機械読解システムの推論過程のベンチマークの構築

井之上直也 [1,2]　Pontus Stenetorp[3,2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所　[3]University College London
{naoya-i, inui}@ecei.tohoku.ac.jp
p.stenetorp@cs.ucl.ac.uk

1 はじめに

機械読解 (MRC) は自然言語理解の主要なベンチマークのひとつであり，近年多種多様なデータセットが公開されている [12, 20, 21, etc.]．しかしながら，これらのデータセットには言語理解力とは直接関係のない回答の手がかりが潜んでおり（例えば，*when* から始まる質問の回答には，記事内に初出の日付を答えればよい，など），システムは自然言語を理解せずとも多くの問題に回答できてしまうことが報告されている [15, 18]．含意関係認識，常識推論等の他の分野でも同様の問題が指摘されており [7, 11]，自然言語理解のベンチマークの品質保証はコミュニティ全体の大きな課題になりつつある．

こうした問題を解消するために，より難易度の高い MRC データセットが構築されてきた．例えば，**マルチホップ QA** と呼ばれるデータセットでは，回答に必要な情報が複数の記事に分散するように質問が作成されている [20, 21]．しかしながら，こうしたマルチホップ QA のデータセットにおいても冒頭で述べたような問題が指摘されており，結局のところ 1 記事の情報が回答の強い手がかりとなっていることが経験的に示されている [3, 10, 14]．

別の方向性として，システムの予測根拠を直接的に評価する試みもある [2, 6, 8, 16, 19, 21]．特に MRC の文脈では，Yang ら [21] による HotpotQA が代表的である．HotpotQA では，図 1 に示すように，質問に対する回答を出力するだけでなく回答の根拠となる文 (**回答根拠**) の集合を同定する必要がある．最終的に，システムの性能は回答の正答率と根拠の正答率の両面から評価される．しかしながら，回答根拠には予測根拠に無関係な情報も含まれている．例えば図 1の回答根拠 [1] において，「*Return to Olumpus* が *Malfunkshun* のアルバムである」という情報は予測根拠とは無関係である．ゆえに，回答根拠の同定精度の評価は，必ずしも予測根拠の評価に繋がらない．

そこで本稿では，$\mathcal{R}^4\mathcal{C}$ という新しい MRC のタスクを提案する[*1]．$\mathcal{R}^4\mathcal{C}$ では，質問への回答に加えて，予測の根拠となる推論の途中経過 (**導出**) を出力することを要求する．図 1 に示すように，導出は，Open Information Extraction [5] のような半構造化形式で表現する．このような半構造化表現を用いることにより，システムの予測根拠を高い解像度で評価できるようになるという利点がある．本稿の貢献は下記のとおりである：

- クラウドソーシングに基づく導出の大規模アノテーションの枠組みを提案する (§3)．
- 4,747 の質問回答ペア，またそれぞれについて 3 つの高品質な参照導出（合計 14,241 導出）が付与された $\mathcal{R}^4\mathcal{C}$ のデータセットを構築する (§4)．データセット，及びアノテーションの枠組みを https://naoya-i.github.io/r4c/ にて公開する．
- 複数の参照導出のアノテーションを用いて，導出の自動評価指標とその正当性を示す (§5.1)．また，これを用いて導出と回答根拠の性質の違いを定量的に示す (§5.2)．

