ルーブリックを用いたタスク指向合成対話コーパスの品質評価手法

粟井修司[1] 吉岡隆宏[1]
[1]富士通株式会社
{awai.shuji,y.takahiro}@fujitsu.com

概要

大規模言語モデルによる合成対話データは広く利用されるが，下流タスクへの貢献度を測ることは困難であった．Perplexity 等の既存指標は，この機能的品質を保証しない．本研究ではこの課題に対し，実データから抽出したルーブリック（評価基準）を軸に，合成データと実データの意味的な分布の乖離を「距離スコア」として定量化する新手法を提案する．満足度分類タスクにおける評価実験の結果，提案する距離スコアは，既存指標よりもモデルの精度と強く相関することを示した．

1 はじめに

近年，大規模言語モデル（LLM）の発展に伴い，タスク指向対話システムの開発プロセスも大きく変化している．特に，人手によるデータ収集のコストと時間を削減するため，LLM を用いて大量の合成会話データを生成し，対話モデルの学習やシミュレーションに活用するアプローチが増えている[1, 2]．しかし，生成された合成データは一見流暢で自然に見えても，実世界のユーザーとのインタラクションが持つ微妙なニュアンスや，タスク達成過程における満足・不満足といった重要な質的パターンを忠実に再現できているとは限らない．この品質の乖離は，学習済みモデルの性能低下に直結する課題である．

この問題の原因として，既存の評価手法の限界があげられる．Perplexity や BLEU[3]，distinct-n[4]といった指標は，主にテキストの言語的な流暢さや表層的な多様性といった生成的品質を測定する指標である．これらの指標は，生成モデル自体の性能評価には有用であるが，そのデータが下流タスクの学習にどれほど有益かという機能的品質を保証するものではない．例えば，Perplexity が低い対話データセットで学習したモデルが，必ずしも満足度予測の精度が高いとは限らない．

さらに，これらの評価は個々の会話単位で行われることが多く，データセット全体が持つ分布の偏りや多様性の欠如といったコーパスレベルでの品質劣化を定量的に捉えることが困難であった．そこで本研究では，合成データセットが，基準となる実データセットのタスク遂行に重要な機能的品質をどれだけ保持しているかを評価するための新しい品質評価フレームワークを提案する．

本研究の貢献は以下の通りである．

1. **ルーブリックベースの品質評価フレームワークの提案**： 実データから抽出したルーブリック（評価基準）[5]を共通の評価軸とし，合成データセットが実データセットの持つ意味的な分布をどれだけ再現しているかを測定する．
2. **提案フレームワークの有効性の検証**： 提案する品質指標が，従来の品質指標よりも，下流タスク（満足度分類）における AI モデルの性能と強く相関することを示す．これにより，本手法が単なる分析ツールではなく，「より優れた AI モデルを作成するための，より優れたデータ」を見分けるための実用的な指標であることを検証する

3 提案手法

本研究では，合成対話コーパスの機能的品質を評価するための新しいフレームワークを提案する．本フレームワークは，「機能的に優れた合成データとは，実データが持つ品質特徴の統計的な分布を忠実に再現するものである」という仮説に基づいており，以下の 3 つのステップで構成される。

3.1 基準ルーブリックの生成

本手法では，まず評価の基準となる共通の評価軸を定義する．Perplexity や BLEU がテキストの表層的な品質を測るのに対し，タスクの成否に直結する意味的な側面を評価するため，基準となる実世界の会話データセットであるD_{ref}を用い，RUBICON[5]の

手法に基づいて会話の品質を測る「満足（SAT）」および「不満足（DSAT）」のルーブリック群$R = \{r_1, r_2, r_3, \dots, r_k\}$を生成する．ルーブリックは，会話がSAT/DSATと判断された理由を評価項目として言語化したものである．例えば，「The assistant presents clear and actionable options.」などがある．また，各会話cが各ルーブリックにどの程度当てはまるのかをスコアとして出力する．スコアは，ルーブリックが当てはまらない場合は 0，当てはまる場合はその度合いに応じて 4 段階のスコアが付与される．このルーブリック群Rをすべての評価における共通の評価軸として用いる．

