Next Sentence Prediction に基づく文脈を考慮した ASR N-best のリランキング

邊土名朝飛 [1] 友松祐太 [1]
[1] 株式会社 AI Shift
{hentona_asahi, tomomatsu_yuta}@cyberagent.co.jp

概要

パイプライン型のタスク指向型音声対話システムにおいて，上流に位置する音声認識システムの音声認識誤りが後続処理に与える影響は大きい．本研究では，ユーザ発話の ASR 出力の N-best 候補と直前のシステム発話のペアを BERT に入力し，Next Sentence Prediction のアプローチでリランキングすることで，ASR システムの音声認識誤りを低減させることを試みる．また，リランキング性能の向上を目的として，質問応答データセットを用いて BERT の fine-tuning を行う．実験では，音声認識誤りを付与したデータセットを作成し，リランキング手法適用前後の音声認識誤り率を測ることで有効性を示す．

1 はじめに

タスク指向型の音声対話システムは，複数のモジュールから構成されるパイプライン型と，1 つのモデルで全ての処理を行う End-to-End 型の 2 つに分類される [1]．一般的なパイプライン型の音声対話システムは，自動音声認識（Automatic Speech Recognition; ASR）システムを用いて音声データであるユーザ発話をテキストに変換し，後段の各モジュールで言語理解（Natural Language Understanding; NLU）や行動決定（Policy），対話状態追跡（Dialogue State Tracking; DST）といった処理を行う [1]．パイプライン型の対話システムは各モジュールが独立しているため，機能ごとにモジュールを個別開発することができるほか，各モジュールの入出力が明確であり解釈しやすいという実応用上の利点がある．

しかし，先行するモジュールから順番に処理される性質上，あるモジュールで生じたエラーが後続のモジュールにも影響を与え，対話能力が低下してしまう欠点が存在する．特に，パイプラインの上流に位置する ASR システムの音声認識誤りが後続処理に与える影響は大きい．音声認識誤りは汎用 ASR システムを使用した場合により顕著に現れるが，[2] のような特定のドメインに特化した ASR システムを開発するためには多大なコストがかかる．そのため，音声認識誤りを改善することを目的として，BERT[3] などの大規模言語モデルを用いて ASR システム出力の N-best 候補をリランキングする手法が数多く提案されてきた [4, 5, 6]．

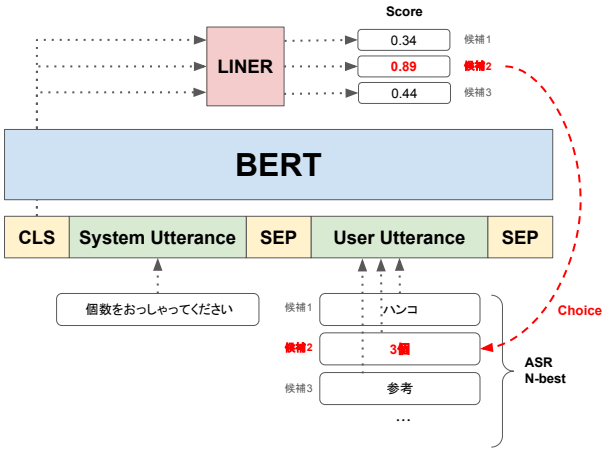

図 1 ASR N-best 候補のリランキングの概要図

本研究では，直前のシステム発話を文脈情報として与え，ユーザ発話の ASR N-best 候補をリランキングすることで，ASR システムの音声認識誤りを低減させることを試みる (図 1 参照)．N-best 候補のリランキングは，BERT を用いて Next Sentence Prediction(NSP) を行うことで実現する．さらに，数が非常に限られている日本語タスク指向型対話のデータセットの代わりに，質問応答データセットを用いて BERT の fine-tuning を行い，リランキング性能を向上させることを試みる．関連研究として，対話状態追跡 (Dialogue State Tracking; DST) タスクにおいては，質問応答データセットを用いて学習することで未知ドメインに対する性能が向上したことが報告されている [7, 8]．