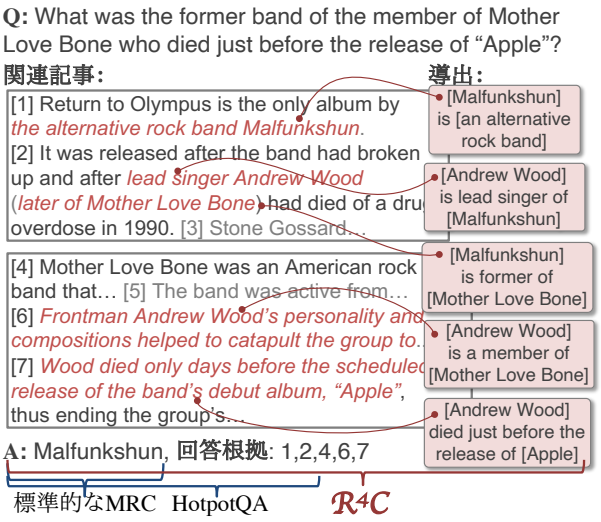

図1: 提案タスク $\mathcal{R}^4\mathcal{C}$ の概要．質問への回答に加え，回答に至る途中経過（導出）を出力することを要求する．文 [1], [2], [4], [6], [7] は回答根拠に対応する文の集合である．回答根拠のうち，導出に関連のある部分をハイライトしている．例は HotpotQA [21] からの引用である．

2 $\mathcal{R}^4\mathcal{C}$

2.1 タスクの定義

$\mathcal{R}^4\mathcal{C}$ は，標準的な MRC の問題設定を踏襲し，導出の出力を要求する．具体的には，入力として質問 q，関連文書集合 R が与えられたとき，回答 a，その根拠となる導出 D を出力するタスクである．

導出については，Open Information Extraction [5] で用いられる半構造化形式 (triplet) を用いる．形式的には，導出 D

[*1] $\mathcal{R}^4\mathcal{C}$ は “Right for the Right Reasons RC.” の略記である．

は**導出ステップ** $d_i \in D$ からなり，$d_i \equiv \langle h_i, r_i, t_i \rangle$ の形式を取るとする．ここで，h_i, t_i はエンティティを表す名詞句，r_i は関係を表す動詞句である（例は図 1を参照のこと）．

2.2 評価指標

導出には半構造化形式を採用するものの，個々の要素は自然言語表現であり，システムが出力する導出の自動評価の方法は自明でない．典型的には，こうした自然言語出力の評価にはクラウドソーシングが用いられるが，時間面でも費用面でも大きなコストがかかってしまい，システムの研究開発が滞ってしまう恐れがある．また，要約等の分野で広く用いられている自動評価指標（ROUGE [13] など）を用いることも考えられるが，これは単一の数値が最終的に与えられるのみであり，結果の解釈が曖昧になってしまう．

そこで $\mathcal{R}^4\mathcal{C}$ では，半構造化形式の利点を活かした評価を行う．具体的には，エンティティレベル，関係レベルでの個別の評価指標を設計し，それぞれの精度を要約の分野で広く用いられている ROUGE により評価する．後述するように，導出は回答根拠の一部の情報を要約したものであるため，ROUGE は適切な評価指標になっているといえる．

より具体的には，システムの導出 D について，n 人のアノテータにより与えられた n 個の参照導出 $G_1, G_2, ..., G_n$ があることを仮定する．複数の参照導出を用いた評価を行うことにより，導出の言語的多様性を吸収して適切な評価ができると考えられる（定量的な実証は 5.1 節を参照のこと）．その後，システムが出力した導出を hypothesis，参照導出を references とし，ROUGE-L Precision/Recall/F1 を計算する[*2]．また，システムのエラーをより正確に把握できるようにするために，2 つの部分的評価指標を利用する：(i) *entity-level*: D, G_i におけるエンティティのみを利用して hypothesis, references を構築し評価する，(ii) *relation-level*: 関係のみを利用して hypothesis, references を構築し評価する．導出を半構造化形式とすることで，単一の評価指標でなく解釈性の高い複数の評価指標を設計することができ，回答根拠の導出よりも挑戦的な課題に取り組めるようになる．

3 データセット構築

大規模な導出のデータセットを構築するために，クラウドソーシングを用いる．先行研究では，導出のアノテーションのような複雑な，特に自由記述形式のタスクでは，アノテーション結果の品質管理が難しいことが知られている [2, 21, etc.]．本稿では，高品質なアノテーションを得るためのクラウドソーシングのタスク設計，および全体のフローを提案する．

