読解問題における論理推論の一貫性評価

川畑 輝 [1]　菅原朔 [2]
[1] 奈良先端科学技術大学院大学　[2] 国立情報学研究所
kawabata.akira.kb3@is.naist.jp　saku@nii.ac.jp

概要

複雑な論理推論に関する自然言語理解モデルの評価が信頼可能であるためには、予測結果の正しさだけでなく、その予測のための妥当な推論根拠への理解も評価することが重要である。しかしこれまでの自然言語理解タスクでは推論結果と推論根拠の一貫的な評価はなされてこなかった。そこで本稿では、既存の選択式の機械読解データセットに対して、その各選択肢が正答または誤答である理由を問う問題を同じく選択式の機械読解データセットとして作成することで、自然言語理解モデルの一貫的な理解評価に取り組んだ。両問題への人間正答率との比較から、現状の自然言語理解モデルには根拠に一貫的に解答することが困難であることがわかった。また、推論タイプと特徴量寄与度の観点から根拠理解の難易度について分析し、寄与度の類似や推論の質的な差異の難易度との関連性を明らかにした。

1　はじめに

自然言語理解モデルが人間のように信頼可能であるためには、予測や生成の正しさだけではなく、その結果が論理的に妥当な根拠を踏まえたものである必要がある。妥当な前提への理解が重要なタスクとして、近年では機械読解においてモデルの論理的推論能力を評価するデータセットが提案されている。ReClor [1] は論理的推論を問うデータセットの代表例であり、GMAT、LSAT から収集された 4 択式の機械読解データセットである。事前学習言語モデル [2, 3, 4] の発展により人間と同程度の精度が報告されているなど、現状の自然言語理解モデルは論理的推論を理解する能力を有するように見える。

しかし既存研究はそうした事前学習言語モデルに共通する弱点として、一貫的な予測の不得手を指摘している [5, 6, 7, 8]。具体的には、元の問題に否定 [8] や置換 [6] などの簡単な操作を施すことで、元の問題と比べ精度が大きく低下したことが報告されている。こうした研究から、言語理解能力評価における一貫性検証の重要性が示唆されている。

論理的推論は妥当な根拠への理解を前提として行われるべきであるが、その規範的要請をモデルが満たしているかどうかは自明ではない。それゆえ論理的推論能力の検証では、推論結果の正誤だけではなく、その根拠の理解も一貫的に評価するタスク設計が望ましい。しかし一方で、そうした根拠はしばしば問題中に明示されていないため、既存研究のように元の問題の簡単な変形を通して一貫性評価のためのデータセットを構築することはできない。そのため、論理的推論の一貫性を検証するためには人手によって元の質問に暗黙的に含まれる根拠を明示化し、収集する必要がある。人手で推論根拠を収集した研究も存在するが、それらは読解における論理的推論は対象としておらず、一貫性を評価できるタスク設計にもなっていない [9, 10]。

そこで本稿では論理的な推論に焦点を当てた一貫性評価データセット RULE (Rationale Understanding for Logical reasoning Evaluation) を構築し、論理的一貫性の評価実験を行なう。RULE は ReClor の問題 4,638 問のうち 600 問を対象に構築された、元の ReClor の問題の正誤根拠理解を問う 4 択機械読解データセットである (図 1)。RULE に含まれる根拠理解問題は ReClor の問題の選択肢それぞれについて作成され、その選択肢の正誤理由が正解選択肢、他の 3 つの選択肢の正誤理由が不正解選択肢となっている。600 問それぞれの選択肢について根拠理解問題を計 2,400 問作成し、後述する品質評価によって最終的に 1,406 問を得た。

実験では、システムの一貫的な論理推論理解の程度を評価するために ReClor で学習したベースラインモデルを RULE で評価した。その結果人間とモデルの精度には大きな乖離が見られた。さらに後続の分析でモデルにとっての一貫的な根拠理解の難易度に影響を及ぼす要素を調査したところ、特に誤答選択肢の根拠理解に改善の余地が大きく残されている

