複数参照ベース評価における参照集合の評価に向けて

渡辺楓大 [1]　佐藤魁 [2]　赤間怜奈 [2,3,4]　吉野幸一郎 [1,4]
[1] 東京科学大学　[2] 東北大学　[3] 国立国語研究所　[4] 理化学研究所
watanabe.f.c44a@m.isct.ac.jp　kai.satou.r8@dc.tohoku.ac.jp
akama@tohoku.ac.jp　koichiro@c.titech.ac.jp

概要

自然言語生成タスクの有望な自動評価方法として、複数の参照に基づく評価方法が存在する。この評価方法が機能するには、複数参照によって想定される出力候補が網羅される必要があるが、既存の評価データセットがそうした網羅性を持つかどうかは明らかではない。本研究では既存の複数参照ベース評価のデータセットに含まれる事例について、参照集合の意味的な網羅の程度を可視化するための方法論を検討する。

1　はじめに

機械翻訳、自動要約、対話応答生成のような自然言語生成タスクにおける代表的な自動評価として、参照ベース評価が挙げられる。参照ベース評価では出力に対する正解をあらかじめ準備し、システム出力と正解との類似度を何らかの指標によって計測することでシステム出力の良さを測る。参照ベースの評価指標には類似度の定義やタスクに応じて様々なバリエーションが存在し、たとえば機械翻訳では BLEU [1] や METEOR [2]、自動要約では ROUGE [3]、対話応答生成では BERTScore [3] などが用いられる。

参照ベース評価において、自然言語の多様性により望ましい出力が複数存在する場合がある。こうした問題を解消するために考案された枠組みのひとつが複数参照ベース評価で、問題毎に複数の正解からなる参照集合を用意しておくことにより、多様性が想定される出力に対しても評価可能となる。複数参照による評価は単一参照による評価と比較して人手評価との相関が強いことが報告されている [4, 5, 6]。

複数参照ベース評価において適切な性能評価を実現するためには、参照集合が「想定される出力候補を過不足なく網羅していること」が理想的である。もし、事例ごとに必要十分な参照集合の規模や構成を事前にある程度予測できれば、評価データセット構築に伴うコストを最小化しつつ効果的な評価基盤を構築することが可能となる。

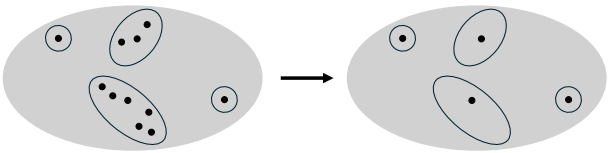

図 1: 類似する参照のクラスタから代表参照を抽出することで網羅性を保った参照集合を構築する

以上の動機から、本研究では、既存の複数参照ベース評価のデータセットに含まれる事例について、参照集合の網羅の程度を定量化するための方法論を提案する。方向性としては、参照集合に対してクラスタリングを行い、類似する網羅性への貢献度が低いサンプルが凝集されることを期待する。具体的には、クラスタリングの閾値に応じて動的に構成されたクラスタの代表点を抽出して新たな参照集合を構築する。その上で、構築した参照集合を用いて複数参照ベース評価を行いスコアの変化を観察することで、元の参照集合の網羅性を特徴づけることを考える。

2　実験仮説

本研究の目的は、複数参照を用いた評価データセットにおいて、用意された参照集合の良さを定量化することにある。参照集合に対する評価観点としては、評価に求められる参照を必要十分に網羅しているかという点があげられる。一般に複数参照ベース評価のデータセットでは、全ての事例で参照集合の大きさが固定されていることが多い。ただし、実際には事例に応じて網羅すべき範囲が異なるため、与えられた参照集合に含まれる一部の参照で十分な評価を行える事例と同時にそうではない事例も存在しうる。例えば、「明日の天気は晴れですか？」というクエリに対する出力としては、最低限「はい」と「いいえ」の二通りの応答が網羅されていれば