3.2 品質特徴量の定義

次に，各会話cを，ルーブリック群Rを用いて品質特徴量ベクトルに変換する．本研究では，以下の 4 つの特徴量を定義する．

- 適用率（Application Rate）
  特定の会話cにおいて，全ルーブリックのうち適用された（スコアが 0 でない）割合．これは，一つの対話において，タスクの成否に関わる意味的イベントがどれだけ多様な種類で発生したかを示す指標である．

$$AR_{SAT}(c) = \frac{|\{r \in R_{SAT} | Score(r,c) \neq 0\}|}{|R_{SAT}|}$$

- スコア平均（Average Score）
  適用されたルーブリックのスコアの平均値．これは，発生した各イベントの質的な強度が平均してどの程度であったかを示す指標である．

$$AS_{SAT}(c) = \frac{\sum_{r \in R_{SAT}} Score(r,c)}{|\{r \in R_{SAT} | Score(r,c) \neq 0\}|}$$

上記 2 式は SAT の例であり，DSAT も同様にAR_{DSAT}とAS_{DSAT}を算出する．

3.3 分布の距離スコア算出

最後に，データセット全体の品質を，D_{ref}との分布の近さとして定量化する．まず，D_{ref}における各満足度ラベルl（SAT/DSAT）の品質特徴量の平均値μ_lと，データセット全体での標準偏差σを計算する．次に，評価対象の合成データセットD_{syn}内の各会話c（ラベル l_cを持つ）について，D_{ref}との距離スコア（Distance Score）を定義する．

$$DS(c) = \sum_{f \epsilon Features} \frac{|Feature_f(c) - \mu_{l_c,f}|}{\sigma_f}$$

この距離スコアは，会話cが D_{ref}の同じラベルを持つ会話群の平均的な品質からどれだけ離れているかを表す．値が小さいほど，その会話は「実データらしい」と見なされる．

4 実験

本章では，提案する品質評価フレームワークの有効性を検証するための実験の詳細を述べる．

4.1 評価方法

評価方法として，提案する品質指標と下流タスク性能との相関を分析する．具体的には，品質の異なる複数の学習データセットを用意し，それぞれで学習したモデルの性能を共通のテストデータで評価する．そして，各データセットの品質指標スコアと，学習したモデルの精度との相関を算出する．

4.2 データセット

顧客とオペレーターの対話ログデータセットである MAIA-DQE[6]を用いた．評価のために，OpenAI の LLM である GPT-5 を用いて顧客満足度を 2 値（SAT/DSAT）に分類しラベルとして用いた．データセットの構成は以下のとおりである．

- 基準データ（D_{ref}）
  MAIA-DQE の学習用データ（183 件）．提案指標で用いる基準ルーブリックの生成と，距離スコアの算出で用いる平均値μと標準偏差σの算出，および BERTScore と BLEU の参照コーパスとして用いた．
- 評価対象データ（D_{syn}）
  D_{ref}をシードとして SynTOD[2]に基づき生成した合成対話コーパス（1,000 件）．
- 学習データセット群（$D_{train\ i}$）
  D_{syn}の各会話に距離スコアを付与し，スコア昇順（品質が高い順）にソートする．本実験では，このソート済みデータを利用し，品質レベルを変化させた複数の学習データセットを構築した．具体的には，取得開始順位を 1, 10, 20, ..., 100 の 11 段階に設定し，各順位から以下のようにデータをサンプリングした．
  検証 1：各取得開始順位から 100 件サンプリングしたデータセット群．
  検証 2：各取得開始順位から 50 件サンプリングし，これにD_{ref}からランダム抽出した 50 件を加えたデータセット群．

- テストデータ（D_{test}）
  MAIA-DQE の評価用データ（183 件）全モデルで共通のホールドアウトテストセットとして用いた．

4.3 比較対象の品質指標

各学習データセット群に対し，提案指標と複数の従来指標を算出し，下流タスク精度との相関を比較した．提案指標としては距離スコア（DS）と，その構成要素である品質特徴量（AR_{SAT}, AR_{DSAT}, AS_{SAT}, AS_{DSAT}）を用いた. 比較対象とする従来指標として，まずテキストの言語的な品質を測る尺度である，gpt2 モデルを用いて算出した Perplexity と，コーパス全体における非重複な n-gram の割合である distinct-n を採用した．また，生成されたデータセットが基準データD_{ref}とどの程度類似しているかを評価するため, BERTScore と BLEU スコアを算出した．これらのスコアは，学習データで用いるD_{syn}の全てのデータとの類似度を計算し，その中の最大値をその会話（または発話）のスコアとした．最終的に，これらのスコアのデータセット全体での平均値を，その学習データセット群の評価値とした．この計算により，生成されたデータが基準データの持つ多様な表現をどの程度網羅しているかを評価できる．

