様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ９ 日現在

機関番号：６２６１５
研究種目：研究活動スタート支援
研究期間：2020～2022
課題番号：２０Ｋ２３３３５

研究課題名（和文）補助問題を備えた説明性の高い機械読解評価基盤の構築

研究課題名（英文）Creating Auxiliary Questions for Explainable Evaluation of Machine Reading Comprehension

研究代表者
菅原　朔（Sugawara, Saku）

国立情報学研究所・コンテンツ科学研究系・助教

研究者番号：１０８５５８９４

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：言語理解システムを着実に開発するには理解の過程に関する精緻な分析と評価が必要であるが、既存のタスクではシステムの能力について十分な説明性が確保されていなかった。本研究は読解問題に焦点を当て、回答に至るまでの根拠についての理解を補助的な問題として課すことで詳細な評価を可能にするデータセットを構築した。既存の選択式の読解問題に対してそれぞれの選択肢が正解・不正解になる根拠情報をクラウドソーシングで収集し、その根拠情報自体が答えになるような選択式の問題を作成することで元々の問題に対して根拠まで含めた一貫性のある理解がなされているかどうかを問えるようにした。

研究成果の学術的意義や社会的意義
言語理解を実現するシステムの構築は自然言語処理における最大の目標のひとつである。システムを着実に開発するには言語理解に関する精緻な分析と評価が必要であり、本研究によって得られたデータセットは読解問題の回答に至るまでのプロセスを分解して補助的な問題として課すことで詳細な評価を可能にした。これにより現状のシステムの限界が示され、本データセットは今後の改善を促進する上で重要な役割を果たす。

研究成果の概要（英文）：Developing natural language understanding systems requires detailed analysis and evaluation of the language understanding process. However, existing tasks have not ensured sufficient accountability for systems' capabilities. This study focused on reading comprehension questions and constructed a new dataset that enables detailed evaluation by testing the understanding of the rationale in the question answering process. We used crowdsourcing to collect rationale texts for the correct and incorrect answers of existing multiple-choice reading comprehension questions, and then used the rationale information to create an auxiliary set of multiple-choice questions that help us to determine whether or not a system correctly answers the question, including the rationale in a consistent manner.

研究分野：自然言語処理

キーワード：自然言語処理　計算言語学　自然言語理解

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

【学術的背景】

自然言語処理分野では、人間のように文章を理解するシステムを構築することがひとつの大きな目標である。システムを評価するためにはベンチマークとなるタスクとして文章題を解かせる**機械読解タスク**があり、回答には既存の自然言語処理で扱われている様々な要素技術（例：共参照解析・常識推論）が総合的に求められるとされる。近年はクラウドソーシングにより大規模なデータセットが多数提案されており、中には人間に近い精度が達成されているものもある。

【問題点】

しかし、既存のデータセットは人間らしい高度な言語理解を要求・評価している（そのための訓練に足る）とは必ずしも言えないという問題点が近年の分析で示されている。例えば Jia and Liang (2017) は問題文に表層的によく似た敵対的な文（ただし本来の正答は変わらない）を文脈文に挿入することでシステムの性能が大きく低下することを発見した。また業績[2]は問題文の疑問詞しか見なくてもシステムが正答できる問題が多く存在することを明らかにした。図 1 に例を示す。文脈文中から正答を抜き出す形式の問題であるが、問題文の when に対して文脈文に時間表現が１つしか存在せず、 when 以降を読まなくても正答できる（また仮に他の時間表現が存在しても、問題文中の名詞を見るだけで容易に正答を特定できる）。こうした問題点は既存データセットが「性能の良いシステムは具体的に何が理解できると言えるのか」という説明性が低いまま設計されていることに由来している。

【学術的問い】

高度な言語理解システムを着実に開発するにはデータセットの説明性の向上が欠かせない。そのためには問題を解くまでのプロセスが詳細に明確化され評価指標として含まれている必要がある。本研究はその要請を「読解のプロセスを陽に記述できるか」と「その記述を読解の精緻な評価に利用できるか」という 2 つの問いに分けて取り組む。

