公開

研 究 終 了 報 告 書

「説明性の高い自然言語理解ベンチマークの構築」

研究期間:2020 年 12 月～2024 年 4 月
研 究 者:菅原朔

1. 研究のねらい

人間のように文章を理解するシステムの開発は自然言語処理研究の目標のひとつである。システムが言語を理解できたかどうかを評価するベンチマークとして近年扱われるタスクに機械読解があり、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が総合的に求められるとされる。しかし既存の機械読解データセットは必ずしも人間らしい高度な理解を要求しているとは言えず、高精度なシステムに何ができるのか説明性が低いという課題があった。本提案はこの課題に評価指標と品質保証という観点から取り組み、両者を解決する説明性の高いベンチマークの構築と、その持続的開発を可能にするロードマップの作成を目標とする。具体的には、人間が言語を理解するのに必要な能力を説明の単位として定義し、質問を正答するのに必要な能力を保証できるようなタスクを設計する。クラウドソーシングを活用しながら大規模で高品質のデータを収集するための方法論を整備し、精緻な評価指標を備えて品質が保証された高度な言語理解のための評価用データセットを構築することで、人工知能技術を信頼できる形で安全に利用できる社会の実現に貢献することを目指す。

2. 研究成果

(1)概要

研究のねらいにあるように、本さきがけ研究で取り組んだことは「A. 持続的開発を可能にするロードマップの作成」「B. データセット収集の方法論」「C. ベンチマークデータセットの提案」の大きく 3 つに分けることができる。

「A. ロードマップの作成」については、心理学の議論を参照しながらベンチマークが満たすべき要件を整理する研究を進めた。1 件目(EACL 2021 採択)の研究では、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。2 件目(ACL 2023 Findings 採択)の研究でこれを具体化し、心理測定学における妥当性議論という枠組みを参照しながら、測定結果の解釈の妥当性を手続き的に向上させるためのステップを自然言語理解ベンチマークに対して提案してチェックリストの形でまとめることに成功した。以上により、「今後どのような要件を満たす評価手法を開発していけばよいか」ということを誘導できるロードマップの作成を一定程度実現した。

「B. データセット収集の方法論」については、 2 件の研究を通してクラウドソーシングでどのように高品質なデータを収集するかという問題に取り組んだ。具体的にはクラウドワーカーの選定やインセンティブの方法と、読解問題の作問に使うべき文章の選択方法に一定の知見を与えた。一連の研究はニューヨーク大学との共同研究として行われ、ACL 2021 と ACL 2022 に論文が採択された。

「C. ベンチマークデータセットの提案」では、上記の研究の知見を活かしながら実際に新たなタスクの提案を行った。具体的には、状況に応じて結末が変わるような文脈における常識推論

(COLING 2022)、賛否のある複雑な議論における論理的推論(EMNLP 2023)、埋め込み的に複雑な状況にある語用論的推論(CoNLL 2023)などを問うためのデータセットを提案した。大規模言語モデルが急速な発展を見せるなか、そうした強力なモデルでも苦手とするような能力・言語現象を問えるようなタスクを提案することに成功している。

以上、成果としては難関国際会議論文 7 件を主著または責任著者として発表している。また、さきがけ研究期間中に言語理解ベンチマークやその周辺研究に取り組む学生の指導にあたり、正式な指導学生ではないが共同して指導に当たる立場として共著で出版した難関国際会議論文が 4 件、国際会議ワークショップ論文が 2 件(すべて査読あり)存在する。