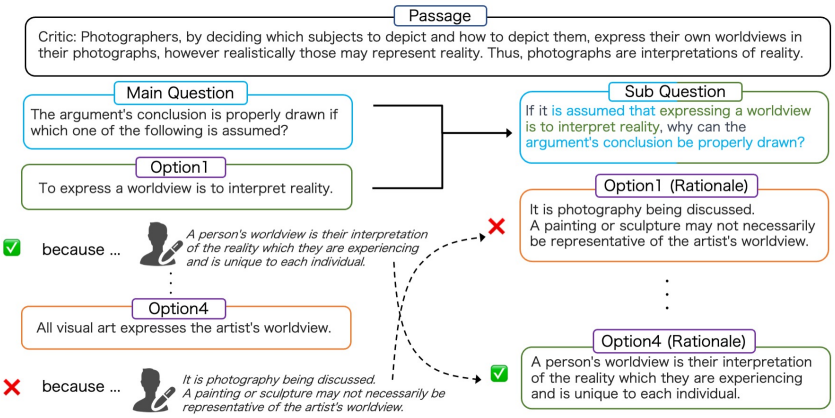

図 1　根拠理解問題の構築イメージ

ことが明らかになった。また、特徴量寄与分布と問題で問われる推論の質的な差異がモデルの根拠理解と関連していることも示唆された。

2　関連研究

本研究と同じように自然言語処理モデルによる推論の意味的な一貫性を変形問題で評価する試みはすでに存在する [5, 6, 7, 8]。たとえば Elazar ら [6] は元の問題をパラフレーズしたデータセットを、Ravichandar ら [8] は元の問題の否定推論箇所を編集することでデータセットを構築し、それぞれの言語理解能力における一貫性を検証している。しかしこれらの既存研究では論理的根拠に焦点を当てた一貫性の評価は行っていない。

他方で、根拠や説明データ自体についての分析やタスク定義はすでに複数なされている [9, 11, 12, 13]。Aggarwal ら [9] は常識推論の QA データセット CommonsenseQA [14] に 4 つの選択肢の正誤理由をアノテートしたデータセット ECQA を構築し、根拠生成と根拠検索タスクを定義した。Sun ら [13] は ECQA に含まれる根拠文を元の問題に連結してモデルに解答させ、問題への解答に根拠文が果たす役割を分析した。これらの研究は根拠データに注目しているものの、本研究と異なり元の問題で獲得された言語能力の一貫的な汎化能力を検証するためにタスクや分析が設計されているわけではない。

3　RULE: データセット構築

本研究では論理的推論の一貫的な根拠理解を評価するデータセット RULE を構築した。RULE に含まれる問題 (以降、**根拠問題**) は ReClor の訓練セットからランダムに選ばれた 600 問 (以降、**主問題**) から作成され、同一の文脈文で各主問題の選択肢の正誤理由を問う。ReClor で学習したモデルを評価するために、根拠問題は主問題と同様に文脈文、質問文、4 つの選択肢から構成される選択式の機械読解タスクとした。3.1 節では選択肢 (根拠文) の収集方法を、3.2 節では質問文の収集方法を説明する。

3.1　正答・誤答根拠の収集

参加者の選定　根拠文を収集するにあたって二段階の資格テストによって事前に参加者を選抜した。まず一段階目の選抜として、応募者に ReClor の問題 10 問を解かせ、その正答率が 8 割以上の場合に合格とした。次に二段階目の選抜として、応募者に ReClor の問題 1 問を提示し、その問題の各選択肢について正誤根拠の執筆を課した。提出された根拠文について著者らが (1) 十分具体的であるか、(2) 問題の意図と合致しているか という二つの基準で評価し、最終的なタスク参加者を決定した。この資格テストは Amazon Mechanical Turk を通じて行われた。最終的な通過人数は 57 人である。

根拠執筆タスク　以上の資格テストを通過した参加者に根拠文の執筆タスクを依頼し、50 人が執筆タスクに参加した。根拠文執筆タスクでは各回ごとに参加者に ReClor の問題が 1 問、各選択肢の正誤と共に提示され、参加者は各選択肢の正誤理由を執筆する。今回の研究では ReClor 600 問に対して正誤根拠を収集し、計 2,400 個の根拠文が集まった。

品質評価　低品質な根拠文を除くために、具体性の観点から評価を行い選別した。具体的でない根拠文　(e.g., “Because this is not mentioned in the passage.”) は 4 択問題の選択肢として機能しないだけでなく、問題に意図されている暗黙的な推論を明示化するという RULE の趣旨から考えても望ましくない。そのため資格テストを通過した参加者に、根拠執筆

者と評価者が重複しない形で評価タスクを依頼した。評価タスクでは参加者は ReClor の問題 1 問と、その問題の選択肢のどれかに対応する根拠文を提示され、根拠文と対応する選択肢を答える。根拠文が具体的でなければ選択肢との対応づけは困難であると考えられるため、この方法で具体性を伴った根拠文を選別した。その結果 1,860 個の根拠文と 475 個の主問題が残った。なお、根拠問題の選択肢は 4 つの根拠文から構成されるため具体性評価によって根拠文が 1 つでも除かれるとその主問題では根拠問題を構成できないが、根拠問題の数を確保するために 4 つ中 3 つ通過した場合は残り 1 つの根拠文を “None of the above choices” に置き換えることで根拠問題を構成した。

