QA タスク回答中の趣旨・補足に対する不適・不足の分析

土田陸斗[1] 飯田頌平[2] 高橋空大[1] 三田寺聖[1] 長谷川遼[1]
謝宇程[1] 宇津呂武仁[1] 林友超[2] 宍戸里絵[2]
[1]筑波大学大学院 システム情報工学研究群
[2]弥生株式会社 AI・データ戦略部 R&D チーム
{s2520796,s2520785,s2420810,s2420791,s2565111}_@u.tsukuba.ac.jp
utsuro_@iit.tsukuba.ac.jp, {shohei_iida,lin_youchao,rie_shishido}_@yayoi-kk.co.jp

概要

本論文では質問応答タスクにおいて，参照回答とLLM により出力される回答との一致度を測る際，出力回答中の「趣旨」と「補足」の情報を区別し，それぞれに対し「不適」「不足」まで考慮した新たな指標を提案している．ここで提案する評価指標は，単なる文章の類似度だけでは考慮できない要素を含んだ評価指標となる．また，そのような基準に基づいた出力回答の分類や実際に人手で分析した結果を最終的に示し，基準の整合性を確認している．

1 はじめに

本論文では，質問応答タスクにおいて，参照回答(正解データ) と大規模言語モデルの出力回答を比較し，「趣旨」「補足」情報ごとの「不適」・「不足」を計測する新たな評価指標を提案している．

「趣旨」「補足」や「不適」「不足」の細かい定義は 3 章や 4.1 節で述べているが，「趣旨」「補足」で区別して出力回答を確認し，参照回答との「不適」「不足」を検出することで，質問応答タスクにおける性能を定量的に評価することが可能となり、単純な文章の類似度だけでは考慮できない要素を含んだ評価指標となる．

4.3 節では実際にこの指標を使い，会計ドメインを使用した質問応答タスクの結果を人手で評価している．出力回答の内、過半数以上が少なからず「不適」「不足」を含んでおり，この部分の原因を分析し改良することで性能向上の余地を確認でき，質問応答タスクの性能向上に向けた分析を行うという点で有効な指標であると言える．

2 関連研究

本論文では質問応答タスクにおける評価の際，参照回答と出力回答を比較しその一致度を測る新たな指標を提案している．既在の指標で，2 つの文の一致度を測る指標として，ROUGE スコアがある．ROUGE スコアの前身として，要約評価のための n-gram ベース手法 [3] があり，その後単語の共起を評価する ROUGE-1 [2] や，最長共通部分列（LCS）ベースで共起している単語の個数で評価する ROUGE-L [4] などが登場している．

その他の指標では，参照文と出力文の一連の単語の並び（n-gram）一致率を計算し、類似度を 0〜1 のスコアで表現する BLUE スコア [6] があり，スコアが高い（1 に近い）ほど参照文に近い出力とみなされ、翻訳や形式が定まったテキストの生成品質を測定できる。

また，文の意味を重視した評価指標として，エンコーダによって参照回答と出力回答から評価値を得る BERT スコア [10] や COMET スコア [7] などが提案されている．

一方，文献 [1] や文献 [5] では長年使用されてきた ROUGE が本当に十分な指標であるのか再検証し，ROUGE ではとらえられない重要な観点を測るより良い評価方法が必要だと示している．

また，本論文のように新たな指標を作成し，提案している関連論文 [9] では 2 つの文書が同じ事実を保持しているかを確かめるために，同じ質問をして同じ答えが返ってくるかという観点で評価する QAGS という指標を提案している．他にも文献 [8] などで従来の評価指標では測れない部分まで考慮した新たな評価指標が提案されている．

しかし、これまで紹介した評価指標では，参照回答や出力回答中の構造的な重要性（趣旨・補足）

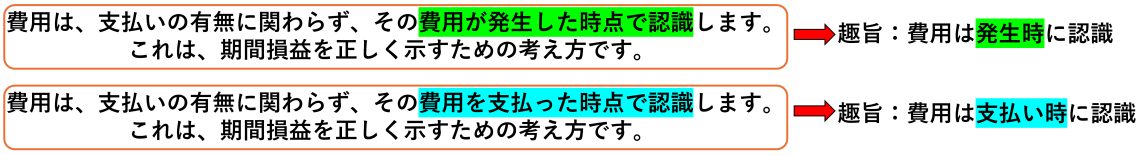

図 1　2 つの文で趣旨情報が異なる例

や、「不適」「不足」といった方向性を扱うことができず，本論文はそれらを考慮しているという点でこれまでの指標とは大きく異なっている。

3　質問応答タスクの回答中に存在する趣旨・補足情報

質問応答タスクでは，参照回答という正解の文にできるだけ出力回答を近づけることを目指している．そこで，質問応答タスクの評価では参照回答と LLM の出力回答を比較し，その類似度を計算する必要がある．しかし，2 つの文の類似度を計算するだけでは完全に考慮できていない要素があると本論文では考えている．それが文の「趣旨」と「補足」の情報である．