3.1 クラウドソーシングインターフェイス

冒頭で述べたように，近年の研究により，高品質な MRC のデータセットが数多く利用可能になっている [20, 21, etc.]．そこで本研究では，既存の MRC のデータセットに対して導出を重層的に付与するインターフェイスを構築する．MRC のデータセットは，(i) 質問，(ii) 回答，(iii) 回答根拠となる文書（**根拠文書**）集合（さらに，回答根拠が付与されていてもよい）の形式で与えられていることを仮定する．

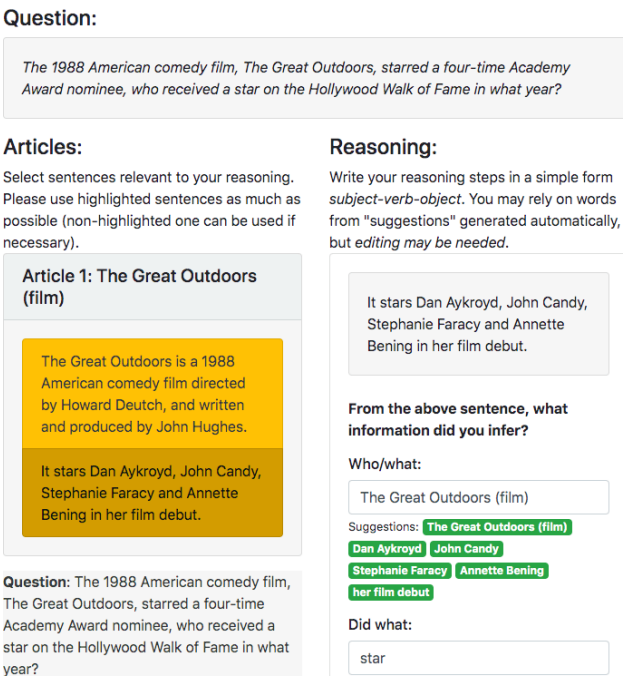

図2: 導出アノテーションのためのクラウドソーシングタスク．記事内の文をクリックし，導出を半構造化形式で書き込む．

まずは，クラウドワーカ（以後，**ワーカ**と呼ぶ）に根拠文書を注意深く読み込ませるために，質問に回答させる．このとき，ワーカの負荷を軽減するために，4 つの回答候補[*3]を提示する．MRC データセットには稀にノイズが含まれているため，「どれにも当てはまらない」という選択肢も追加する．

つぎに，ワーカは導出を入力する（図 2を参照）．まずは根拠文書中の文をクリックし（画面左側），半構造化形式により導出を入力する（画面右側）．ワーカの負荷の軽減，およびアノテーションの一貫性の担保のために，テキストボックスの下側に入力候補を提示する．これらの入力候補には，あらかじめ定義された前置詞，根拠文書から自動抽出[*4]された名詞句，動詞句を含む．MRC データセットが回答根拠のアノテーションを持つ場合には，回答根拠をハイライトして表示する．

3.2 ワークフロー

低品質なワーカを排除するために，まずは資格テストを実施する．資格テストでは，3.1 節で説明したタスクと同様のタスクを実施し，高品質なアノテーションができるワーカを人手により同定する．最終的なアノテーションは，これらの**有能ワーカ**のみにより実施される．

クラウドソーシングのプラットフォームとして，Amazon Mechanical Turk (AMT)[*5]を用いる．本稿では，5,000 以上のタスクをこなし，かつタスク承認率が 95.0% 以上であるワー

[*2] https://github.com/Diego999/py-rouge, apply_best=True.