実験では，音声認識誤りを付与したデータセットを作成しリランキング後の ASR 出力の音声認識誤り率を測ることで，NSP および質問応答データセットを用いた fine-tuning の有効性を検証する．

2 手法

2.1 BERT NSP

最も適した ASR N-best 候補を選択するためのアプローチとして NSP を採用する．NSP は，2 つの文を入力として与え，2 文目が 1 文目に続く文章であるかどうかを識別するタスクである (図 2 参照)．BERT-NSP の概要を図 1 に示す．本研究では，直前のシステム発話を 1 文目，後続のユーザ発話の ASR N-best 候補を 2 文目とみなし，NSP と同様の枠組みで N-best 候補の “適切さ”のスコアを定量的に推定する．ASR システムが出力した N 件の認識候補 (N-best) に対し，各候補のスコアをそれぞれ推定した後，スコアが高い順に N-best 候補を並び替える．最終的に，1 番目の候補 (i.e., 最もスコアの高い候補) を ASR 出力として選択する．BERT は事前学習として NSP タスクを解いているため，fine-tuning を行わない状態でもある程度は文脈的に適切な N-best 候補を選択できることが期待できる．

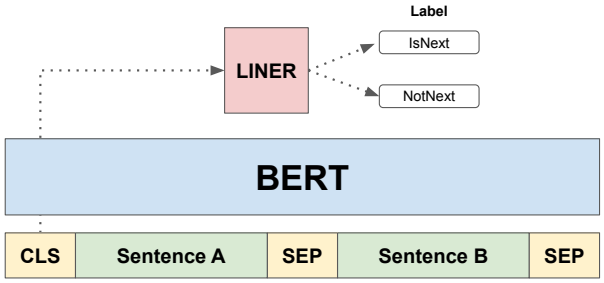

図 2 Next Sentence Prediction

2.2 QA BERT NSP

質問応答データセットを使用し，質問文を 1 文目，回答文を 2 文目とみなして NSP タスクを解くことで fine-tuning を行う．タスク指向型対話のシナリオの多くは「何日に予約しますか？」⟶「12 日の月曜日でお願いします」のように質問と回答が交互に行われることで進んでいくことから，質問応答タスクの一種と考えることができる．そこで本研究では，質問応答データセットから「会社の最高責任者を何というか？」⟶「社長」のような質問・回答ペアを作成し，それらのデータを用いて BERT を fine-tuning する．これにより，タスク指向型対話に適したモデルとなり，単純な NSP タスクで事前学習したモデルよりもシステム発話と N-best 候補の間の関係をより適切に捉えることができるようになると考えられる．

