEC に着目する．Web 検索を用いることで，修正候補の検討範囲に LLM 自身の知識だけでなく，最新の語彙を含めることが可能となる．最新の語彙を含むデータセットを作成して実験を行い，これらの語彙に対する Web 検索の有効性を確認した．さらに訂正の際に用いる周辺文脈の範囲が LLM モデルによって異なることが分かった．

1 はじめに

自動音声認識（automatic speech recognition; ASR）技術は発展を続けているが，希少な語彙あるいは固有名詞においては依然としてエラーが発生し，これが要約や質問応答などの下流の自然言語処理タスクに影響を与える．ASR が一般的な会話ではなく，企業や学会等の特定の組織で使用された場合には，このような未知語に出会う確率が高くなる．そのため，この場合には ASR は未知語に対処する方法を備える必要がある．

近年，大規模言語モデル（LLM）の強力な言語理解を ASR 出力の修正に用いる，生成的誤り訂正（generative error correction; GEC）が注目されている．しかし，希少な語彙あるいは固有名詞に対する知識を網羅することは難しいので，LLM を用いても依然としてこれらを正しく認識することは難しい．

本研究は，希少な語彙あるいは固有名詞に対する ASR エラーを対象にした，Web 検索を用いた GEC に着目する．誤り訂正の際に Web 検索を用いることで，修正候補の検討範囲に LLM 自身の知識だけでなく，最新の語彙を含める．

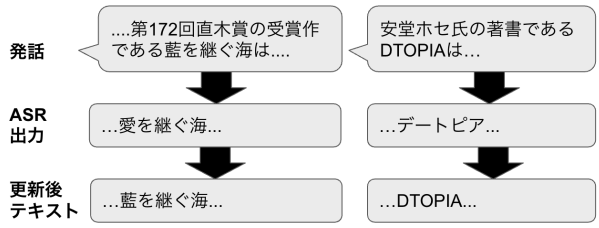

図 1: 本研究のタスクの例

図 1 に本研究のタスクの例を示す．“藍を継ぐ海” や “DTOPIA” は 2024 年後半以降に出現した語彙であり，これを含む発話を ASR で正しく認識することが難しい．例えば，“DTOPIA” は “デートピア” や “ディストピア” などと誤認識される．この様な希少な語彙あるいは固有名詞を正しく “DTOPIA” と認識することが本研究の目的である．

実験では最新の語彙を含むデータセットを作成し，それに対する ASR の結果に対して Web 検索を用いた GEC を適用した．その結果，Web 検索を用いることで LLM の知識がない語彙に関連する ASR エラーに対して対処することができた．さらに多く GEC への入力量を変化させて性能を測り，入力量が増えることで性能が向上する傾向が見られる一方で，LLM モデルによって最適な文脈範囲は 異なることがわかった．

2 関連研究

ASR の結果に対して後処理を行い，その性能を向上させる試みは以前から行われてきた．この分野の初期には，ユーザーの介入を前提とし，訂正作業を効率化する支援ツールや手法が提案された [1].

深層学習などの基盤技術の発展あるいはコーパスの増加に伴い，ユーザーの介入なしに訂正作業を行う研究が提案された．Cucu et al. [2] は統計的機械翻訳の手法に基づいて ASR エラーを訂正する手法を

提案した．Mani et al. [3] は，Transformer ベースのモデルを利用して，自己回帰的な方法で ASR 訂正モデルを学習した．Dutta et al. [4] は seq2seq モデルである BART を用いた．

近年，LLM の高い言語能力が着目され，LLM を用いた ASR の誤り訂正の手法が提案されている．ASR システムは通常，単一の最良仮説（1-best）に加え，複数の代替候補である N-best 仮説を出力することができる．LLM がこの N-best 仮説を入力として受け取り，誤り訂正を行う手法が提案されている．[5, 6, 7]．N-best 仮説を生成することに計算コストがかかり，なおかつ一般的な ASR システムではサポートされていない場合があるため．Li et al. [8] は，複数の ASR システムから 1-best 仮説を取得し，LLM への入力とした．

