

ALBERT および LSTM ベースのモデルによる対話破綻検出

戎 淳¹ 曹 婷² 木村 優平¹ 片倉 多智¹ 早坂 絵央¹ 吉原 圭祐¹ 西村 百之輔¹ 延澤 志保³*¹京都市大学知識工学部²早稲田大学大学院基幹理工学研究科³京都市大学情報工学部

1 対話システムの評価

品質の良い対話システムを開発するには数多くの問題を解決する必要がある。その中でも、対話システムの発話の品質を確保することは重要と考えられる。対話システムの発話に関して、さまざまな品質の要求がある。タスク指向型の対話システムの発話品質においては、発話がユーザのタスクの達成率と対話に対してユーザーの満足度などを評価基準として判断されるのは一般的である。一方、非タスク指向型対話システムの発話品質においては、タスク指向型の対話システムは特定なタスクがないため、非タスク指向型の評価基準では適用できない。そのため、対話の継続性を評価基準として評価することが多いと考えられる。人間と対話システムとの間で生じる対話破綻を自動検出できる技術が確立されると、対話システムは発話を行う前に出力予定の発話が対話破綻につながるかどうか分かり、生成する発話が破綻になるのを回避することが可能となる。

1.1 対話破綻検出

対話破綻とは、対話システムがユーザの発話を正しく理解していないまま破綻のある発話を生成し、ユーザと対話システムの会話が継続できないことを指す。対話破綻検出はダイアログコンテキストで対話システムの発話が対話破綻を引き起こす可能な内容を検出することである [1]。DBDC データセット¹のサンプルを表 1 に示す。

表 1: 対話破綻のサンプル [2]

System:	買い物は一人が楽です
User:	確かに気が楽ですね
System:	買い物は長いです
User:	まあ見るのも楽しいし
System:	買い物は一緒に楽しいですね

表 1 に見られるとおり、対話システムは「買い物は一人が楽です」と発話したにも関わらず、3 回目の発話は「買い物と一緒に楽しいです」と発話し、1 回目の発話と食い違っている。サンプルのように、対話システムがユー

ALBERT and LSTM-based Models for Dialogue Breakdown Detection.

Bo Rong (戎 ボツ)¹, Ting Cao (曹 テイ)², Yuhei Kimura¹, Taichi Katakura¹, Kaio Hayasaka¹, Keisuke Yoshihara¹, Momonosuke Nishimura¹, and Shiho Hoshi Nobesawa³*

¹ Faculty of Knowledge Engineering, Tokyo City University

² Graduate School of Fundamental Science and Engineering, Waseda University

³ Faculty of Information Technology, Tokyo City University

¹DBDC データセット, <https://sites.google.com/site/dialoguebreakdown-detection4/datasets>.

ザの発話を理解できないため、正しくない発話は対話の破綻とみなすことができる。

1.2 機械学習による対話破綻検出

BERT モデルに基づいて、分析を通じて、センテンスベクトルは推定に効果的であった。Sugiyama らの提案したモデルは対話破綻検出で優れたパフォーマンスを発揮した [3]。Sugiyama らは DBDC3 データセットと DBDC4 データセットの 2 つトレーニングデータセットでの実験結果の比較から、トレーニングデータセットが主要なパフォーマンスであることを示した [3]。Sugiyama らは BERT 事前学習モデルは対話コーパスでトレーニングされていないため、モデルの予測は最良の結果ではない可能性があるとして述べた。Hendriksen らは、LSTM モデルに基づいて、対話破綻検出のための多項式シーケンス分類器を提案すると同時に、モデルアーキテクチャと単語埋め込みモデルのパフォーマンスについても調査した [4]。

2 対話破綻の検出

本研究では、非タスク指向型対話システムの対話品質を向上させる方法として、オープンメインでユーザーとの対話の継続性に影響する対話破綻の検出方法を提案する。本研究では事前学習モデル ALBERT を用いて対話コンテンツを理解させて、Bi-LSTM を使って学習させ、人間のように対話破綻を認識できることを目指している。

2.1 提案手法の概要

本研究の処理の流れを図 1 に示す。

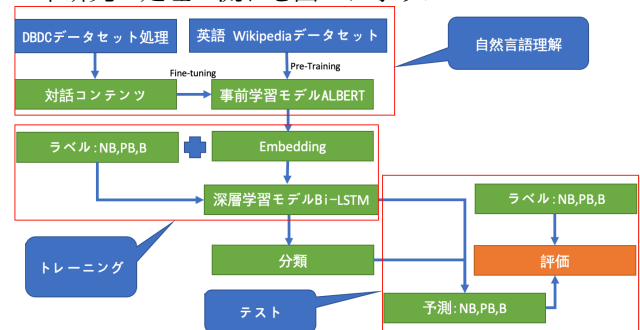


図 1: 本研究の処理の流れ

図 1 のように、英語 Wikipedia データセットで事前学習させたモデル ALBERT に、対話コンテンツをインプットとして、事前学習モデル ALBERT を微調整した embedding を取得する。取得した embedding と人間が評価したラベルを深層学習モデル Bi-LSTM に学習させ