よい。これに対し、「明日の天気を教えてください」というクエリに対する出力としては二通りよりも多くの応答が想定される。

本研究では、参照集合の網羅の程度を可視化する方法として、クラスタリングによる手法を検証する。図 1 に示す通り、参照集合に対しクラスタリングを適用すると、同じクラスタに含まれる参照はクラスタリングで用いた距離尺度に応じて類似したものとなる。言い換えれば、構築されたクラスタ内の集合は、少なくともクラスタリングに用いた尺度の観点、例えば意味空間上での距離や単語表層の近さ、においては冗長なサンプルであるという見方ができる。本研究では特に参照文同士の意味的な近さに着目する。参照応答集合に対し意味的クラスタリングを適用し、各クラスタから代表的な参照文を抽出して新たな参照集合を構築することで、元の参照集合よりは小規模ながら意味的な網羅性を保った参照集合を構築する（図 1）。この新たに構築した参照集合を用いて、複数参照ベース評価を行った際のスコアを観察する。この時、参照数の減少に対するスコアの変化によって、元の参照集合の特徴付けが可能であると考えられる。少ない参照数でもスコアを維持できている事例は一部の参照で十分な評価を行える事例であり、逆に参照数の減少とスコアの減少の相関が強い事例は、十分な評価を行うためには多くの参照が必要な事例であると考えられる。

3 実験設定

前章の仮説を検証するために、複数参照ベース評価データセットに基づく分析を行った。

3.1 分析に用いたデータセット

本研究では複数参照ベース評価のデータセットとして SummEval [7] を分析した。SummEval は要約タスクにおけるメタ評価のためのデータセットであり、100 件のニュース記事それぞれに対して 11 個の人手による参照要約と 16 個の機械生成要約が付けられている。今回はその中でも前半の 20 件を分析対象とした。

3.2 手順

クラスタリングによって新たな参照集合を構築する手順を示す。各記事に付属している 11 個の人手参照要約を意味的埋め込みに変換し、クラスタリングを適用して意味的に類似する参照要約を凝集させた。埋め込み表現には OpenAI 社が提供している text-embedding-3-large を使用した。クラスタリングアルゴリズムは DBSCAN [8] を使用した。DBSCAN はクラスタ数を指定するパラメータは存在しないため、閾値パラメータ ϵ を十分に細かい刻み幅で複数設定し、各設定に対してクラスタリングを実行することで、1 から 11 までの全てのクラスタ数のクラスタリングが形成される時点を評価の対象とした。

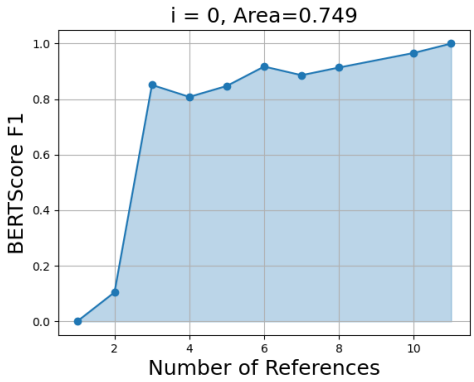

図 2: 参照集合の大きさを変化させた時の BERTScore の F1 値（正規化後）の変化の様子

DBSCAN の距離行列は次のように定義した。二つの参照 r_1 と r_2 がそれぞれ n 個、m 個の文から構成されているとする。

$$r_1 = (x_1, x_2, \cdots, x_m), r_2 = (y_1, y_2, \cdots, y_n)$$

参照 r_1 と r_2 の距離行列は以下の距離の定義に基づいて構築した。

$$D(r_1, r_2) = \frac{1}{2m}\sum_{i=1}^{m}\min_{y\in r_2}(1 - x_i \cdot y) + \frac{1}{2n}\sum_{i=1}^{n}\min_{x\in r_1}(1 - x \cdot y_i)$$