LLM の内部知識に加え，外部知識源を利用する試みが提案されている．この試みは，近年の LLM の研究で注目されている Tool-use のパラダイムと関連している．このパラダイムでは，LLM が自身の推論能力だけでなく，検索エンジンや API などの一般的なツールを用いることで，より複雑なタスクや最新情報を含むタスクに対応することを目指す．Rasooli et al. [9] は，固有表現の訂正に，訓練データから抽出した固有表現を検索して利用する DARAG を提案している．Yamashita et al. [10] は，先行研究 [11] で作成された語彙データベースに対して音響近傍埋め込みで音響的に類似した語彙を検索し，LLM にヒントとして与えることで，希少な語彙の認識精度を向上させた．ここで用いられている知識源はこれらの研究のために作られたクローズなものである．

GEC においてよりオープンな知識源あるいはツールを利用する研究として，Sugano et al. [12] の研究がある．ここでは形態素解析を用いてデータセットを構築し，ASR 出力の固有名詞に対して Web 検索を呼び出すように LLM を学習させた．本研究との違いは 2 点である．1 点目に本研究では LLM の学習を行なっていない．2 点目に LLM への入力である周辺文脈の範囲や形態と性能との関係を考察していることである．

3 提案手法

本研究は，希少な語彙あるいは固有名詞に対する ASR エラーを対象にした，Web 検索を用いた GEC に着目する．提案手法は，ASR 出力 A およびその中に含まれる認識誤りが疑われる箇所の集合 B が得られていることを前提とし，以下の 2 段階のステップを経て訂正を行う．はじめに，LLM m は，ASR 出力 $a_i \in A$ と認識誤り $b_{i,j} \in B_i$ の周辺文脈を元に，検索クエリ q を生成する．次に，LLM m と 検索クエリ q を Web 検索サービスを使って，ASR 出力 $a_i \in A$ を更新し，訂正済みテキスト $c_i \in C$ を生成する．

検索クエリ Q を生成する際に用いる周辺文脈の範囲には次のバリエーションを考えた．

- 局所的文脈: 未知語を含む 1 文のみを用いる
- 大域的文脈: 対話全体の書き起こしを用いる
- 超大域的文脈: 対象の対話に加え同一の未知語を含む全ての対話を統合して用いる

超大域的文脈を得る際に，複数の対話が同一の語彙を含むことを判断するために ASR 出力における誤り箇所の文字列の一致を用いる．具体的には，対象の誤り文字列（$b_{i,j}$）と同一の文字列が出現するすべての対話をデータセット全体から検索・抽出し，それらを結合することで超大域的文脈を構築する．

さらに周辺文脈として会話をそのまま用いるのではなく，会話を要約することを試みた．[13] は RAG を使った質問応答システムを構築する際に，知識源となる文書をそのまま LLM に入力するよりも，その文書の要約を用いることで性能が向上したと報告している．本研究では，ASR の誤り訂正においても，会話の要約を用いることで性能が向上するか検証する．要約を生成する際には，LLM を用いて文脈情報からその単語が本来持つ意味的・音響的特徴を推測して記述させる．例として，ASR が “DUKTIG” を “リュクティグ” と誤認識した時，周囲の文脈を入力として「# “リュクティグ” の特徴記述分析## 文脈から推測される意味的・音響的特徴### 1. 意味的特徴 **製品カテゴリー**- 子ども向けのおもちゃ・遊具の一種- “おままごとキッチン” という玩具の製品名またはシリーズ名 - IKEA 社が販売する商品ラインの名称...」という要約が生成された．

4 実験

4.1 実験準備

実験に用いるデータセットは次のステップで作成した．

1. 2024 年以降の語彙である 30 語を人手で選ぶ．
2. Web 検索を用いた LLM により語彙の説明を導出する．さらに LLM を使ってランダム選定し

た異なる 2 つの営業と顧客のペルソナをそれぞれ作成する．最後に LLM を使って 2 人の話者による 20 ターン以上の日本語の対話テキストを 10 対話生成する．
3. 音声合成を用いて対話テキストに基づいて対話音声を生成する．
4. ASR を用いて対話音声から ASR 出力を得る．
5. 人手で対象の未知語の誤り箇所をアノテーションする．

