物語文評価における評価モデル統合手法としての多次元項目反応理論の検討

有馬士央[1] 菊池英明[2]
[1]早稲田大学大学院 [2]早稲田大学
s.arima@suou.waseda.jp
kikuchi@waseda.jp

概要

本研究では，物語文の自動評価において複数の評価指標を統合する手法として，多次元項目反応理論(IRT)の適用を検討した．METEOR，BERTScore，COMET，Perplexity を対象に，人手評価との相関を単体評価，単純平均，IRT 統合で比較した．その結果，順位相関では単体指標や単純平均が強力なベースラインとなる一方，識別力パラメータを用いた重み付け統合では線形相関が向上することが示された．さらに，多次元 IRT により評価指標の役割の違いを分析し，Perplexity の特異性を明らかにした．

1 はじめに

近年，自然言語生成技術の発展に伴い，自動生成された文章の品質を適切に評価する手法の重要性が高まっている．特に，物語生成や長文生成の分野においては，文法的な正しさのみならず，内容の一貫性や自然さ，さらには人間が感じる主観的な評価をどの程度反映できているかが重要な課題となっている．従来の自動文章評価では，BLEU[1]や METEOR[2]に代表される参照文との類似度に基づく指標や，BERTScore[3]や COMET[4]のように文脈表現を用いた高度な評価指標が広く用いられてきた．一方で，これらの評価指標はそれぞれ異なる観点から文章を評価しており，単一の指標のみでは人手評価を十分に再現できない場合があることが指摘されている．

また，参照文を必要としない評価手法として，言語モデルの Perplexity[5](以下，PPL)が広く利用されている．PPL は文章の流暢さを測る指標として有用であるが，物語としての面白さや内容的妥当性を直接評価するものではない．そのため，PPL 単体で人手評価と高い整合性を示すとは限らない．

また，物語に特化しているモデルとして StoryER[6]がある．StoryER は Ranking/Rating/Reasoning を統合し，対人相関が高いが，学習データやカテゴリに依存があり評価の精度にもまだ課題が残る．評価モデルごとに観点とスケールが異なるため，単独・単純集約では一貫した品質判断が難しい．

このような背景から，本研究では複数の自動文章評価指標を統合する枠組みに着目する．具体的には，心理測定学の分野で用いられてきた項目反応理論(IRT)を用いて，複数の評価指標を統合し，文章の潜在的な品質を表す能力値を推定する．そして，推定された能力値と人手評価との相関を分析することで，IRT による統合評価の有効性を検証する．

本研究の貢献は以下の 2 点である．

(1) 複数の参照あり・参照なし自動評価指標を IRT により統合する手法を提案する．

(2) 統合された能力値と人手評価との相関を定量的に評価する．

2 先行研究

2.1 参照あり自動評価指標

BLEU や METEOR は，生成文と参照文との n-gram の一致度に基づいて評価を行う指標であり，機械翻訳評価において広く利用されてきた．METEOR は語順や同義語を考慮する点で BLEU よりも柔軟な評価が可能であるとされている．近年では，分散表現を用いた評価指標が提案されている．BERTScore は，BERT による文脈埋め込みを用いて生成文と参照文の意味的類似度を評価する手法である．また，COMET は事前学習済み言語モデルを用いて，人手

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

評価との一致度を学習することで，高い評価性能を示している．

2.3 参照なし自動評価指標

参照文を必要としない評価指標として，言語モデルの Perplexity がある．PPL は文章が言語モデルにとってどれだけ予測しやすいかを示す指標であり，文法的流暢さの評価に適している．しかし，内容の妥当性や物語性といった高次の要素を直接評価することは困難である．

また物語特化型の StoryER では人間評価がすでに行われている文章を学習し，人間評価との相関が既存の手法よりも高い結果になった．しかし，学習済みモデルの公開やそれぞれの文章につけた総合的なスコア等は公開されておらず，再現性が低い．

2.2 評価指標の統合

複数の評価指標を統合する試みとしては，単純な平均や重み付き平均を用いる方法が一般的である．しかし，これらの手法では各指標の特性や信頼性の違いを十分に考慮できない．

項目反応理論は，テスト項目の難易度や識別力を考慮しながら，被験者の潜在能力を推定する枠組みであり，評価指標統合への応用が期待されている．小論文自動採点の分野では IRT を用いた複数モデルの統合の有用性が報告されている[7]．そのため本研究では，IRT を用いて自動文章評価指標を統合し，人手評価との整合性を検証する．