(2)詳細

研究テーマ A「説明性の高いベンチマークの持続的開発を可能にするロードマップの作成」

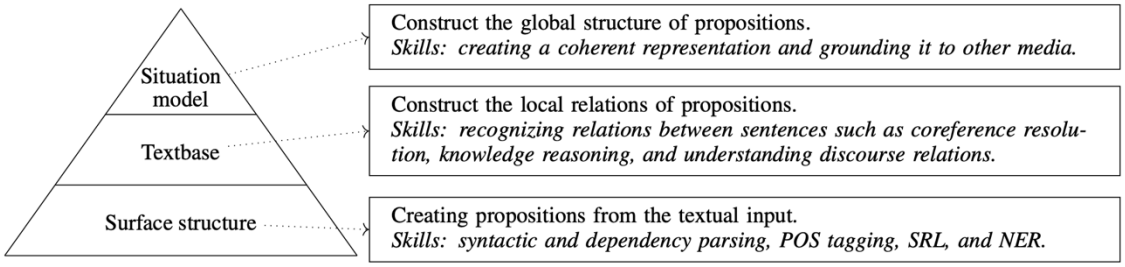

図 1 心理学の計算論的な読解のモデル化での表現のレベルと自然言語理解能力の対応。

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(図 1)。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満たすべき要件についてチェックリストを整備した(図 2)。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。本研究は計算言語学分野の難関国際会議である EACL 2021 と ACL 2023 Findings に採択され、発表を行った。

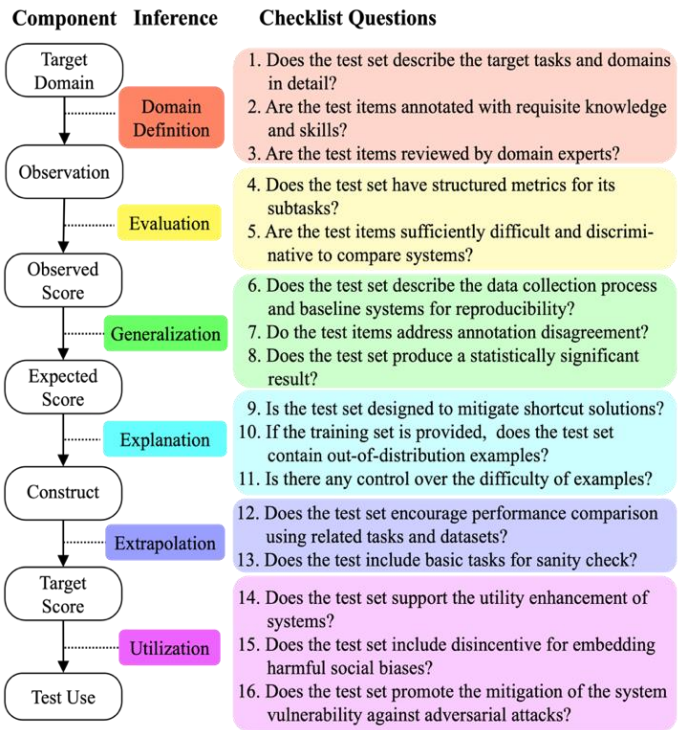

図 2 妥当性議論における 16 個のチェックリスト。

研究テーマ B「クラウドソーシングにおける高品質なデータセット収集の方法論」

クラウドソーシングにおいて大規模にデータを収集する際、単純な指示を出して依頼するだけでは品質の高い質問を集めることは困難である。そこで 2021 年度は、クラウドソーシングにおいてどのようなインターフェース・作問指示・報酬の仕組みが質問の良し悪しに寄与するかを明らかにし、今後効率的に高品質なデータを集めるための知見を蓄積することを目指した。具体的には、質問収集手法を収集した質問の難易度（人間とシステムの正答率の差分）の観点から比較することで、質問収集においてどのような介入が有効か調査した（図 3）。4つの収集プロセスについて 1000 問以上ずつ収集し、データセットとして公開した。結果として、作業者に繰り返しのフィードバックを与えることが有用であることがわかったが、これは作業者相互よりも専門家によって行われることが望ましいことが明らかになった。