新たな参照集合は、構築された各クラスタからランダムに一つずつ参照を選択することで構築した。DBSCAN で外れ値と判定された参照は、要素が一つのクラスタであるとみなし、必ず選択するものとした。ランダム選択に伴う偶然性や選択の偏りを軽減するため、同一のクラスタリング結果に対してこの操作を 100 回繰り返し、それぞれ異なる 100 通りの参照集合を構築した。

3.3 評価指標

前節の処理の結果、記事毎に参照数 $i(1 \le i \le 11)$ の新たな参照集合が 100 通り構築される。各参照集合を使用して、データセットに付属する 16 個の機械生成要約に対する BERTScore [9] を計算し、F1 値の平均値を scores$[i]$ とした。この結果を図示すると

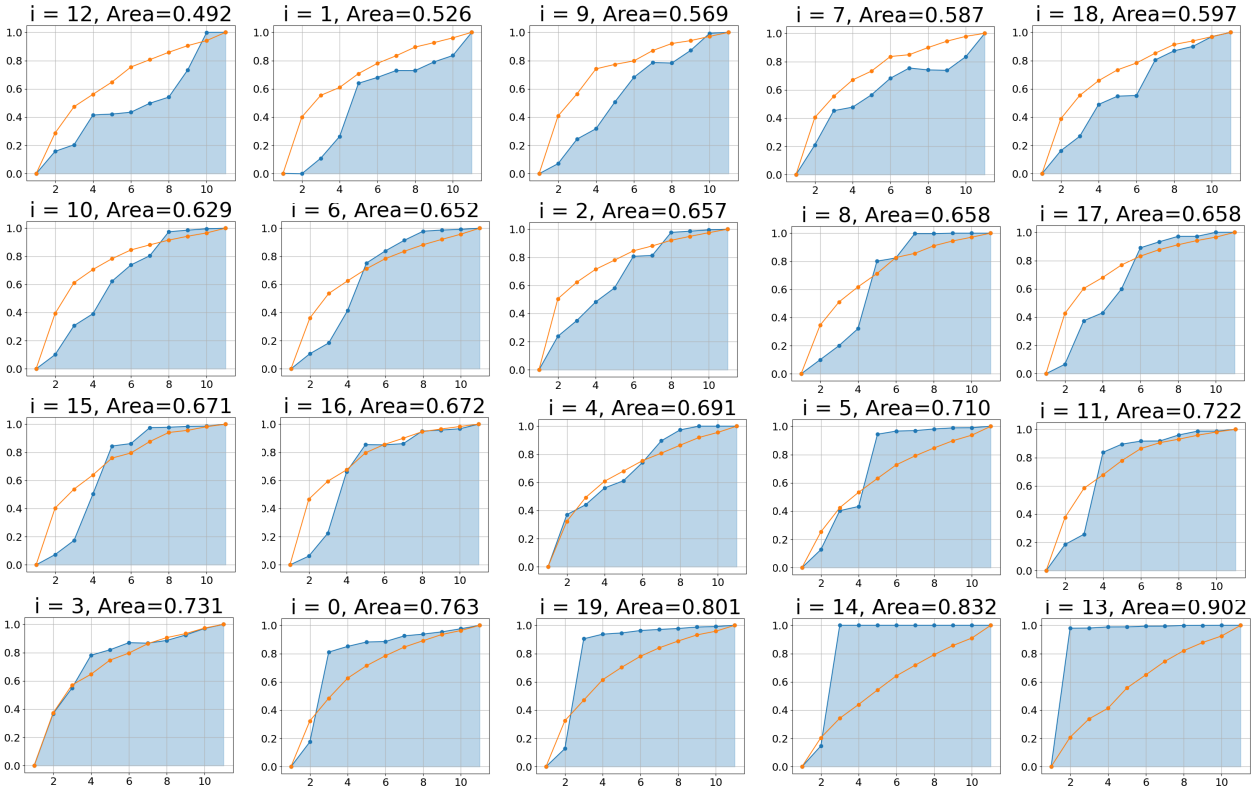

図 3: 参照数を変化させた時のスコア推移（青線：クラスタからの選択によって構築した参照集合によるスコア推移、橙線：ランダム選択によって構築した参照集合によるスコア推移）