3 提案手法

本研究では，複数の自動文章評価指標を項目反応理論（IRT）により統合し，文章の潜在的品質を表す能力値 θ を推定する．本章では，使用する評価指標および IRT に基づく統合手法について説明する

3.1 使用する評価指標

本研究では，METEOR/BERTScore/COMET/Perplexity の 4 つの自動評価指標を用いる．METEOR，BERTScore，COMET は参照文を用いた評価指標であり，生成文と参照文の類似性を異なる観点から評価する．一方，PPL は参照文を必要としない指標であり，文章の流暢さを評価する．なお本研究では再現性の低さや評価尺度の違いから StoryER[6]は用いない．

3.2 項目反応理論による統合

IRT では，各評価指標を「項目」，各文章を「被験者」とみなし，文章の潜在能力値 θ を推定する．本研究では，評価指標の出力を 5 段階に離散化し，順序尺度として扱う．IRT モデルにより推定された能力値 θ は，各評価指標の特性を考慮した統合スコアであり，単一指標では捉えきれない文章品質を表現できると期待される．本研究では 1-5 段階というデータ特性と評価モデル間の特性差(厳しさ・識別力)を同時推定し揃えることを踏まえて Graded Response Model (以降 GRM)が最適であると判断した．

4 実験

本研究では，自動文章評価指標を項目反応理論により統合し，推定された能力値と人手評価との整合性を検証する．本章では，使用したデータセット，評価指標，および実験手順について説明する．

4.1 データセット

本研究では，StoryER で公開されている文章データセットを用いて実験を行った．データセットの各文章には，人間の評価者による 15 段階の主観評価が与えられており，文章の総合的な品質を表している．本実験では，この人手評価の平均値を人手評価スコアとして用いた．評価対象となる文章数は 806 件であり，各文章に対して参照文が 1 件以上付与されている．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

4.2 使用する評価指標

本研究ではMETEOR/BERTScore/COMET/Perplexity の 4 つの自動評価指標を用いた．METEOR，BERTScore，COMET は参照文を用いた評価指標であり，生成文と参照文の意味的類似性を異なる観点から評価する．一方，PPL は参照文を必要とせず，事前学習済み言語モデルに基づいて文章の流暢さを評価する指標である．

4.3 評価指標の前処理(離散化)

各自動評価指標の出力は連続値であるため，IRT モデルへの入力として使用するために 5 段階評価に離散化した．具体的には，各指標のスコアに対して順位に基づく分位点分割を行い，1（低評価）から 5（高評価）までのカテゴリ値に変換した．なお，Perplexity は値が小さいほど良い評価であるため，他の指標と評価方向を揃える目的で，カテゴリ化の際に順位を反転させた．

4.4 実験手順

実験の手順を説明する．まず各文章に対して 4 つの評価モデルのスコアを算出した．次に各指標のスコアを 5 段階評価に変換し，変換後のスコアに IRT を適用して各文章の能力値 θ を推定した．最後に推定された θ と人手評価平均値との相関を算出した．相関の評価には，Spearman の順位相関係数，Kendall の順位相関係数，および Pearson の相関係数を用いた．

5 結果

本章では，人手評価との一致度を，順位に基づく相関（Spearman，Kendall）および線形相関（Pearson）の 3 種類の相関係数を用いて評価する．評価観点の違いを明確にするため，相関係数ごとに結果を示す．

5.1 実験設定と評価方法

評価対象は 806 本の物語文であり，人手評価は 1–5 段階の平均値を用いた．自動評価指標として，参照文を用いる METEOR，BERTScore，COMET を用いた．各指標の値は単純平均を算出する際には正規化を行った．一方，IRT による統合では指標間のスケールを揃えずに分析を行った．また，本研究では単純平均，IRT による潜在能力値 θ，および IRT で推定された識別力パラメータ a を用いた重み付け統合を比較する．これにより，単純な集約とモデルベースの統合手法の違いを明確にする．

5.2 Spearman 順位相関による比較

表 1　人手評価との Spearman の順位相関係数 (n = 806)

手法	Spearman ρ
METEOR	0.489
BERTScore	0.425
COMET	0.412
Perplexity	0.056
単純平均(正規化)	0.460
IRT(GRM)	0.457
IRT(a 重みづけ)	0.480

表 1 に, 各評価手法と人手評価との Spearman 順位相関係数を示す．Spearman 相関では，METEOR が最も高い値を示した．一方，単純平均および IRT による統合手法はいずれも単体指標と同程度の相関を示しており，順位付け性能の観点では大きな差は見られなかった．識別力パラメータ a を用いた重み付けを行った IRT 統合は，統合手法の中では最も高い値を示したが，単体指標を明確に上回るものではなかった．