4.4 評価タスク

指標の有効性を異なる条件下で検証するため，2 種類の満足度分類タスクを設定した．結果の頑健性と再現性を担保するため，各学習データセットについて 100 回の独立した学習試行を実施した．各試行は異なる乱数シードを用いて行い，得られた 100 個の macro F1 スコアの平均値を，そのデータセットにおける最終的なモデル精度とした．

- 検証 1：ルーブリックベース検証
  提案指標算出で用いるルーブリックを用いたタスクにおいて，データの学習のしやすさを表現できているか検証することを目的とする．具体的には，各会話のルーブリックスコアを特徴量として入力し，4 つの機械学習モデル（LogisticRegression， SVM， RandomForest, LightGBM）を用いて満足度を予測した．特徴量には RFE （Recursive Feature Elimination）[7]を適用し，上位 20 次元を選択して使用した．
- 検証 2：汎用モデル検証
  ルーブリックスコアを直接利用できない，より公平かつ実用的な条件下で，提案指標がデータセットの本質的な学習のしやすさを表現できているか検証することを目的とする．具体的には，各会話の生テキスト全体を事前学習済み言語モデル（deberta-v3-base）[8]に入力して特徴量ベクトルを抽出し，検証 1 と同様の 4 つの機械学習モデルで満足度を予測した．

5 結果と考察

5.1 各検証タスクにおける相関分析

各検証タスクにおける各品質指標とモデル精度との相関を算出した結果を表 1 に示す．

検証 1（ルーブリックベース）では，提案指標の DistanceScore はモデル精度に対し-0.82 から-0.86 という極めて強い負の相関を示した．この結果は，提案指標がモデルの学習データとしての適性を適切に捉えていることを示唆する．一方，BLEU スコアも 0.71 から 0.87 という強い正の相関を示した. これは，本検証がルーブリックスコアそのものを入力特徴量としていることに起因し，テキストの表層的類似度が高いデータが結果的にルーブリックの付与パターンの基準データと近くなったためだと考えられる．

これに対し，検証 2（汎用モデル）では，線形モデル（LogisticRegression, SVM）における, DistanceScore の相関は-0.70 および-0.61 と高い値を維持したのに対し，BLEU の相関は 0.56 と 0.57 に低下した．これは，下流タスクの性能予測において，単なるテキストの表層的類似性よりも，提案指標の方が，より本質的かつ頑健な評価尺度であることを強く示唆しているといえる．

5.1 各モデルと品質評価・精度の関係

図 1，図 2 に各機械学習モデルの精度との関係について示す．図 1 が示すように，検証 1 では全モデルで取得開始順位が上がるに伴い精度が低下している．しかし，図 2 が示す検証 2 では，異なる傾向が観測された．

表 1 では，検証 2 において決定木ベースの非線形モデル（RandomForest, LightGBM）の場合，ほとんどの品質指標との相関が弱くなっている．これは，検証 2 の学習データが「品質の高い実データ 50 件」と「品質が変動する合成データ 50 件」が含まれていることが影響していると考えられる．

DistanceScore が高い学習データセットほど，含ま

表 1　各検証タスクにおける品質指標とモデル精度（macro F1 スコア）の相関

モデル	提案指標					既存指標			
	DS	AR_{SAT}	AR_{DSAT}	AS_{SAT}	AS_{DSAT}	perplexity	distinct-2	BERT Score F1	BLUE-4
検証 1									
LogisticRegression	**-0.86**	-0.45	0.84	-0.74	0.64	0.58	0.62	0.20	0.77
SVM	-0.86	-0.48	0.85	-0.75	0.58	0.69	0.66	0.40	**0.87**
RandomForest	**-0.83**	-0.36	0.81	-0.71	0.57	0.56	0.67	0.15	0.73
LightGBM	**-0.82**	-0.38	0.81	-0.71	0.54	0.58	0.67	0.19	0.71
検証 2									
LogisticRegression	-0.70	-0.62	**0.74**	-0.69	0.66	0.69	0.45	0.49	0.57
SVM	-0.61	-0.48	0.65	-0.51	**0.71**	0.57	0.32	0.47	0.56
RandomForest	0.07	-0.25	-0.08	0.00	**-0.31**	-0.05	0.04	0.11	0.03
LightGBM	0.17	**-0.36**	-0.12	0.02	0.10	-0.21	0.12	-0.15	-0.26