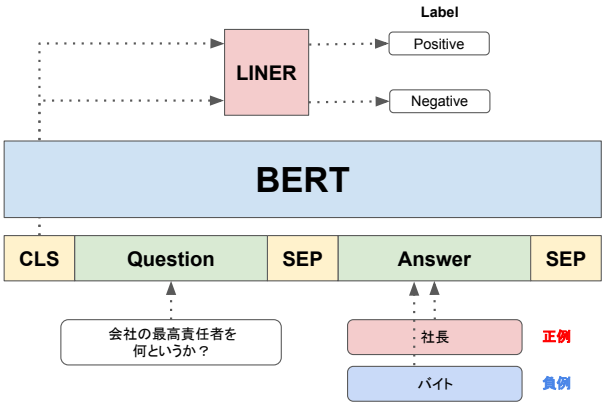

図 3 QA BERT NSP モデルの fine-tuning

3 実験

本章では，2 章で紹介した手法の評価実験を行う．

3.1 実験設定

3.1.1 評価用データセット

評価用データセットとして日本語質問応答データセットである JGLUE[9, 10] の JCommonsenseQA を採用し，dev set（データ数：1,119 件）を実験に用いた．JCommonsenseQA は，1 つの質問につき 5 つの回答選択肢が提示され，そのうち 1 つが正解となっている Multiple choice task データセットである．検証にあたり，JCommonsenseQA の質問文をシステム発話，回答文をユーザ発話とみなす．ただし，回答文には音声認識誤りは含まれていないため，音声合成エンジンを用いて回答文を音声に変換し，その音声を ASR システムに入力することで音声認識誤りを付与した．音声合成エンジンは pyopenjtalk[1)] を，ASR システムは Google Cloud Speech-to-Text[2)] を用いた．音声合成時には，弊社で運用している電話自動応答サービス[3)] の状況に近い音声認識誤りを付与するために，合成音声データのサンプリングレートを一般的な電話音声のものと同じ 8kHz にダウ

1) https://github.com/r9y9/pyopenjtalk
2) https://cloud.google.com/speech-to-text
3) https://www.ai-messenger.jp/voicebot/

表 1　評価結果

	w/ ground truth		w/o ground truth	
	WER [%]	CER [%]	WER [%]	CER [%]
Oracle	-	-	15.0	11.0
1-best	-	-	48.2	37.7
Random	-	-	93.3	70.3
BERT-NSP	42.7	32.4	64.7	49.5
QA-BERT-NSP	41.7	31.7	58.5	44.6

ンサンプリングし，音声品質を低下させた．また，ASR システムが出力する N-best 候補の数は最大 10 件（10-best）出力するように設定し，音声認識に失敗したサンプルは除外した．最終的に評価に用いたデータは 958 件となった．

3.1.2　モデル

実験では，質問応答データセットで fine-tuning した BERT(**QA-BERT-NSP**) および fine-tuning 無しの BERT(**BERT-NSP**) の 2 つのモデルを用いた．ベースとなる BERT モデルは，東北大学が公開している日本語 BERT[4] を採用した．

QA-BERT-NSP の学習には，JGLUE の JCommonsenseQA の `train set`（データ数：8,939 件）を使用し，質問・回答テキストのペアを作成した．質問・回答テキストのペアの正例と負例の比率は 1:1 に設定し，負例サンプルは正解を除いた 4 つの回答選択肢からランダムに 1 件選択することで獲得した．これにより，17,878 件の学習データを取得した．また，このうち 90% を学習用データ，残りの 10% を検証用データセットとした．学習時のパラメータについては，batch size は 32，learning rate は 5×10^{-5}，warmup ratio は 0.1，epoch 数は 5 に設定し，Validation loss が最小となったモデルを評価実験に使用した．その他のパラメータについては huggingface transformer のデフォルト設定に従った．

3.1.3　評価指標

適切な ASR N-best 候補を選択できたかを測る評価指標として Word Error Rate (**WER**) と Character Error Rate (**CER**) を採用した．WER 計算時に用いる Tokenizer は，§3.1.2 で説明した日本語 BERT のものを使用した．また，音声認識結果には基本的に含まれない句読点（、。）・疑問符（？）・感嘆符（！）はテキスト中から削除して評価を行った．

3.2　結果と考察

実験結果を表 1 に示す．ここで，表中の Oracle，1-best，Random について説明する．

- **Oracle**　ASR 出力の N-best 候補中から WER，CER が最小となる候補を選択した際の性能．すなわち，リランキング性能の上限値を表している．
- **1-best**　ASR 出力の N-best の第 1 候補を選択した際の性能．
- **Random**　ランダムに N-Best 候補を選択した際の性能．

次に，w/ ground truth，w/o ground truth について説明する．

- **w/ ground truth**　ASR 出力の N-best 候補中に正解データ (Ground truth) を含ませた場合の評価結果．
- **w/o ground truth**　ASR 出力の N-best 候補中に正解データ (Ground truth) を含めない場合の評価結果．

BERT-NSP と QA-BERT-NSP の WER，CER を比較すると，QA-BERT-NSP のエラー率が全ての評価項目で低いことがわかる．このことから，ASR 出力の N-best 候補のリランキングタスクにおいては，質問応答データセットで fine-tuning を行うことは有効であることが示唆される．ここで，w/ ground truth と w/o ground truth の結果を比較すると，w/o ground truth における BERT-NSP と QA-BERT-NSP の性能差よりも w/ ground truth における性能差が小さいことがわかる．この結果から，質問応答データセットによる fine-tuning では文脈的に正しい N-best 候補を選択する能力への寄与は小さいということが示唆さ