図 2 のような図が得られる。横軸は参照集合が持つ参照の数であり、基本的にはこれが大きくなるに従い縦軸の BERTScore も高くなる。ただし、クラスタリングの過程である点において十分な数の参照集合が得られた場合、参照数を増やしても BERTScore が高くならないことが考えられる。

上記のような参照数と参照ベース評価スコアの関係を可視化するため、以下の式で定義される AREA という指標を定義した。

$$\text{AREA} = \frac{1}{11} \sum_{i=1}^{11} \text{norm}(\text{scores}[i])$$

これは図 2 の青枠部分の面積に対応する。AREA が高い事例は、参照数の減少に対してスコアの減少が緩やかである、すなわち少ない参照数で評価に必要な参照を網羅できている可能性が高いことを意味する。

3.4 参照のランダム選択

クラスタリングによる参照選択との比較実験として、参照をランダム選択した実験も行った。ここでは元の参照集合から指定した数の参照をランダムに選択する操作を 100 回行い、各参照集合を使って評価して得られる BERTScore の F1 値の平均値の変化を計算した。

4 結果と分析

図 3 は実験により得られた全 20 事例におけるスコア変化を示す（左上から右下にかけて AREA スコア昇順）。図中の青線はクラスタリングによって構築した参照集合で評価した時のスコア変化、橙線はランダム選択によって構築した参照集合で評価した時のスコア変化を表す。ランダム選択による参照集合スコア変化は事例間で大きな差が見られず、参照数の増加に伴ってスコアが上昇する傾向を示した。一方で、クラスタリングによるスコア変化、特に AREA が高い事例では一定の参照数でスコアが頭打ちになる様子が確認できる。この頭打ちとなる結果は、クラスタリングによって同じクラスタに含まれる参照を追加しても冗長なサンプルとなるという仮説を支持するものである。

4.1 参照の寄与度の分析

最も高い AREA が得られた事例 (i=13) のクラスタリング結果とスコアの変化について分析する。表 1

表 1: AREA 最大事例 (i=13) のクラスタリング結果

参照数	クラスタリング結果
11	[0], [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]
10	[6, 8], [0], [1], [2], [3], [4], [5], [7], [9], [10]
9	[6, 8], [7, 10], [0], [1], [2], [3], [4], [5], [9]
8	[6, 7, 8, 10], [0], [1], [2], [3], [4], [5], [9]
7	[6, 7, 8, 9, 10], [0], [1], [2], [3], [4], [5]
6	[1, 4], [6, 7, 8, 9, 10], [0], [2], [3], [5]
5	[1, 4], [5, 6, 7, 8, 9, 10], [0], [2], [3]
4	[1, 4, 5, 6, 7, 8, 9, 10], [0], [2], [3]
3	[1, 3, 4, 5, 6, 7, 8, 9, 10], [0], [2]
2	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10], [0]
1	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

表 2: AREA 最小事例 (i=12) のクラスタリング結果

参照数	クラスタリング結果
11	[0], [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]
10	[2, 4], [0], [1], [3], [5], [6], [7], [8], [9], [10]
9	[0, 9], [2, 4], [1], [3], [5], [6], [7], [8], [10]
8	[0, 2, 4, 9], [1], [3], [5], [6], [7], [8], [10]
7	[0, 2, 4, 9, 10], [1], [3], [5], [6], [7], [8]
6	[0, 2, 4, 7, 9, 10], [1], [3], [5], [6], [8]
5	[0, 1, 2, 4, 7, 9, 10], [3], [5], [6], [8]
4	[0, 1, 2, 4, 6, 7, 9, 10], [3], [5], [8]
3	[0, 1, 2, 4, 6, 7, 8, 9, 10], [3], [5]
2	[0, 1, 2, 3, 4, 6, 7, 8, 9, 10], [5]
1	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