る。テストはモデルにラベルを推測させ、推測したラベルと人間が評価したラベルと比べ、モデルの評価を行う。

2.2 DBDC データセット

本研究ではDBDC(Dialogue Breakdown Detection Challenge) データセット [1] を用いた。DBDC データセットには、対話のコンテンツはユーザーとシステムの対話コンテンツが含まれている。システムが生成した発話は人手によって評価され、表 2 のラベルが付けられている。これは Higashinaka らの提案した破綻 (B), 破綻かもし

表 2: 対話コンテンツのラベル [1]

NB:	対話をスムーズに続けることが可能である。
PB:	対話をスムーズに継続することは困難である。
B:	対話を続けることは困難である。

れない (PB), 破綻ではない (NB) の 3 種類のラベル [1] に対応する。本研究では DBDC データセットの IRIS データセットを用いた。合計 100 の対話があり、それぞれに 20 の発話、合計 2000 の発話、1000 のラベルがある。

2.3 学習モデル

2.3.1 事前学習モデル ALBERT

事前学習モデル ALBERT(A LITE Bidirectional Encoder Representation from Transformers)[5] は多くの自然言語処理タスクに使用されている。ALBERT モデルは SQuAD2.0(Stanford Question Answering Dataset)¹ 読解問題のタスク [6] で人間よりも高いスコアを取得した。ALBERT が破綻検出のパフォーマンスを向上させることが期待できる。英語 Wikipedia データセットは多くの知識や単語が含まれているので、本研究は英語 Wikipedia データセットを使い、事前学習モデル ALBERT に事前学習させたモデルを用いた。事前学習モデル ALBERT を DBDC データセットを用いて微調整した。

2.3.2 深層学習モデル Bi-LSTM

深層学習モデル LSTM(Long Short-Term Memory)[7] は長期の時系列データを学習することができる。自然言語処理に応用されて、大きな成果をあげている。本研究は深層学習モデル Bi-LSTM(Bi-directional Long Short-Term Memory) を用いた。

3 実験結果と考察

ベースラインプログラムは、各発話に含まれる単語を特徴 (Bag-of-Words) として使用し、条件付き確率場 (CRF) を使用して対話破綻を検出する [2]。正確率, F-measure, JSD(Jensen-Shannon Divergence), MSE(Mean squared error) を用いて評価した。実験結果を表 3 に示す。

表 3 が示したように、本提案はベースラインより正確率が高く、JSD や MSE もベースラインより小さい。JSD は 3 つのラベルの予測分布と、Jensen-Shannon Divergence によって計算された最良ラベルの予測分布との間の距離

表 3: 本研究とベースライン [8] の実験結果

	baseline	ALBERT+LSTM
Accuracy	0.4635	0.6300
F-measure (B)	0.3421	0.4556
JSD	0.4381	0.0630
MSE	0.2237	0.0314

である。MSE は 3 つのラベルの予測分布と、平均二乗誤差によって計算された最良ラベルの予測分布との間の距離である [2]。

4 まとめ

品質の良い対話システムを開発するため、対話システムの発話内容の品質を確保することは重要であると考えられる。本研究では、雑談対話システムを対象として、ユーザとの会話の継続性の向上のため、システムの対話破綻検出の精度を向上させる方法を検討する。提案手法では、ALBERT を用いた事前学習と、LSTM を用いた深層学習による対話破綻検出のモデルとを用いる。DBDC データセットで実験を行い、事前学習モデル ALBERT を微調整することで、提案モデルはベースラインのパフォーマンスのモデルよりも優れていることが示された。

参考文献

- [1] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3146–3150. European Language Resources Association (ELRA), may 2016.
- [2] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. Overview of the dialogue breakdown detection challenge 3. *Proceedings of Dialogue System Technology Challenge 6 (DSTC6) Workshop*, 2017.
- [3] Hiroaki Sugiyama. Dialogue breakdown detection using BERT with traditional dialogue features. In *Proc. Dialogue system technology challenge 6*, 2019.
- [4] Mariya Hendriksen, Artuur Leeuwenberg, and Marie-Francine Moens. LSTM for dialogue breakdown detection: exploration of different model types and word embeddings. In *Proc. Dialogue system technology challenge 6*, 2019.
- [5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net, 2020.
- [6] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. Association for Computational Linguistics, July 2018.
- [7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, 1997.
- [8] Ryuichiro Higashinaka, Luis F. D'Haro, Bayan Abu Shawar, Rafael E. Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Joao Sedoc. Overview of the dialogue breakdown detection challenge 4. *Proceedings of WOCHAT*, 2019.

¹SQuAD2.0, <https://rajpurkar.github.io/SQuAD-explorer/>.