２．研究の目的

上記の問いに答えるため、以下の目的を設定した。

① 読解のプロセスを明確にする方法を考案し、実際に既存のデータセットにおいて収集する

② 収集した読解のプロセスについての情報を利用して新たな補助的な問題を作成し、評価用データセットとして構築する

３．研究の方法

【根拠情報の収集】

① については、まず既存の読解問題データセットである ReClor (Yu et al., 2020) から選択式の問題 600 問（以降、主問題と呼ぶ）を取り出た。クラウドソーシングにおいて事前タスクで選抜された参加者に対して、主問題の各選択肢をその正誤とともに提示し、正誤理由の執筆を依頼した。収集後、具体的でない・意味が明瞭でないような低品質な根拠文をさらにクラウドソーシングによるチェックを通してフィルタリングし、1,860 件の根拠文を得た。

【評価用データセットの構築】

② について、収集した根拠文が 3 ないし 4 件ある主問題について、各選択しの正誤の根拠を問うような質問文を GPT-3 を利用して生成した。具体的には、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。その後、品質評価として人間の正答率を測定し、クラウドワーカー3 人中 2 人が正解または 3 人とも正解なしを選んだ事例 1,406 件を補助問題として作成した。これを用いて人間と言語理解システムの精度を比較したところ、全体で人間の正答率が 89.14% だったのに対して、 GPT-3 (text-davinci-003) の正答率は 61.66% であった。現状のシステムには根拠問題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、推論タイプと文章中の特徴量寄与度の相互関係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。

４．研究成果

以上の研究結果は、主問題・補助問題あわせて 2000 件を超えるデータセットとして一般に公開することで関連研究者がシステムの評価に利用できるように準備している最中である。この内容は言語処理学会第 29 回年次大会で発表し、第一著者の川畑は若手奨励賞を受賞した。国内で類似の取り組みはあるもののデータセットの構築に至っている例はなく、また国外を見ても同様のデータセットで複雑な論理的推論が必要な補助問題を作成しているはほとんどなく、今回の取り組みは新規性が高いと考えられる。今後、データセットの拡張やより発展的な題材において同様の手法で補助問題が作成できるか検討し、さらなる精緻な評価の実現に向けて取り組んでいく予定である。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　3件／うち国際共著　1件／うちオープンアクセス　4件）

１．著者名 Shinoda Kazutoshi, Sugawara Saku, Aizawa Akiko	４．巻 -
２．論文標題 Improving the Robustness of QA Models to Challenge Sets with Variational Question-Answer Pair Generation	５．発行年 2021年
３．雑誌名 Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop	６．最初と最後の頁 197-214
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2021.acl-srw.21	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Shinoda Kazutoshi、Sugawara Saku、Aizawa Akiko	４．巻 -
２．論文標題 Can Question Generation Debias Question Answering Models? A Case Study on Question?Context Lexical Overlap	５．発行年 2021年
３．雑誌名 Proceedings of the 3rd Workshop on Machine Reading for Question Answering	６．最初と最後の頁 63-72
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2021.mrqa-1.6	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Saku Sugawara, Pontus Stenetorp, Akiko Aizawa	４．巻 1
２．論文標題 Benchmarking Machine Reading Comprehension: A Psychological Perspective	５．発行年 2021年
３．雑誌名 Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume	６．最初と最後の頁 1592-1612
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 川畑輝，菅原朔	４．巻 -
２．論文標題 読解問題における論理推論の一貫性評価	５．発行年 2023年
３．雑誌名 言語処理学会 第29回年次大会 発表論文集	６．最初と最後の頁 2914-2919
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 川畑輝，菅原朔
2．発表標題 読解問題における論理推論の一貫性評価
3．学会等名 言語処理学会 第29回年次大会
4．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関