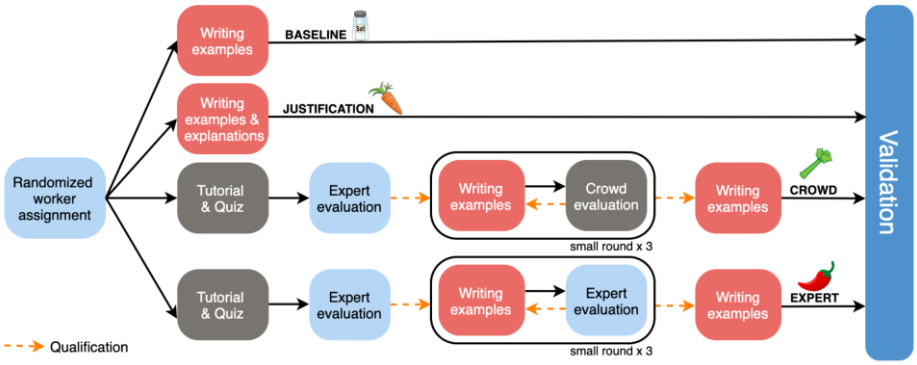

図 3 クラウドソーシングにおいて比較されたデータ収集手法。

もう一つの問題として、作問に用いる文章が結果として執筆される質問にどのような影響を与えるのかが明らかでなかった。これが不明瞭なままだと、望ましい難易度・能力について質問を作成する上でどのような文章を用いればよいかが自明ではなく、効率のよいデータ収集の妨げになる可能性があった。そこで複数の異なるソースの文章を利用し、文章の読みやすさとドメインごとの違いが収集される質問の難易度・要求される能力の違いにどのような変化をもたらすか分析した。結果として、文章の難易度は質問の難易度とほとんど相関しないこと、異なる文章のドメインを利用することで多様な性質の質問を集めることが可能なことが示唆された。

以上の研究はニューヨーク大学の研究グループとの共同研究として行われ、計算言語学分野の難関国際会議である ACL 2021 と ACL 2022 に論文が採択されて発表を行った。

研究テーマ C.「説明性の高いベンチマーク用データセットの構築」

以下、構築した中で代表的なもの 2 件について述べる。

1. 既存の常識推論のデータセットは一般的な状況下でも成り立つ常識的知識を問うことが多く、文脈に依存するような柔軟な推論を評価することが十分ではなかった。そこで既存のデータセットにおける短い物語文を利用し、その結末となる文を複数パターン用意した。それぞれの結末が答えになるような質問文を作成することで、質問という文脈に依存しながらもすべての選択肢が正答になりうるようなタスクを設計した（図 4）。およそ 4,500 問の質問を作成したところ、テストセットにおいて教師なしの状況下ではその時点でもっとも高性能なシステムでも 6 割ほどの正解率しか達成できなかった（人間の正解率は 9 割程度）。本論文は計算言語学分野の難関国際会議である COLING 2022 に採録されて発表を行った。

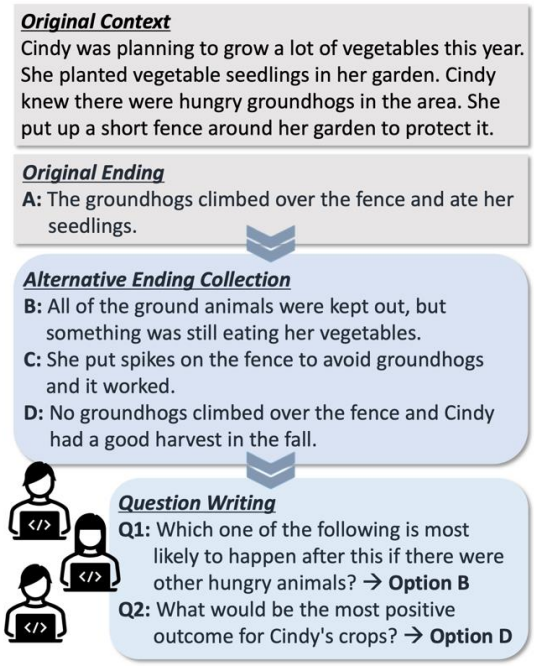

図 4 文脈的な理解が要求されるタスクの例。与えられた文章に続く異なる結末を複数作成して選択肢（A–D）にし、異なる質問ごとに文脈を与えて正答の選択肢を変化させる。

2. 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。そこで本データセットでは、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についての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り、とくに誤答となる選択肢がなぜ誤答なのかについての根拠を理解していると言えないことがわかった。本論文は計算言語学分野の難関国際会議である EMNLP2023 に採択された。