3.2 質問文の生成

具体性評価テストに通過した根拠文それぞれについて、主問題の質問文と選択肢から、その選択肢の正誤の根拠を問う質問文（根拠質問文）を生成した。生成には GPT-3 (text-davinci-003) を使用し、主問題の選択肢と質問文を入力としてその選択肢の正誤理由を問う質問文を出力とするようなプロンプトを用いた。また、根拠質問文に選択肢の正誤を反映させるために、正答選択肢と誤答選択肢で使用するプロンプトを分け、誤答のプロンプトには生成対象の根拠質問文に “not” などの否定表現を入れることで誤りの理由を問う根拠質問文を生成させた。生成例として、否定プロンプトを入力した GPT-3 に質問文 “What mistake does the argument commit in its reasoning?” と 選択肢 “It confuses probability and certainty.” を与えると、“What evidence is there that the argument does not commit the mistake of confusing probability and certainty?” という根拠問題用の質問文が生成される。

品質評価　収集された根拠文と根拠質問文から根拠問題を構成した。人間の正答可能性が保証された根拠問題を選別するために、最終的な品質評価として人間の正答率を測定した。資格テストを通過した 57 人に対して根拠問題の解答を依頼し、それぞれの問題について 3 人分の解答結果を収集した。3 人中 2 人以上が正解または 3 人全員が “None of the above choices” を選択した問題 1,406 問を解答可能な問題とみなし、評価問題として採用した。

主問題として使用した ReClor の問題と RULE は、文脈文の数はどちらも 467、質問数はそれぞれ 467 と 1,406、質問文の平均の長さ (単語数) はそれぞれ 13.4 と 29.6、選択肢の平均の長さは 21.5 と 16.7 であった。ReClor と比べて RULE の質問文は平均的に長いが、これは RULE の質問文が主問題の質問文と選択肢の内容を含むためである。

4 評価実験

4.1 実験設定

評価には DeBERTa-V3-Large [4, 15] と GPT-3 (text-davinci-003) [16] を用いた。紙面の都合上、性能の高かった GPT-3 の結果を報告する。両者の比較は付録 A に示す。モデルの根拠理解問題への一貫的な汎化能力を評価するため、評価データには今回作成した RULE 1,406 問と主問題 467 問の 2 種類を用いる。

4.2 実験結果

RULE における汎化性能　モデルの一貫的な論理的推論能力を評価するために ReClor で訓練したモデルを RULE で評価した。モデルが根拠を正確に把握して主問題に正答しているのであれば、ReClor での正答率と RULE での正答率に大きな差はないはずである。精度評価の結果を表 1 に示す。RULE 全体での精度を見ると、ReClor での精度と乖離があるだけでなく、下線を付した $\Delta = 27.48$ から、ReClor において人間に匹敵するほどの精度を記録できるモデルであっても、根拠理解を評価の考慮に入れると人間の理解度とは未だ大きな開きがあることがわかる。

分析 1: 正答主問題における一貫性　ここでは正しく答えられた主問題の根拠問題に精度評価の焦点を当て、 “正しく答えられた主問題の根拠問題には同様に正しく答えられるべき” という要請をモデルが人間に比べどれほど満たしているか評価する。具体的には、モデルが正答した主問題それぞれの根拠問題 $N_i (i = 1, \ldots, 467)$ 個のうちモデルまたは人間が正答できた問題数 $C_i (C_i \leq N_i)$ の分布を算出することで、正答した主問題における両者の根拠理解の程度を測る。集計された C_i の分布は付録 B に示す。GPT-3 と比べ、人間の方が各 N_i についてより大きな C_i の占める割合が多く、正答した主問題の根拠問題にも一貫して正答できるという結果となった。

分析 2: 正答・誤答の根拠と正答率　ここでは根拠問題への解答可能性と主問題の正誤関係との関連を調査する。具体的には、根拠問題の正答率をその主問題への解答結果（主問題正答・誤答）と主問題

表 1 根拠問題と主問題（ReClor、下部）における精度 (%)。主問題への解答結果と作成元選択肢の正誤によって分けて集計し、件数を括弧で示した。

主問題	選択肢	人間	GPT-3	Δ
正解	正答	$93.28_{(893)}$	$86.72_{(276)}$	6.56
	誤答	$89.17_{(2493)}$	$51.93_{(746)}$	37.24
	全体	$90.67_{(3386)}$	$61.23_{(1022)}$	29.44
不正解	正答	$83.18_{(220)}$	$79.78_{(95)}$	3.40
	誤答	$82.16_{(612)}$	$57.09_{(289)}$	25.07
	全体	$82.93_{(832)}$	$62.74_{(384)}$	20.19
全体	正答	$91.28_{(1113)}$	$84.93_{(371)}$	6.30
	誤答	$87.72_{(3105)}$	$53.35_{(1035)}$	34.37
	全体	$89.14_{(4218)}$	$61.66_{(1406)}$	27.48
主問題		$79.08_{(1401)}$	$73.02_{(467)}$	6.06