例えば，図 1 では文字だけを見ると大半の部分は参照回答と出力回答が一致しているように思えるが，文の内容まで考慮したときに，参照回答中で重要な情報である「趣旨」情報が出力回答中には入っていないことが分かる．このように，参照回答中の重要な情報が異なる場合と，単なる「補足」情報が異なる場合では，質問応答タスクにおける性能の評価は区別する必要がある．

そこで，本論文では，質問応答の参照回答には，重要な情報である「趣旨」とそれを補足する情報である「補足」の 2 つが存在していると考え，出力回答と参照回答との一致度を考える際に，「趣旨」と「補足」を考慮した評価指標を提案する．

また，本論文で述べている「趣旨」情報というのは参照回答や出力回答中で、質問に対する答えとして特に重要な情報である．一方、「補足」情報というのは趣旨を補強する情報、具体例、背景説明など、重要度が低い情報である。

4　出力回答中の不適・不足の分析

4.1　概要

図 2 の上部で示すように，出力回答中には参照回答に対して誤った情報が加えられていたり，正しい情報が追加されていることがある．また，逆に参照回答と比較して不足している部分が存在している．そのような部分について，参照回答より情報が多い部分は「不適」，参照回答より情報が少ない場合は「不足」と定義する．本論文では，「不適」についてはさらに細かく情報が誤りである場合と，正しい情報が過剰に足されている場合の 2 つが考えられるがここではまとめて「不適」とし，情報の正誤は考えていない．ここで定義した「不適」「不足」について，本論文では 3 章で述べた参照回答中の趣旨・補足のそれぞれの情報に対して，その不適度や不足度を細かく数値化した指標を提案する．さらに，その指標に基づいて不適部分や不足部分を細かく分析することで，質問応答タスクでの性能を向上させることが可能となる．

実際の出力回答を用いた不適・不足の具体的な例を図 3 に示す．

図 3 中では，確定申告の方法を聞く質問に対する参照回答に 2 つの重要な情報，「趣旨」が含まれている．1 つ目は確定申告では 1 年間の所得と所得税を計算し，税務署に報告することで，2 つ目は確定申告の方法は青色申告と白色申告があるということである．そのような参照回答に対し，青色申告と白色申告に関する記載が無い出力回答は LLM の回答が不十分，つまり「不足」となる．加えて，不足している情報は参照回答中の「趣旨」であるため「趣旨」が「不足」となる．一方，申告にはマイナンバーカードが必要であること（趣旨）や，マイナンバーカードの申請方法（補足）まで書かれている回答は参照回答と比較して過剰であり，「趣旨」と「補足」の両方が「不適」となる．よって，LLM の出力回答には以下の 4 つの特性があると考えられる．

- 「趣旨」が「不適」
- 「趣旨」が「不足」
- 「補足」が「不適」
- 「補足」が「不足」

本論文では，全ての出力回答にはこれらの特性があると考え，4 つの特性それぞれに特性の大きさを表す数値を割り当てる．4.2 節で上記の 4 つの特性を考慮した評価指標を作成し， 4.3 節で分析手順と結果を示す．