[*3] 正しい回答と，関連文書のタイトルからランダムに選んだ 3 つの誤答からなる．

[*4] Spacy: https://spacy.io/

[*5] https://requester.mturk.com/

分割	# QA	導出の分布				
		ステップ数 2	3	4	≥ 5	合計
train	2,612	5,448	1,708	476	204	7,836
dev	2,702	5,322	1,997	623	164	8,106
dev (fin.)	2,135	4,279	1,549	466	111	6,405

表1: $\mathcal{R}^4\mathcal{C}$ コーパスの規模．合計 4,747 の質問回答ペアが含まれ，それぞれには 3 つの導出が付与されている (合計 14,241 導出)．

# ref.	Entity P/R/F	Relation P/R/F	Full P/R/F
1	82.0/77.8/76.9	62.3/49.5/50.1	80.6/72.4/74.4
2	83.6/87.1/83.4	65.9/65.3/62.0	80.8/81.6/80.1
3	84.7/90.0/85.7	71.3/73.9/69.5	83.2/84.9/83.0

表2: “正しい” 導出の評価値と参照導出の数の関係．複数の参照導出を持つことの重要性が確認できる．

カにのみ，資格テストを受けることを許可した．資格テストでは，1 事例あたり ¢15 を報酬として支払った．また，最終的なアノテーションでは，1 事例あたり ¢30 を報酬として支払った．1 事例あたり，3 人のワーカを割り当てた．

3.3 データセット

3.1 節で述べた基準を満たす MRC データセットは複数存在する（例えば，SQuAD [17], WikiHop [20]）．本稿では，マルチホップ QA の研究で広く用いられているデータセットの一つである HotpotQA [21] を用いる[*6]．マルチホップ QA をベースとすることにより，導出が複数の記事に分散し，より挑戦的な研究課題を含むデータセットになることを狙いとする．

アノテーションでは，HotpotQA の 90,564 の訓練事例より 3,000 事例を，7,405 の開発事例より 3,000 事例をサンプルした（つまり，18,000 のタスクが AMT 上で発行された）．資格テストとインターフェイス開発のために，これらとは別に 300 訓練事例をサンプルし利用した．また，HotpotQA に付与されている 回答根拠をアノテーションに利用した．

本稿で構築するデータセットは，主としてシステムの予測根拠の評価を目的としたものであるが，システムの fine-tuning を行えるよう訓練事例にもアノテーションを行っている．

3.4 統計

資格テストには，256 名のワーカが参加し，45 名の有能ワーカを得た．低品質なアノテーションを排除するために，(i) 質問への回答が誤っている事例，(ii) 「どれにも当てはまらない」という回答を持つアノテーションを排除した．最終的には，1 事例につき 3 つの導出が付与されている事例のみを残しコーパスを構築した．表 1 に統計を示す（train, dev を参照）．

4 品質評価

4.1 方法

付与した導出から元の回答を復元できるかどうかをチェックするために，AMT を利用した評価を実施した（**回答復元性チェック**）．本タスクでは，HotpotQA の質問が与えられたとき，導出**のみ**に基づいて質問に回答できるかを 3 レベル (Yes, Likely, No) により評価する．

評価は開発セットの 8,106 導出を対象とし，1 質問について 3 人のワーカを割り当てた．評価の信頼性を担保するために，3.4 節で集めた有能ワーカのみにタスクを依頼した．本タスクでは，¢15 を報酬として支払った．

また，対応する回答根拠に導出の情報が含まれているかを二人の専門家により評価した（**導出可能性チェック**）．ここでは，50 個の導出ステップを評価対象とした．

4.2 結果

回答復元性チェックでは，評価結果の一致率として Krippendorff's α 0.266 を得た (a fair agreement)．また，多数決投票により次のような分布を得た: Yes: 94.8%, Likely: 2.3%, No: 1.5% （票割れ: 1.5%）．導出可能性チェックでは，回答復元性チェックにおいて Yes と判断された事例が 44 事例あり，このうち 90.0% (40/44) の導出について，二人のアノテータより「導出可能」という回答が得られた[*7]．以上の結果より，導出のアノテーションの複雑さにも関わらず，提案したアノテーションの枠組みにより高品質な導出のアノテーションが可能であることが示唆された．