表 3: 事例 (i=13) における各参照の寄与度

参照番号	0	1	2	3	4	5	6	7	8	9	10
寄与度	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

表 4: 事例 (i=12) における各参照の寄与度

参照番号	0	1	2	3	4	5	6	7	8	9	10
寄与度	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

は事例 (i=13) のクラスタリング結果である。このクラスタリング結果と図 3 のスコア変化を照らし合わせると、特定の参照が選択されている間はスコアが維持され、選択されなくなるとスコアが下がることが確認された。このことから BERTScore の計算に強く寄与する参照の存在が示唆される。

この寄与度の定量化として、各参照について自身を除いた 10 参照を用いた BERTScore の F1 値と、全参照を使った BERTScore の F1 値の差分を計算したところ、特定の参照の寄与度が著しく高い結果となった（表 3）。また同様に、AREA が最も低い事例（i=12）についてもクラスタリング結果および寄与度を計算すると、特定の参照の寄与度が高く、寄与度が高い参照が大きいクラスタに含まれていた（表 2 表 4）。

これらの結果から示唆される可能性は 2 通りあり、仮説通り本質的に参照数が不足しており十分に網羅が行われていない事例のほかに、寄与度が高い参照が参照されず AREA のスコアに影響が出ている可能性がある。

4.2 評価対象文書の多様性

また、こうした一部の事例において参照集合中の特定の参照が BERTScore に強くする現象の原因として、SummEval に含まれる機械要約の質の問題もある可能性がある。つまり、評価対象文書の多くが特定の参照に類似していることで、少数の参照の評価スコアへの寄与が過度に大きくなった可能性である。SummEval に付与された機械要約は大規模言語モデルが文生成の主流となる以前に付与されたものであり必ずしも高い質のものばかりではない。こうした評価対象となる機械要約の質と多様性についても今後検討を行う必要がある。

5 おわりに

本研究では、既存の複数参照ベース評価のデータセットに含まれる事例について、参照集合の網羅の程度を特徴づける手法を提案した。提案手法は、参照集合の網羅性の可視化を実現し、実験仮説を裏付けるような可視化結果が得られた。しかし詳細な分析を行ったところ、実験仮説が成立すると言うためにはさらなる分析が必要なことが示唆された。

今回の実験では、評価対象文書として SummEval に付属する機械生成要約に限定して実験を行ったが、これらの文書が正解出力の分布を十分に網羅しているとは限らず、評価対象側の分布に偏りが存在する可能性が残されている。今後は、大規模言語モデル等を用いて多様な評価対象文書を合成し、評価スコアの変動を分析することで、参照集合に起因する要因と評価対象に起因する要因を分離した評価を行う予定である。さらに、クラスタリングにおける閾値の決定方法を検討するとともに、その結果として得られる参照集合自体の網羅性をどのように評価すべきかについても議論していく。

謝辞

本研究は JST さきがけ（JPMJPR24TC）と JST BOOST（JPMJBY24A1）の支援を受けた。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Tianyi Tang, Hongyuan Lu, Yuchen Jiang, Haoyang Huang, Dongdong Zhang, Xin Zhao, Tom Kocmi, and Furu Wei. Not all metrics are guilty: Improving NLG evaluation by diversifying references. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 6596–6610, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[5] Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 918–932, Online, November 2020. Association for Computational Linguistics.

[6] Xianfeng Zeng, Yijin Liu, Fandong Meng, and Jie Zhou. Towards multiple references era – addressing data leakage and limited reference diversity in machine translation evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 11939–11951, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 391–409, 2021.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**, KDD'96, p. 226–231. AAAI Press, 1996.

[9] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.