4）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

れる．

fine-tuning により文脈的に正しい N-best 候補を選択する能力があまり改善しなかったのにも関わらず，N-best の中に正解データを含めない w/o ground truth において性能に差がでた要因として，fine-tuning により名詞単体の N-best 候補を優先的に選択するようになったことが考えられる．質問応答データセットの回答は名詞 1 単語であることが多いため，それらのデータを用いて fine-tuning したことにより，名詞らしくない N-best 候補を選択しないような挙動になる．評価データも同じく質問応答データセットを用いているため，名詞らしくない候補を選ばないようにするだけでもエラー率が改善したと考えられる．

一方，1-best を見ると BERT-NSP と QA-BERT-NSP と比較してエラー率が 10 ポイント前後低いため，より文脈的に正しい N-best 候補を選択できるよう BERT モデルを改善する必要がある．Oracle を見ると，1-best と比較してエラー率が 1/3 以下であることから，ASR 出力の N-best 中には正しい出力候補が含まれている可能性が高いことを示しており，リランキングによる改善の余地が大きいといえる．

4 おわりに

本研究では，文脈情報として直前のシステム発話と，ユーザ発話の ASR 出力の N-best 候補のペアを BERT に入力し，Next Sentence Prediction のアプローチでリランキングすることで ASR システムの音声認識誤りを低減させることを試みた．また，リランキング性能の向上を目的として，質問応答データセットを用いて BERT の fine-tuning を行うことの有効性を検証した．音声認識誤りを付与したデータセットを用いた実験の結果，文脈的に正しい N-best 候補を識別する能力への fine-tuning の効果は限定的であることが示唆された．一方で，質問応答データセットで fine-tuning した BERT の方が fine-tuning していない BERT よりもエラー率は低下した．今後の課題として，より詳細なエラー分析と，多様な音声認識誤りパターンを生成し学習に用いることで音声認識誤りにロバストなリランキングモデルを構築することを検討していきたい．

謝辞

本論文の作成にあたりご協力頂きました，株式会社 AI Shift の杉山雅和氏，戸田隆道氏，東佑樹氏，二宮大空氏，下山翔氏にこの場を借りて厚く御礼申し上げます．

参考文献

[1] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. **SIGKDD Explor. Newsl.**, Vol. 19, No. 2, p. 25–35, nov 2017.

[2] Yong Zhao, Jinyu Li, Shixiong Zhang, Liping Chen, and Yifan Gong. Domain and speaker adaptation for cortana speech recognition. In **2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 5984–5988, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. Rescorebert: Discriminative speech recognition rescoring with bert. In **ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 6117–6121, 2022.

[5] Shih-Hsuan Chiu and Berlin Chen. Innovative bert-based reranking language models for speech recognition. In **2021 IEEE Spoken Language Technology Workshop (SLT)**, pp. 266–271, 2021.

[6] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. Effective sentence scoring method using bert for speech recognition. In Wee Sun Lee and Taiji Suzuki, editors, **Proceedings of The Eleventh Asian Conference on Machine Learning**, Vol. 101 of **Proceedings of Machine Learning Research**, pp. 1081–1093. PMLR, 17–19 Nov 2019.

[7] Shuyang Gao, Sanchit Agarwal, Di Jin, Tagyoung Chung, and Dilek Hakkani-Tur. From machine reading comprehension to dialogue state tracking: Bridging the gap. In **Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI**, pp. 79–89, Online, July 2020. Association for Computational Linguistics.

[8] Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, and Pascale Fung. Zero-shot dialogue state tracking via cross-task transfer. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7890–7900, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp.

2957–2966, Marseille, France, June 2022. European Language Resources Association.
[10] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 言語処理学会第 28 回年次大会, 2022.