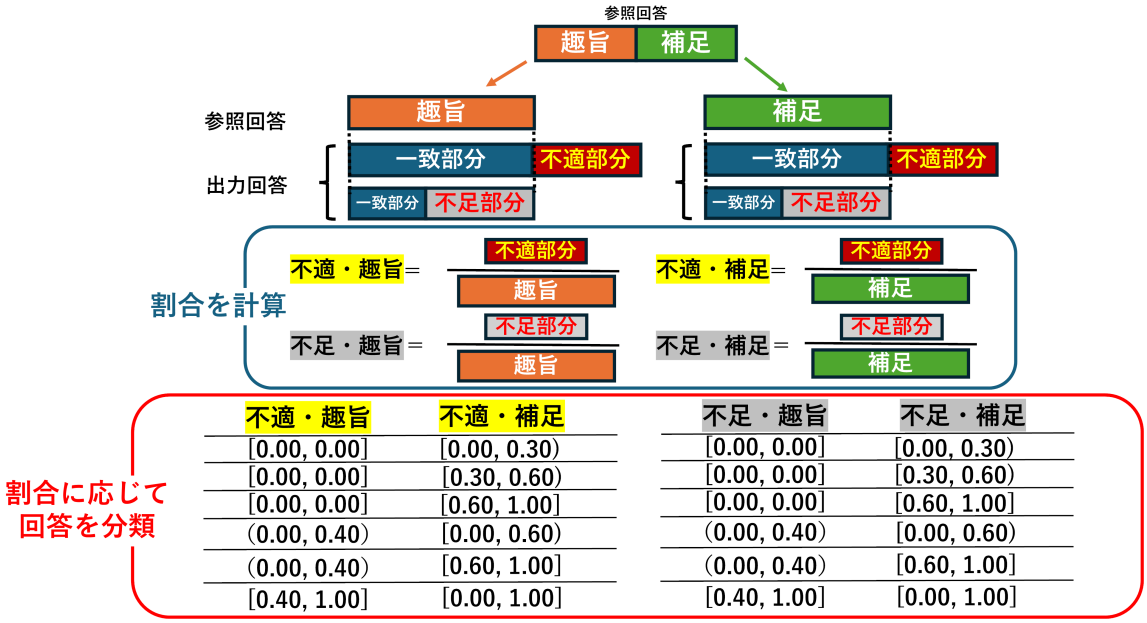

図 2　不適・趣旨、不適・補足、不足・趣旨、不足補足の定義と出力回答の分類条件

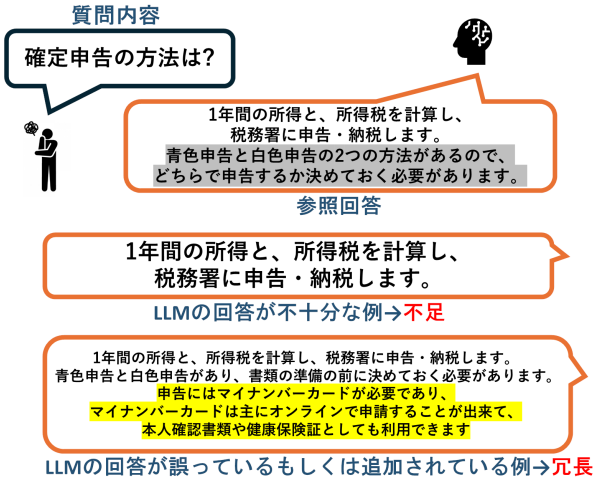

図 3　質問応答タスクにおける LLM の出力回答の不適・不足の例

4.2　出力回答中の不適・不足の評価指標

4.1 説では出力回答には 4 つの特性があることを述べた．本節ではその 4 つの特性を考慮した評価指標を提案する．

それぞれの特性には 4.3 説で述べられている手順によって，数値が計算される．そこで割り当てられる数値は値が高いほどその特性が大きいことを示し，1 つの出力回答に対し 4 つの数字が計算される．この時，「不適」の部分が全く無い出力回答であってもその出力回答は「不適」の特性が 0 であると考えることで，すべての出力回答が 4.1 節で述べている特性を持っているとすることが出来る．

続いて，4 つの特性の数値によって出力回答を分類する．分類の軸は「不適」と「不足」の 2 軸であり，図 2 に示すように，「不適」と「不足」それぞれに 6 つの項目があり，表 2 のように 6 × 6 の 36 種に分類される．

表 1　「不適」「不足」の有無のみを考慮した分析結果

	不適なし	不適あり	計
不足なし	12	23	35
不足あり	14	1	15
計	26	24	50

図 2 中の 6 つの分類項目では，「趣旨」の「不適」「不足」が「補足」の「不適」「不足」よりも重大だという考えに基づいている．よって，「不適・趣旨」や「不足・趣旨」の割合が 0.00 である時，LLM の出力回答に「趣旨」に関する大きな「不適」「不足」は無いとして，「補足」の細かい誤りで 3 つに分類している．続いて，「趣旨」の「不適」「不足」の割合が 0.00 より大きく 0.40 より小さい場合について，「補足」の割合でさらに 2 つに分類している．つまり，「趣旨」と「補足」それぞれに少し「不適」「不足」がある出力回答と，「趣旨」のわずかな「不適」「不足」と「補足」の大きな「不適」「不足」がある出力回答を区別している．最後に，「趣旨」の「不適」「不足」が 4 割以上ある場合，質問応答タスクの回答としてすでに致命的であるため，「補足」に関する条件は無く，0.00 から 1.00 までの全範囲としている．

表 2　「趣旨」「補足」情報を考慮した出力回答の「不適」「不足」分析結果

			不適						計
			趣旨∈ [0.00, 0.00]			趣旨∈ (0.00, 0.40)		趣旨∈ [0.40, 1.00]	
	趣旨∈	補足∈	補足∈ [0.00, 0.30)	補足∈ [0.30, 0.60)	補足∈ [0.60, 1.00]	補足∈ [0.00, 0.60)	補足∈ [0.60, 1.00]	補足∈ [0.00, 1.00]	
不足	[0.00, 0.00]	[0.00, 0.30)	13	3	13	1	0	3	33
		[0.30, 0.60)	5	0	0	0	0	0	5
		[0.60, 1.00]	8	0	0	0	0	0	8
	(0.00, 0.40)	[0.00, 0.60)	1	0	0	0	0	0	1
		[0.60, 1.00]	0	0	0	0	0	0	0
	[0.40, 1.00]	[0.00, 1.00]	0	0	0	1	0	2	3
計			27	3	13	2	0	5	50