アノテーションの結果，全体で 309 箇所，1 単語あたり 10.30 箇所の誤りが得られた. 重複を除いたユニークな誤りは 204 種類，1 単語あたり 6.87 種類となった．

語彙の説明を生成する際に Gemini-2.5-flash[1] を，対話テキストの生成の際に Claude Sonnet 4.5 (Claude)[2] をそれぞれ用いた．また，音声合成システムとして Gemini-2.5-pro-tts[1] を，ASR システムとして Whisper-large-v3[3] をそれぞれ用いた．対話テキストの生成のプロセスは，Tao et al. [14] の手法を参考にしており，ペルソナを用いることで，対話テキストの多様性を増やす狙いがある．

実験対象の語彙には，“KPop Demon Hunters”，“Pokemon LEGENDS Z-A”，“Terra Xross 1”，“すっぽり収納着る毛布 (N ウォームラビットタッチ ミドル LMO)”，“香酢が効いた旨辛たれビャンビャン麺” などが含まれる．これらは特定のトピックに依存しない評価を行うため，家具，エンタメ，食品，経済用語，サービスなど多様なドメインから収集した．

LLM のモデルの能力差を検討するために，LLM m として 4 つの LLM モデルを用いた; Claude[2]，Gemini-2.5-pro (Gemini)[1]，Llama 4 Maverick (Llama4)[4], Qwen3-235B-A22B (Qwen3)[5]．Claude, Gemini はクローズなモデルであり，Llama, Qwen3 はオープンなモデルである．

評価指標として正解率を用いて，GEC の誤り訂正能力を測った．各対話における対象の語彙の出現回数が異なるため，データセット全体の性能を測る際にはマクロ平均を使用した．また，3 回試行の平均値を計測した. 対象語彙と GEC の出力を比べる際にはカタカナやアルファベットの表記ゆれや部分一致を許容した．例えば，“鬼滅の刃 無限城編 第一章 猗窩座再来” という語彙に対しては，“鬼滅の刃”“無限城編”“猗窩座再来” の部分一致も正答とした．また，“DUKTIG” のように公式ウェブサイトでカタカナが併記されている一部の他言語表記については，“ドゥクティグ” などの日本語表記も正答として定義した．

また，本実験における評価は誤り箇所が既知であるという前提に基づいており，ASR 出力からの誤り検出性能は評価の対象外である．

4.2 Web 検索の有効性検証

Web 検索の有無が性能に及ぼす影響を確認するために，表 1 に周辺文脈と LLM のモデルを変更した際の結果を示す．

外部知識の利用効果を検証するため，Search ON(検索あり) と Search OFF(検索なし・LLM の内部知識のみ) の 2 条件を設定した．

周辺文脈の範囲や LLM のモデルによらず，Web 検索を用いることで，すべての条件で Web 検索なしよりも高い性能が得られている．本研究が対象とする「2024 年以降の語彙」という，LLM の内部知識では修正困難な語彙に対して，Web 検索によって得られた最新の知識が有効に機能したことを示す．

クローズなモデルである Gemini と Claude が，オープンなモデルである Llama や Qwen3 と比較して，全体的に高い性能を示している．最も高い性能は，Gemini が超大域的文脈と Web 検索を用いた場合であり，その値は 0.797 である．さらに Web 検索を用いない場合において，Gemini や Claude が Llama よりも高いスコアを出している．これらのことから，本実験においては，クローズなモデルの方が希少語彙に対する LLM の内部知識量あるいは文脈からの推論能力が優れていた．

局所的文脈よりも大域的文脈の性能が低い場合は 1 件のみであり，多くの LLM モデルにおいては大域的文脈を使うことで性能が向上した．大域的文脈と超大域的文脈の比較では，Web 検索を使用した場合に性能が向上したケースが 2 件であり，Web 検索を使用しない場合に，性能が向上したケースが 3 件である．これらの結果から，文脈の範囲が広がることで性能が向上する傾向が見られる一方で，モデルによって最適な文脈範囲は異なると判断できる．例えば，Gemini は，Web 検索を使用した場合において，文脈が広がるほど性能が向上した．具体的には，局所的，大域的，超大域的のそれぞれに対する性能は