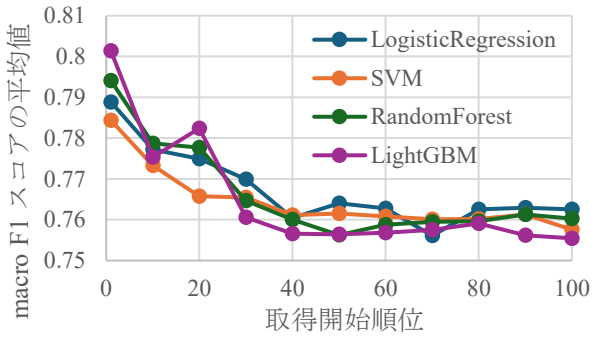

図 1　検証1におけるモデル精度

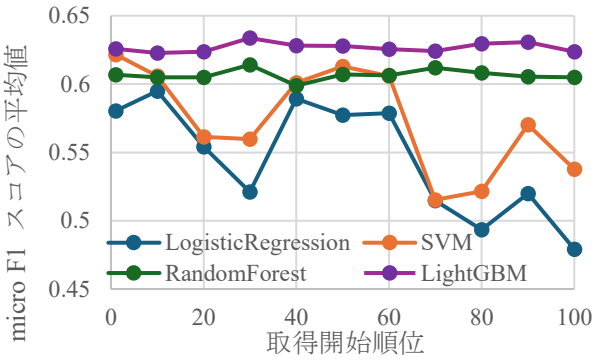

図 2　検証2におけるモデル精度

れている実データと，品質の低い合成データとの品質の乖離が大きくなる．線形モデルは，データ全体の分布から学習するため，この品質の乖離に強く影響され，品質の低い合成データが含まれることで精度低下したと考えられる．これに対し，決定着ベースのモデルは，データセット内の局所的なパターンを捉える能力に長けている．そのため，品質の高い実データが形成するパターンと，品質の低い合成データが形成するパターンを，ある程度分離して学習することが可能である．結果として，品質の低い合成データが含まれることによる影響を緩和できたため，DistanceScore との相関が弱まり精度低下を防げたと考えられる．

ただし，この無相関という結果は，本実験で用いた合成データの品質が相対的に低かったことに起因する可能性が高い．品質の乖離が小さい合成データを用いた場合，決定木ベースのモデルにおいても提案指標との相関を観測することが期待される．

6 おわりに

本研究では，合成対話データが下流タスクの学習にどれほど有益か，という「機能的品質」を測定するための新しいフレームワークを提案した．既存の指標が捉えきれなかったコーパスレベルでの品質劣化を評価するため，実データから抽出したルーブリックを共通の評価軸とし，合成データが実データの持つ意味的な分布をどの程度再現しているかを定量化した．

実験の結果，提案する品質指標は，Perplexity や BLEU といった既存の指標よりも，下流タスク（満足度分類）における AI モデルの性能と強く相関することが確認された．この結果は，本手法が「より優れた AI モデルを作成するための，より優れたデータ」を見分けるための実用的な指標となり得ることを示唆している．

今後の展望として，他ドメインへの適用性の検証や，データ生成プロセスを改善するためのフィードバックループへの応用があげられる．

参考文献

[1] Megan Leszczynski et al. Talk the walk: Synthetic data generation for conversational music recommendation. arXiv preprint arXiv:2301.11489, 2023.

[2] Chris Samarinas et al. Simulating task-oriented dialogues with state transition graphs and large language models. arXiv preprint arXiv:2404.14772, 2024.

[3] Kishore Papineni et al. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002.

[4] Jiwei Li et al. A diversity-promoting objective function for neural conversation models. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016.

[5] Param Biyani et al. Rubicon: Rubric-based evaluation of domain-specific human ai conversations. Proceedings of the 1st ACM International Conference on AI-Powered Software, 2024.

[6] John Mendonca et al. Dialogue quality and emotion annotations for customer support conversations. Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), 2023.

[7] Isabelle Guyon et al. Gene selection for cancer classification using support vector machines. Machine learning, Vol. 46, No. 1, pp. 389-422, 2002.

[8] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543, 2021.