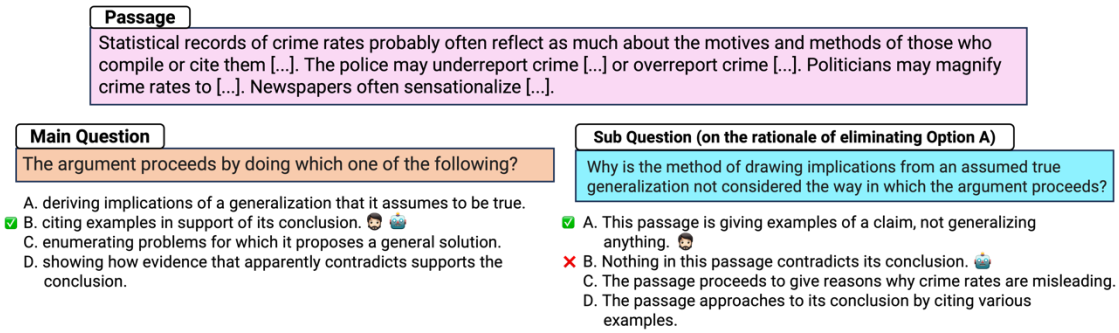

図 5 選択問題で機械が自らの判断の根拠を正確に指摘できない例。選択肢ごとにそれが正答ないし誤答となる根拠を収集し、補助問題の選択肢とする。機械は主問題に正答できても補助問題を容易に誤答することを発見した。

3. 今後の展開

大規模言語モデルの登場によって、言語理解のベンチマーク環境は様変わりしている。しかし評価の妥当性を左右する構成要素は本質的には変わらず、今後も本研究で整備したロードマップ（チェックリスト）に沿って評価方法を改善していく研究を継続することが望ましいと考えている。現状のモデルが未だに実現できていない能力は未だに多く、人間のような振る舞いを達成できているとは言い難い。弱点を可視化しながらさらなるシステムの発展を促進できるようなベンチマークの開発に取り組んでいく必要がある。こうした成果は、大規模言語モデルが実応用や社会実装される際の評価データ・手法を提供するものとしての価値があり、社会における利用が拡大しつつあるなかにおいてはすぐさま活用が期待されるものである。技術発展が急速であることから、遅くとも 2 年から 3 年をひとつのスパンとして開発と社会実装の好循環を繰り返していくことが望ましい。

4. 自己評価

当初の研究計画において想定していた研究項目は何らかの形でいずれも複数件の難関国際会議論文として発表しながら成果を公開しており、一定の達成がなされたと捉えている。こうした成果を支えていたのはさきがけの財源で雇用されたリサーチ・アシスタント（常時 2-3 名程度）であり、柔軟な研究実施体制で進めることができた。また、研究費執行は適切に行われた。研究成果の科学技術及び社会・経済への波及効果については、とくに今後数年間は国内においても日本語に強い大規模言語モデルの開発が進むなか、本研究で基礎づけられた評価手法をそうした開発へ還元することを通して社会・経済へポジティブな効果をもたらすことを期待している。

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：13件

1. Akira Kawabata, Saku Sugawara. Evaluating the Rationale Understanding of Critical Reasoning in Logical Reading Comprehension, *Proceedings of the 2023 Meeting of Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp.116-143, 2023.
論理的な議論の前提や帰結を推論する必要がある選択式読解問題について、既存のデータセットでは正答・誤答から得られる情報が限られており、たとえ正解できたとしてもどのような推論を経て正解に至ったのか説明性が低かった。そこで、選択肢のそれぞれについてそれが正答ないし誤答になる根拠をクラウドソーシングで記述させ、その根拠自体が選択肢となる補助タスクを作成した。
2. Saku Sugawara, Nikita Nangia, Alex Warstadt, Samuel R. Bowman. What Makes Reading Comprehension Questions Difficult? *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pp.6951-6971, 2022.
クラウドソーシングによる文章読解問題の収集では、作問に用いる文章が結果として執筆される質問にどのような影響を与えるのかが明らかでなかった。そこで複数の異なるソースの文章を利用し、文章の読みやすさが収集される質問の難易度・要求される能力の違いに与える変化を分析した。結果として、文章の難易度は質問の難易度とほとんど相関しないことが可能なことが示唆された。

3. 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)*, pp.1592-1612, 2021.
読解の定義とその評価手法という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。

(2)特許出願

研究期間全出願件数： 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)

特になし