における選択肢としての正誤（正答選択肢・誤答選択肢）という 2 つの条件で分けて集計した (表 1)。

表 1 を見ると、モデルは誤答選択肢の根拠問題に弱いが、一方で正答選択肢の根拠問題では人間に肉薄するほど良い精度を記録していることがわかる。

誤答根拠問題への弱さに関しては、複数選択肢の問題を多クラス分類として解くことに起因する個々の選択肢の推論へのフィードバックの不十分さと、否定的な根拠推論という推論の複雑さが原因として考えられる。正答選択肢に関しては、その根拠問題を解くことと主問題の正答を答えることが類似しているため、ReClor での学習が正答根拠問題にも有効だったのだと考えられる。

分析 3: 推論タイプごとの精度比較 ここではモデルの根拠理解に関する特徴を、問題で問われる推論の種類 (推論タイプ) の観点から定性的に分析する。ReClor の問題で問われる推論には 17 の種類があり、例えば “ Which one of the following would most seriously weaken the argument?” という質問文は、文脈文の主張を弱めるような選択肢を尋ねる Weaken という推論タイプに分類される。

ReClor と RULE での推論タイプごとの精度を集計したところ (付録 C)、全体的な傾向として、ReClor では推論タイプごとに精度に大きくばらつきがあるが、RULE に関しては比較的差は小さかった。RULE では全ての問題が一様に根拠を問う推論であるという面があるため、推論タイプごとの差が顕著には現れなかったのだと考えられる。また、より暗黙的な推論への理解を求められる推論タイプほど ReClor と RULE (正答選択肢) での精度差が顕著であった。例えば文脈文の趣旨を問う Implication、Conclusion/Main point、 文脈文の主張の不備を指摘する Weaken、Explain or Resolve では他の推論タイプに比べ ReClor と RULE 間で精度差が大きい。考えられる理由として、これらの推論タイプの主問題ではモデルは文脈文に明示されていない情報を把握する必要があるが、RULE ではその暗黙的な推論が明示化されているため、回答が容易になったと思われる。この結果は、提示された根拠文をモデルが問題への理解形成に利用できている可能性を示唆する。

分析 4: 特徴量分布の差異比較 人間が正答選択肢の根拠問題を答える際には、主問題を答える際に利用する情報と近い情報を文脈文に求めると考えられる。この仮説がシステムにも当てはまるか調べるため、DeBERTa の主問題予測時と正答根拠問題予測時の文脈文の寄与分布を Integrated Gradients [17] で求め、両分布のコサイン距離を計算した。

同時に、今回は根拠理解の難易度と分布距離の関係を調べた。具体的には、モデルの根拠問題正答数に基づいて主問題を EASY（$C/N \geq 2/3$）、MEDIUM（$2/3 > C/N > 1/3$）、HARD（$C/N \leq 1/3$）の 3 セットに分け、それぞれの問題集合に含まれる主問題と正答選択肢の寄与分布のコサイン距離を平均することで、根拠理解の難しさとの関連を調査した。対象サンプル数はそれぞれ 39、136、181 件である。

結果は EASY が 0.22、MEDIUM が 0.16、HARD が 0.11 であり、モデルが多くの根拠問題を解けている主問題 (EASY) では、モデルがその正答根拠問題を解く際に主問題を解く場合と比較的類似した情報を利用していることがわかる。これは根拠問題への頑健さと、モデルが主問題と根拠問題で利用する特徴量の共通性が関係していることを示唆する。

5 結論

本稿では一貫的な根拠理解という観点から論理的推論能力の信頼性を評価するデータセット RULE を構築し、精度評価と分析を行った。実験の結果、RULE における精度は人間と大きな乖離があるが、精度を下げている要因は誤答選択肢への低い頑健さにあり、また、主問題と根拠問題間の特徴量寄与分布の類似性が根拠問題の正答率と関連しているという示唆が得られた。根拠理解への更なる分析、一貫性を改善するための具体的な方策等については今後の課題としたい。

謝辞

本研究は JST さきがけ JPMJPR20C4 の支援を受けたものです。また、NAIST の渡辺太郎教授から貴重なご意見をいただきました。感謝いたします。

参考文献

[1] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In **International Conference on Learning Representations**, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXv preprint 1907.11692, 2019.