1) https://gemini.google.com/
2) https://www.anthropic.com/claude
3) https://github.com/openai/whisper
4) https://llama.meta.com/
5) https://github.com/QwenLM/Qwen

表 1: 周辺文脈の範囲および Web 検索の有無に応じたマクロ平均正解率のモデル間比較

Context	Search	Claude	Gemini	Llama	Qwen3
局所的	OFF	0.351	0.378	0.101	0.177
	ON	0.534	0.621	0.447	0.540
大域的	OFF	0.499	0.459	0.290	0.265
	ON	0.718	0.750	0.492	0.531
超大域的	OFF	0.282	0.476	0.339	0.320
	ON	0.362	0.797	0.380	0.612

表 2: モデルおよび文脈範囲別の誤り要因分析 (件数および割合)

Model	Context	総失敗数	検索失敗 (%)	推論失敗 (%)
Claude	大域的	65.0	20.3 (31.2%)	44.7 (68.8%)
	超大域的	136.3	31.0 (22.7%)	105.3 (77.3%)
Gemini	大域的	63.3	16.0 (25.3%)	47.3 (74.7%)
	超大域的	50.3	6.3 (12.5%)	44 (87.5%)
Llama	大域的	112.6	74.3 (66.0%)	38.3 (34.0%)
	超大域的	133.3	113.3 (85.0%)	20.0 (15.0%)
Qwen3	大域的	104.0	65.7 (63.2%)	38.3 (36.8%)
	超大域的	96.3	34.0 (35.3%)	62.3 (64.7%)

0.621，0.750，0.797 であった．しかし，Llama では，大域的文脈の値が 0.492，超大域的文脈の値が 0.380 であり，文脈範囲が広がったことにより性能が劣化した．

誤りの要因を分析するため，誤りを「検索失敗」と「推論失敗」に分類した．前者は適切なクエリ生成や知識取得の失敗，後者は知識に基づく導出の失敗と定義する．なお，検索結果内に正答の単語が含まれていない場合に検索失敗とした．また，検索失敗時は正解導出が困難となるため，検索成功例のみを対象に推論失敗を判定した．表 2 に，モデルおよび文脈範囲別の誤り要因の分類結果を示す．分析の結果，Claude は，超大域的文脈において推論失敗数が 44.7 件から 105.3 件へ大幅に増加しており，長大な文脈処理に課題が見られた．Llama は超大域的文脈において検索失敗数が 74.3 件から 113.3 件へ増加しており，文脈量の増加に伴いクエリ生成能力が悪化したと考えられる．Qwen3 は，文脈拡張により検索失敗は減少したが推論失敗が増加するトレードオフを示した．一方，Gemini は超大域的文脈で検索失敗および推論失敗ともに減少し，総失敗数は最小となった．以上より，超大域的文脈は検索精度向上に寄与しうる反面，モデルによっては推論能力の低下を招くことが示唆された．

表 3: 周辺文脈を要約したときのマクロ平均正解率のモデル間比較

Context	Search	Claude	Gemini	Llama	Qwen3
大域的	ON	0.754	0.667	0.562	0.332
超大域的	ON	0.694	0.686	0.569	0.324

4.3 周辺文脈の要約の有効性検証

本研究では，より洗練した情報を得るために周辺文脈として会話をそのまま用いるのではなく，会話を要約することを試みる．大域的な周辺文脈と超大域的な周辺文脈に対して要約を適用して GEC を行った結果を表 3 に示す．局所的な周辺文脈は対象の語彙が出現した 1 つの文であり，これ以上要約できないことから，今回の実験対象としなかった．

表 1 と表 3 を見ると，周辺文脈の要約の効果は LLM モデルによって異なることがわかる．周辺文脈の要約は Claude と Llama においては性能向上に寄与するが，Gemini と Qwen3 の性能を劣化させる．

