研 究 終 了 報 告 書

「言語理解の能力に基づく機械読解ベンチマークの構築」

研究期間: 2019年10月～2021年3月
研 究 者: 菅原 朔

1. 研究のねらい

読解問題を解くのに必要となる能力がラベル付けされた機械読解データセットを構築することで、システムの性能について精緻な評価指標を与える。ここで能力は自然言語処理で存立している基礎技術（照応解析・意味役割解析等）を単位とし、問題の作成にあたり自然言語の形式表現体系を新たに提案する。このテストの応用により、実世界における言語理解システムの説明性・頑健性の向上が期待される。

2. 研究成果

（1）概要

本研究の大目標は、解答のために意図した能力が要求されることが明確化されたベンチマークデータセットを構築することである。必要な能力は人手によって保証されることが理想的であるが、その前段階として能力をある程度自動的に判定する手法が存在することが望ましい。そこで、入力情報の削除という観点からシステムの振る舞いを分析する手法を検討した。

また提案するベンチマークタスクの有効性を高めるためには、読解にどのような能力が必要になるか、またどのようにその能力の保証をタスクデザインとして行うかという点に何らかの理論的な裏付けが必要である。そこで、心理学における読解研究や心理測定学を参照して、提案内容の拠り所となるような基盤を論じた。

（2）詳細

研究テーマ 1.
機械読解データセットにおいて必要となる言語理解能力の半自動的な分類手法の提案

【背景と課題】

これまでに提案されている機械読解データセットにおいては、問題に正しく回答するためにどのような能力が必要とされているかの分析が少なく、成績の良いシステムがどのようなシステムを実際に備えていると言えるかについて根拠付けることが難しかった。先行研究では人手によるアノテーション等を行って小規模のデータを分析していたが、昨今使われる大規模なデータセットへの適用が難しく、自動的な分析手法が望まれていた。

【手法と評価】

「ある能力を行使するために必要となる要素」を課題文中から落としてもシステムが正答できるのであれば、その問いはその能力を行使しなくても解ける可能性が高いと言える。たとえ

ば課題文の全体の語順をランダムに並び替えても依然としてシステムが正しく回答できるのであれば、システムはその回答の根拠として課題文の語順の理解を必要としなかったのではないかと推測できる。そのような問いにおいては、語順を理解することを含む能力、たとえば統語構造の理解などが必要とされない可能性があり、言語理解の評価には足りえない場合がある。以上の考察から、「言語理解に必要な要素を落とす」という手法を合計 12 個の能力に対応させて提案し、既存の 10 のデータセットに適用した。結果として、多くのデータセットですでに解けている問いでは高度な言語理解が要求されているわけではなさそうである、という観察を得た。またその観察を小規模なデータでの人手の評価を通して検証した。より高度な言語理解の評価を行うためには、問いのデザインをより洗練させることが必要である。

【成果】
本研究の論文は人工知能分野のトップ国際会議のひとつである The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020) に採択された。

研究テーマ 2.
機械読解データセットがベンチマークタスクであるために求められる要件の整理

【背景と課題】
上記プロジェクトで取り組んだように、機械読解における既存のデータセットではどのような評価が行われているかの意義付けが弱く、成功したシステムが備える能力について説明性が弱いという課題があった。実践的な試みや分析は様々になされているが、こうした取り組みを支えるための理論的な基盤はまだ整備されていない。

【提案】
読解とは何か、それをどのように評価するか、という 2 つの観点から理論的な基礎を整備した。具体的には、心理学における読解の研究から計算論的な言語理解のモデルを援用し、読解とは何かについての心理学的な知見と既存の機械読解データセットにおける能力の単位の対応付けを行った。後者の評価手法については、心理測定学における妥当性の構成概念を援用し、既存のデータセット設計がそれをどれほど満たしているかについて検討を行った。結果として、「他の知覚情報との結びつきや新規な状況の理解を問うものがさらに必要」と「回答のプロセスを中間的に適切に評価できるようなタスクデザインが必要」という結論を得た。こうした結論は、今後当分野で求められる要件の提示とその理論的な根拠を与え、研究者の参照先になりうるという意義がある。