最終的な $\mathcal{R}^4\mathcal{C}$ データセットでは，回答復元性チェックにおいて Yes と判断された事例のみを開発セットとした．統計を表 1に示す（dev-final を参照）．本コーパスは複数の参照導出が付与された初めての MRC データセットである．最も関連深い研究として，科学ドメインの質問応答データセットに導出を付与した WorldTree コーパス [9] があるが，(1) 問題設定が異なる（根拠文書が添付されていない），(2) 規模が異なる（1,680 事例の質問-回答ペアに対し，一つの参照導出が付与されている），(3) 参照導出が自然言語である，という点で異なる．

5 分析

5.1 複数の参照導出の必要性

2 節では，複数の参照導出により導出の正確な評価が可能であることを仮定した．この仮説を検証するために，参照導出数と “正しい” 導出の評価値の間に正の相関が見られるかを確認した．ここで，正しい導出とは有能ワーカにより新たに書き下された開発セットの 100 問の導出とする．

表 2 に実験結果を示す．参照導出の数 (# ref.) を増やすことが評価値の改善に繋がることから，複数の参照導出を持つことが適切な導出の評価に繋がることが示唆された．同時に，有能ワーカのアノテーション品質が高いことも再確認することができた．また，このデータセット上でシステムが目指すべき上限値 (human upperbound) が参照数 3 の場合の評価値に対応することも示している．

また，entity-level の性能向上よりも，relation-level の性能

[*6] https://hotpotqa.github.io/ 2020 年 1 月現在，40 以上の結果の投稿がある．

[*7] ただし，判断が割れた事例についても，タイプミスなどの軽微なミスを含む事例であり，本質的なエラーはなかった．

モデル	Entity P/R/F	Relation P/R/F	Full P/R/F
IE	20.5/63.0/26.7	40.0/61.5/41.3	20.7/58.4/27.2
CORE	60.7/69.1/62.4	74.4/40.3/46.6	72.2/56.4/61.5

表3: ベースラインモデルの導出正解率の評価．図 2 の human upperbound とは大きく乖離があり，回答根拠からコア情報を抜き出す (CORE)，またはあらゆる情報を抜き出す (IE) だけでは導出にはならないことを示している．

向上幅が高いことから（8.8% v.s. 19.4%），関係の表現はエンティティの表現よりも多様性に富むことが示唆される．実際に，位置を示す関係として *is in*, *is a town in*, *is located in* といった異なる言語表現が付与されていた．

5.2 タスクの性質

$\mathcal{R}^4\mathcal{C}$ の性質を明らかにするため，IE, CORE の二つのベースラインモデルを評価する．まず，IE モデルは，回答根拠からあらゆる関係を抽出する[*8]．次に，CORE モデルは，回答根拠のコアな情報を抽出する．具体的には，各回答根拠（その記事タイトルを t とする）について，依存構造木に基づいてルート動詞 v とその最初の右側の子供 c_r を抽出し，$\langle t, v, c_r \rangle$ を導出として出力する．

表 3 に実験結果を示す．まず，ベースラインモデルの性能は 5.1 節の human upperbound の性能には遠く，$\mathcal{R}^4\mathcal{C}$ は HotpotQA の回答根拠検出とは本質的に異なるタスクであることが示唆された．すなわち，回答根拠から単にあらゆる関係を抽出する，またはコアな情報を抽出するだけでなく，推論に関連した箇所を適切に出力してこそ高い評価を得られるタスクになっており，MRC システムの内部挙動のより精緻な評価を可能にすることが確認できた．

6 おわりに

MRC システムの内部挙動の評価に向けて，新たな MRC タスク $\mathcal{R}^4\mathcal{C}$ を提案し，データセットを大規模かつ高品質に構築できる枠組みを提案した．評価実験により，$\mathcal{R}^4\mathcal{C}$ データセットは正しい導出を適切に捉えられること，および既存の HotpotQA の回答根拠検出タスクにはない新しい研究課題を提供することを定量的に示した．構築したデータセット，アノテーションシステム，および自動評価スクリプトは https://naoya-i.github.io/r4c/ にて公開予定である．

今後の課題として，構築したデータセットを用いて既存の MRC システムの内部挙動を評価することが挙げられる．また，説明可能な MRC システムの開発に向けて，近年顕著な成果を上げている事前訓練済み言語モデル [4] を利用し，導出の組み立てと回答の探索を同時にモデル化することを検討している．

謝辞

本研究は JSPS 科研費 19K20332 の助成，JST，CREST，JPMJCR1513 の支援を受けたものである．

[*8] 本実験では Stanford OpenIE [1] を用いた．

参考文献

[1] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. “Leveraging linguistic structure for open domain information extraction”. In: *In Proc. of ACL-IJCNLP* 1.1 (2015), pp. 344–354.