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv preprint 2006.03654, 2020.

[5] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6174–6184, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1012–1031, 2021.

[7] Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. BECEL: Benchmark for consistency evaluation of language models. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3680–3696, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[8] Abhilasha Ravichander, Matt Gardner, and Ana Marasović. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. arXiv preprint 2211.00295, 2022.

[9] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 3050–3065, Online, August 2021. Association for Computational Linguistics.

[10] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7716–7740, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[12] Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6740–6750, Online, July 2020. Association for Computational Linguistics.

[13] Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. Investigating the benefits of Free-Form rationales. arXiv preprint 2206.11083, 2022.

[14] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv preprint 2111.09543, 2021.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, pp. 3319–3328. JMLR.org, August 2017.

A　DeBERTa と GPT-3 の比較

GPT-3 は 5-shot で、DeBERTa は根拠問題作成に利用した主問題 467 問を抜いた ReClor の train データ 4,171 件でファインチューニングした。DeBERTa は 10 epoch 学習させた中で ReClor の dev データ 500 件における accuracy が最も高い epoch のモデルを使用した。

表 2　DeBERTa と GPT-3 の精度比較 (%)

	DeBERTa	GPT-3
ReClor	71.09	73.02
RULE	36.84	61.66

B　正答主問題への一貫性評価

表 3　根拠問題の正解数 (GPT-3) に応じた ReClor の問題数の分布 (%) と件数 (括弧内)

#N/C	0	1	2	3	4	total (%)
1	$50.0_{(15)}$	$50.0_{(15)}$				100.0
2	$20.3_{(16)}$	$53.2_{(42)}$	$26.6_{(21)}$			100.0
3	$8.5_{(8)}$	$33.0_{(31)}$	$37.2_{(35)}$	$21.3_{(20)}$		100.0
4	$0.0_{(0)}$	$16.7_{(23)}$	$20.3_{(28)}$	$42.0_{(58)}$	$21.0_{(29)}$	100.0

表 4　根拠問題の正解数 (人間) に応じた ReClor の問題数の分布 (%) と件数 (括弧内)

#N/C	0	1	2	3	4	total (%)
1	$12.6_{(11)}$	$87.4_{(76)}$				100.0
2	$2.0_{(5)}$	$19.3_{(47)}$	$78.7_{(192)}$			100.0
3	$0.3_{(1)}$	$4.4_{(13)}$	$20.5_{(61)}$	$74.7_{(222)}$		100.0
4	$0.2_{(1)}$	$1.3_{(6)}$	$6.5_{(31)}$	$15.4_{(74)}$	$76.7_{(368)}$	100.0

C　推論タイプごとの精度比較

17 種類ある推論タイプのうち、サンプル数が 30 以上の推論タイプ (9 種類) のみを集計対象とした。

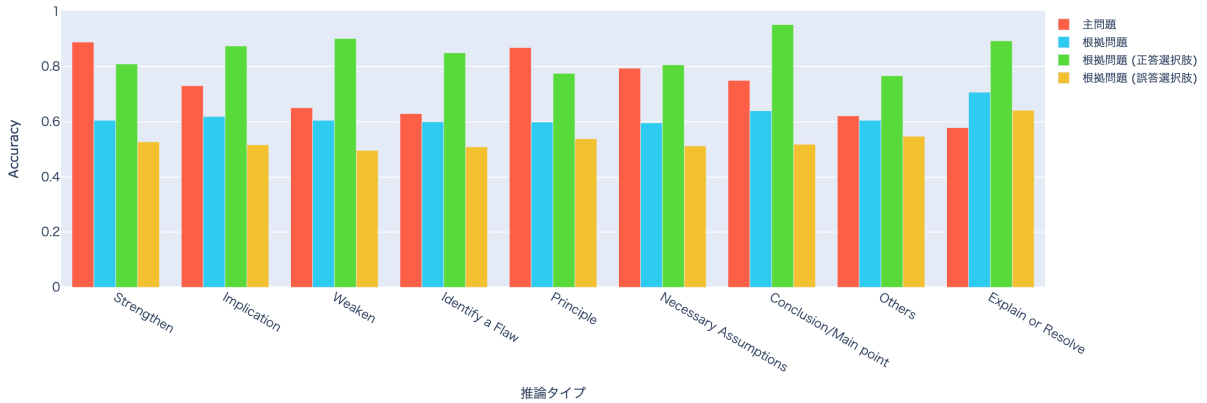

図 2　推論タイプごとの精度