5.3 Kendall 順位相関による比較

表 2 に，各評価手法と人手評価との Kendall 順位相関係数を示す．Kendall 相関においても，Spearman 相関と同様の傾向が確認された．特に，局所的な順位の一致度を評価する Kendall 相関では，METEO

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

R が比較的高い値を示しており，単体指標が安定した順位付け性能を持つことが示唆される．IRT による統合手法および単純平均は，単体指標と同程度の相関にとどまった．

表 2　人手評価との Kendall の順位相関係数 (n = 806)

手法	Kendall τ
METEOR	0.364
BERTScore	0.317
COMET	0.307
Perplexity	0.040
単純平均(正規化)	0.342
IRT(GRM)	0.340
IRT(a 重みづけ)	0.361

5.4 Pearson 相関による比較

表 3　人手評価との Pearson の相関係数 (n = 806)

手法	Pearson r
METEOR	0.658
BERTScore	0.739
COMET	0.719
Perplexity	0.215
単純平均(正規化)	0.673
IRT(GRM)	0.551
IRT(a 重みづけ)	0.756

表 3 に，各評価手法と人手評価との Pearson 相関係数を示す．Pearson 相関では，識別力パラメータ a を用いた IRT 統合が最も高い値を示した．これは，各評価指標の人手評価との線形的一致度を考慮した重み付けが有効であることを示している．一方で，IRT モデルが直接推定する潜在能力値 θ は，単純平均よりも低い相関にとどまった．

5.5 多次元 IRT による分析

多次元 IRT を用いて評価指標の構造を分析した結果，意味的評価を主に反映する因子と，流暢さに関連する因子が分離される傾向が確認された．特に，Perplexity は意味的評価因子とは異なる次元に強く寄与しており，人手評価が重視する観点とは異なる側面を捉えていることが示唆された．

6　考察

本研究の結果から，複数の自動評価指標を統合することで人手評価との一致度は一定程度向上するが，順位相関においては単純平均および IRT 統合は単体指標と同程度の性能にとどまることが分かった．このことは，正規化された指標が同一の評価観点を測定している場合，単純平均が強力なベースラインとなることを示している．

一方で，識別力パラメータ a を用いた重み付けでは，Pearson 相関において最も高い値が得られた．この結果は，IRT 解析を通じて信頼性の高い指標を特定し，線形的一致度を高める統合が可能であることを示している．ただし，この手法は IRT モデルが直接出力する潜在能力値 θ とは異なり，IRT 解析結果を利用した後処理的な統合である点に留意が必要である．

また，多次元 IRT の結果から，Perplexity は意味的評価指標とは異なる次元を捉えていることが明らかとなった．これは，Perplexity が物語評価において無効であることを意味するものではなく，人手評価が重視する評価観点とは異なる特性を持つため，順位付けへの寄与が限定的であったと解釈できる．

さらに，単純平均では指標間のスケールを事前に揃える必要があるのに対し，IRT では正規化を前提とせずに指標を統合できる点も重要な特徴である．以上より，IRT は評価指標の役割や特性を分析し，解釈可能な形で統合するための枠組みとして有用であると考えられる．

7　おわりに

本研究では，物語評価における複数自動評価指標の統合手法を比較し，IRT を用いた分析の有効性と限界を明らかにした．相関係数の最大化という観点では単純平均や重み付け統合が有効であった一方，IRT は評価指標の特性や役割を分析的に理解するための枠組みとして有用であることが示された．今後は，人手評価の観点構造をより詳細にモデル化し，評価目的に応じた指標統合手法の設計が課題である．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Papineni, K., Roukos, S., Ward, T., & Zhu, W-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. ACL.FirstName LastName. Title of the article. **Journal of Natural Language Processing**, Vol. 13, No. 1, pp. 251-258, 2006.

[2] Satanjeev Banerjee, Alon Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proceedings of the ACL Workshop on In- trinsic and Extrinsic Evaluation Measures, pp. 65–72, 2005.

[3] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi, BERTScore: Evaluating Text Generation with BERT, arXiv preprint arXiv:1904.09675, 2019.

[4] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for Machine Translation Evaluation. EMNLP.

[5] Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity — a measure of difficulty of speech recognition tasks. The Journal of the Acoustical Society of America.

[6] Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, Hideki Nakayama, StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), pp. 1739‑1753, 2022.

[7] 青見樹,堤瑛美子,宇都雅輝,植野真臣:「項目反応理論を用いた自動採点モデルの統合手法」, 第35 回人工知能学会全国大会論文集,2F3-GS-10g-03,2021 年.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).