【成果】
本研究の論文は計算言語学分野のトップ国際会議のひとつである The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021) に採択された。

3. 今後の展開

ベンチマークデータセットの素材となる質問の収集には、クラウドソーシングによる大規模化が欠かせない。今後はまずクラウドソーシング手法の比較検討を通してより高品質・挑戦的な質問を収集するための手法を確立する。なお本研究課題は JST さきがけ「信頼される AI の基盤技術」の研究課題「説明性の高い自然言語理解ベンチマークの構築」(2020 年度採択)において継続・発展させる。

4. 自己評価

評価用データセット構築に向けてその分析手法・理論的基礎を与える研究が進展し、2本の論文が採択されたことは一定の進展・成果と言える。またデータセットの素材となる質問を集めるためのクラウドソーシング手法の比較研究も進展しており、当初の計画通りに研究費を利用しながら、研究者・開発者が広く利用できるようなデータセット構築に向けて着実に研究を進めている。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:3件

1. Saku Sugawara, Pontus Stenetorp, Kentaro Inui, Akiko Aizawa. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20), 2020, 8918-8927.
Existing analysis work in machine reading comprehension (MRC) is largely concerned with evaluating the capabilities of systems. However, the capabilities of datasets are not assessed for benchmarking language understanding precisely. We propose a semi-automated, ablation-based methodology for this challenge; By checking whether questions can be solved even after removing features associated with a skill requisite for language understanding, we evaluate to what degree the questions do not require the skill. Experiments on 10 datasets (e.g., CoQA, SQuAD v2.0, and RACE) with a strong baseline model show that, for example, the relative scores of the baseline model provided with content words only and with shuffled sentence words in the context are on average 89.2% and 78.5% of the original scores, respectively. These results suggest that most of the questions already answered correctly by the model do not necessarily require grammatical and complex reasoning. For precise benchmarking, MRC datasets will need to take extra care in their design to ensure that questions can correctly evaluate the intended skills.
2. Saku Sugawara, Pontus Stenetorp, Akiko Aizawa. Benchmarking Machine Reading Comprehension: A Psychological Perspective. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), 2021, accepted.

Machine reading comprehension (MRC) has received considerable attention as a benchmark for natural language understanding. However, the conventional task design of MRC lacks explainability beyond the model interpretation, i.e., reading comprehension by a model cannot be explained in human terms. To this end, this position paper provides a theoretical basis for the design of MRC datasets based on psychology as well as psychometrics, and summarizes it in terms of the prerequisites for benchmarking MRC. We conclude that future datasets should (i) evaluate the capability of the model for constructing a coherent and grounded representation to understand context-dependent situations and (ii) ensure substantive validity by shortcut-proof questions and explanation as a part of the task design.
3. Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), Dec 2020, pp.6609-6625.
A multi-hop question answering (QA) dataset aims to test reasoning and inference skills by requiring a model to read multiple paragraphs to answer a given question. However, current datasets do not provide a complete explanation for the reasoning process from the question to the answer. Further, previous studies revealed that many examples in existing multi-hop datasets do not require multi-hop reasoning to answer a question. In this study, we present a new multi-hop QA dataset, called 2WikiMultiHopQA, which uses structured and unstructured data. In our dataset, we introduce the evidence information containing a reasoning path for multi-hop questions. The evidence information has two benefits: (i) providing a comprehensive explanation for predictions and (ii) evaluating the reasoning skills of a model. We carefully design a pipeline and a set of templates when generating a question-answer pair that guarantees the multi-hop steps and the quality of the questions. We also exploit the structured format in Wikidata and use logical rules to create questions that are natural but still require multi-hop reasoning. Through experiments, we demonstrate that our dataset is challenging for multi-hop models and it ensures that multi-hop reasoning is required.

（2）特許出願

研究期間累積件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

研究科長賞，東京大学大学院情報理工学系研究科，2020 年 3 月