最も性能が向上したのは Claude および超大域的文脈の条件であり，0.362 から 0.694 と値が約 2 倍近くになった．Sonnet-4.5 は超大域的文脈の値が大域的文脈よりも低く，今回のタスクにおいては長大な文脈を処理することが苦手であると判断できる．そのモデルに対しては，長大な文脈を要約することが効果的である．

全条件中で最も高い性能は要約を用いない Gemini であり，その値は 0.797 である．Gemini は Test-time Scaling を使ったモデルであるため，入力に要約を用いなくても，その推論の過程で要約と同等の情報が導出されたと考えられる．

5 結論

本研究は，希少な語彙あるいは固有名詞に対する ASR エラーを対象にした，Web 検索を用いた GEC に着目した．Web 検索を用いることで，修正候補の検討範囲に LLM 自身の知識だけでなく，最新の語句を含めることを狙った．実験の結果，Web 検索を用いることで LLM の知識がない語彙に関連する ASR エラーに対処できることが分かった．さらに，LLM への入力量が増えることで性能が向上する傾向が見られる一方で，LLM モデルによって最適な文脈範囲は異なることがわかった．そのため，今後の課題として，LLM モデルの特性に応じた文脈範囲を自動的に選ぶことが挙げられる．

参考文献

[1]Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. Automatic speech recognition errors detection and correction: A review. **Procedia Computer Science**, Vol. 128, pp. 32–37, 2018. 1st International Conference on Natural Language and Speech Processing.

[2]Horia Cucu, Andi Buzo, Laurent Besacier, and Corneliu Burileanu. Statistical error correction methods for domain-specific asr systems. In **Proceedings of the First International Conference on Statistical Language and Speech Processing**, SLSP'13, p. 83 – 92, Berlin, Heidelberg, 2013. Springer-Verlag.

[3]Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. Asr error correction and domain adaptation using machine translation. In **ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 6344–6348, 2020.

[4]Samrat Dutta, Shreyansh Jain, Ayush Maheshwari, Ganesh Ramakrishnan, and Preethi Jyothi. Error correction in ASR using sequence-to-sequence models. **CoRR**, Vol. abs/2202.01157, , 2022.

[5]Renjie Ma, Mufan Qian, Mark Gales, and Kevin Knill. ASR Error Correction Using Large Language Models. **IEEE Transactions on Audio, Speech and Language Processing**, pp. 1389–1401, 2025.

[6]Zeyu Liu, Junqi Wang, Jingyi Li, Jun Huang, Fei Huang, and Ruolan Xu. Multi-stage Large Language Model Correction for Speech Recognition. **arXiv**, 2024.

[7]Norihiro Yamashita, Munir Yamamoto, Haruya Kokubo, and Yoichi Kawaguchi. LLM-based Generative Error Correction for Rare Words with Synthetic Data and Phonetic Context. In **Interspeech 2025**, pp. 3653–3657, 2025.

[8]Sihan Li, Chen Chen, C. Y. Kwok, Chu Chu, Eng Siong Chng, and Hisashi Kawai. Investigating ASR Error Correction with Large Language Model and Multilingual 1-best Hypotheses. In **Interspeech 2024**, 2024.

[9]Sabyasachi Ghosh, Mohammad Sadegh Rasooli, Misha Levit, Peng Wang, Jing Xue, Dinesh Manocha, and Jiatao Li. Failing forward: Improving generative error correction for asr with synthetic data and retrieval augmentation. In **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 2466–2482, 2025.

[10]Mohammad Sadegh Rasooli, Sabyasachi Ghosh, Peng Wang, and Jiatao Li. Retrieval Augmented Correction of Named Entity Speech Recognition Errors. **arXiv preprint arXiv:2409.06062**, 2024.

[11]Christophe Van Gysel, Mirko Hannemann, Ernest Pusateri, Youssef Oualil, and Ilya Oparin. Space-efficient representation of entity-centric query language models, 2022.

[12]Ryo Sugano, Hidenori Sato, Akio Sakuma, Minoru Kumano, Yoshihiro Kawai, and Shinji Watanabe. 音声認識における固有名詞対策のための Tool-use, 2025.

[13]Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. **ArXiv**, Vol. abs/2310.04408, , 2023.

[14]Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2025.