また，本論文で使用したデータでは「不適」と「不足」が混在している出力回答はわずかであったが，質問応答タスクにおいての性能向上へ向けた分析を行う際に有効であると考え 36 種まで細分化している．

4.3　分析手順・結果

本節では 4.2 節で数値化された値に応じた出力回答の分類まで行う実際の手順を示す．

初めに，参照回答について「趣旨」「補足」情報を分析する．参照回答の内容を確認し，質問の答えとして特に重要である情報を「趣旨」、「趣旨」を補足する情報や具体例，そこまで重要ではない情報を「補足」と分類し，その情報の個数を計測する．ここで述べている個数というのは文単位や句読点で句切っているものではなく，意味単位で計測する個数である．図 3 の例では参照回答には「趣旨」が 2，「補足」が 0 となる．

次に，出力回答と参照回答との不適度、不足度を計測する．これらは先ほど分析した参照回答の「趣旨」「補足」ごとに行われる．図 3 の例では，4.1 節でも述べているように 1 つ目の出力回答は「趣旨」が 1 つ「不足」しており，2 つ目の出力回答では「趣旨」が 1 つ「不適」かつ「補足」が 1 つ「不適」となっている．この「不適」「不足」の個数を参照回答に対する割合に変換する．つまり，図 3 の場合，参照回答の「趣旨」が 2 つ、「補足」が 0 であるのに対し，1 つ目の出力回答は「趣旨」情報の「不足」が 1 つあるため，「趣旨」情報の「不足」が 1 / 2 = 0.50 となる．2 つ目の出力回答は「趣旨」と「補足」の「不適」が 1 つずつあるため，「趣旨」情報の「不適」が 1 / 2 = 0.50，「補足」情報の「不適」が 1 / 0.1 = 10 となる．ここで，分母が 0 になる場合のみ 0 は 0.1 として計算している．

以上の手順で 4 つの特性に対し数値を計算し，各出力回答を 4.2 節で述べているように分類する．

分析結果について，まず「不適」「不足」の有無だけで分類した結果を表 1 に示す．今回使用するデータは会計ドメインの質問応答タスクにおける LLM の出力回答 50 件を対象としている．表 1 を見ると，今回の使用した 50 件の内 38 件が少なからず「不適」「不足」の部分を持っていることが分かる．従来の質問応答タスクの評価では，参照回答と差異が無い 12 件と差異がある 38 件という 2 値で分類することが一般的であるが，この 38 件の誤りを細かく分析し，性能向上に役立っているという意味でも本論文で提案する指標が有効であると確認することが出来る．

実際に，参照回答とは完全に一致していない 38 件の出力回答について分析した結果を表 2 に示す．表 2 を見ると，「不適」と「不足」が混在する出力回答は非常に少ないことが分かる．これは今回使用したデータが少数であることや，質の悪い出力回答がほぼ存在しなかったためであるが，対象のデータによっては表 2 のすべてのマスが埋まることが想定され，有効な分析を行うことが出来る．

5　おわりに

本論文では質問のタスクにおいて，参照回答や出力回答の「趣旨」と「補足」情報を考慮し，「不適」「不足」まで詳細に分析する新たな評価指標を提案した．また 4.3 節では，参照回答と出力回答の差異の有無のみで比較する一般的な評価方法に比べ，本論文で提案する指標が優れている点を示している．よって，本論文で提案した評価方法を適用することで LLM の出力回答の誤りを細かく分析することが可能になり，性能向上へ多く貢献すると考えている．

謝辞

本研究は筑波大学と弥生株式会社の共同研究として実施した.

参考文献

[1] Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In **Findings of ACL**, pp. 1547–1560, 2022.

[2] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[3] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. 2003.

[4] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. 2004.

[5] Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In **Findings of ACL**, 2025.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of ACL**, July 2002.

[7] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of EMNLP**, pp. 2685–2702, 2020.

[8] Shuqian Sheng, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, and Chenghu Zhou. RepEval: Effective text evaluation with LLM representation. In **Proceedings of EMNLP**, November 2024.

[9] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In **Findings of ACL**, p. 5008–5020, 2020.

[10] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with BERT. In **ICLR Workshops**, 2020. Preprint available at arXiv:1904.09675.