[2] Oana-maria Camburu et al. “e-SNLI : Natural Language Inference with Natural Language Explanations”. In: *Proc. of NIPS*. 2018, pp. 1–13. arXiv: 1812.01193.

[3] Jifan Chen and Greg Durrett. “Understanding Dataset Design Choices for Multi-hop Reasoning”. In: *Proc. of NAACL: HLT, Volume 1 (Long and Short Papers)*. 2019, 4026–4032”.

[4] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proc. of NAACL (to appear, preprint is available at arXiv:1810.04805)*. 2019.

[5] Oren Etzioni et al. “Open information extraction from the web”. In: *Communications of the ACM* 51.12 (2008), pp. 68–74.

[6] Angela Fan et al. “ELI5: Long Form Question Answering”. In: 2019, pp. 3558–3567. arXiv: 1907.09190.

[7] Suchin Gururangan et al. “Annotation Artifacts in Natural Language Inference Data”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (2018), pp. 107–112.

[8] Peter A. Jansen. “Multi-hop Inference for Sentence-level TextGraphs: How Challenging is Meaningfully Combining Information for Science Question Answering?” In: *Proc. of TextGraphs-12*. 2018, pp. 12–17. arXiv: arXiv:1805.11267v1.

[9] Peter A. Jansen et al. “WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-Hop Inference”. In: *Proc. of LREC*. 2018, pp. 2732–2740. arXiv: 1802.03052.

[10] Yichen Jiang et al. “Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension”. In: (2019). arXiv: 1906.05210.

[11] Pride Kavumba et al. “When Choosing Plausible Alternatives, Clever Hans can be Clever”. In: *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. 2019, pp. 33–42.

[12] Tomáš Kočiskỳ et al. “The NarrativeQA Reading Comprehension Challenge”. In: *Trans. of ACL* 6 (2018), pp. 317–328. arXiv: 1712.07040.

[13] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Proc. of the Workshop on Text Summarization Branches Out*. 2004, pp. 74–81.

[14] Sewon Min et al. “Compositional Questions Do Not Necessitate Multi-hop Reasoning”. In: 2019. arXiv: 1906.02900.

[15] Pramod K. Mudrakarta et al. “Did the Model Understand the Question?” In: *Proc. of ACL*. 2018, pp. 1896–1906. arXiv: 1805.05492.

[16] Fatema Nanzneen Rajani et al. “Explain Yourself ! Leveraging Language Models for Commonsense Reasoning”. In: *Proc. of ACL*. 2019, pp. 4932–4942. arXiv: arXiv:1906.02361v1.

[17] Pranav Rajpurkar et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: *Proc. of EMNLP*. 2016, pp. 2383–2392. arXiv: 1606.05250.

[18] Saku Sugawara et al. “What Makes Reading Comprehension Questions Easier?” In: *Proc. of EMNLP*. 2018, pp. 4208–4219.

[19] James Thorne and Andreas Vlachos. “Automated Fact Checking: Task formulations, methods and future directions”. In: *Proc. of COLING*. 2018, pp. 3346–3359.

[20] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. “Constructing Datasets for Multi-hop Reading Comprehension Across Documents”. In: *Trans. of ACL* 6 (2018), pp. 287–302. arXiv: 1710.06481.

[21] Zhilin Yang et al. “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering”. In: *Proc. of EMNLP*. 2018, pp. 2369–2380. arXiv: 1809.09600.