氏　　名　HO THI XANH

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2528 号

学位授与の日付　2024 年 9 月 27 日

学位授与の要件　複合科学研究科　情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Exploring and Evaluating the Reasoning Steps in Multi-hop Machine Reading Comprehension

論 文 審 査 委 員　主　査　相澤　彰子
情報学コース　教授
武田　英明
情報学コース　教授
井上　克巳
情報学コース　教授
菅原　朔
情報学コース　助授
Florian Boudin
ナント大学　准教授

Summary of Doctoral Thesis

Name in full: HO THI XANH

Title: Exploring and Evaluating the Reasoning Steps in Multi-hop Machine Reading Comprehension

Machine reading comprehension (MRC) aims to teach computers to read and understand unstructured text to answer a given question. Various directions have emerged from the foundational structure of the MRC task, including multi-hop MRC and conversational MRC. We note that multi-hop MRC requires models to undergo multiple steps to derive the ultimate answer, making it inherently involve reasoning steps within the question-answering (QA) process. This characteristic offers valuable avenues for exploration and model evaluation. Consequently, this thesis endeavors to explore and assess the reasoning steps within multi-hop MRC.

There are two primary issues associated with the multi-hop MRC task. The first pertains to reasoning shortcuts, wherein a machine learning model can answer questions without engaging in multi-hop reasoning. Consequently, the model may lack a genuine understanding of the question's context and underlying concepts, thereby constraining its capacity to offer meaningful responses. This compromise undermines the reliability and effectiveness of the model, particularly in tasks demanding comprehensive understanding and reasoning. The second concern pertains to the absence of explanatory information. Such information is crucial for comprehending the rationale behind models' predictions for specific questions. However, the majority of existing datasets lack comprehensive tasks designed to provide explanations for predicted answers. While there is one dataset available (the R4C dataset), its limited size precludes its utility as a multi-hop dataset with comprehensive explanations for training end-to-end systems.

Our thesis mainly focuses on exploring and evaluating the reasoning steps in multi-hop MRC; it includes seven chapters. In Chapter 1, we discuss two prevalent issues: reasoning shortcuts and missing explanations, which serve as the primary motivations for our research. In Chapter 2, we provide an overview of the multi-hop MRC task, encompassing the definition of reasoning steps and various representations thereof. Additionally, we explore the intricacies of reasoning shortcuts within multi-hop MRC, delineating their definition and existing methods for detection. In Chapter 3, we construct a new multi-hop dataset, 2WikiMultiHopQA, which is a large and high-quality dataset providing comprehensive explanations for predictions. Through experiments, we demonstrate that our dataset poses a challenge for multi-hop models, ensuring that multi-hop reasoning is required. In Chapter 4, we analyze the effectiveness of the reasoning steps in the question-answering process in the triple-form. We consider three

aspects in our analyses: QA performance, reasoning shortcuts, and robustness. Our analyses show that the reasoning steps can help improve QA performance and reduce reasoning shortcuts, but they do not make the models more robust against adversarial questions. In Chapter 5, we enhance the existing reasoning steps (triple-form) in comparison questions, a sub-type of multi-hop questions by introducing three probing tasks: extraction, reasoning, and robustness. The current reasoning steps in triple-form are limited by the fixed schedule of Wikidata; therefore, we represent the reasoning steps in this chapter in a sub-QA format. We then evaluate top-performing models on our dataset. The results reveal that the models may not possess the ability to subtract two dates even when fine-tuned on our dataset. Additionally, we analyze the effectiveness of the reasoning steps in the sub-QA format and find that they can help to improve QA performance. In Chapter 6, we examine the behaviors of large language models (LLMs) across three stages in the QA process: question decomposition, subproblem solving, and composition. To comprehensively investigate LLMs, we consider three different scenarios for the input context: without context, with unstructured context, and with structured context. Our main finding is that the current LLMs fail to decompose complex questions into sub-questions, which is key in the QA process. In Chapter 7, we conclude our research and discuss about potential future works. The proposed datasets and analyses in this thesis contribute to building a more explainable QA system in the future. They also contribute to a better understanding of the models in the QA process when dealing with multi-hop questions.

（別紙 Separate Sheet）

Results of the Doctoral Thesis Defense

博士論文審査結果

Name in Full
氏　　名　HO THI XANH

Ｔｉｔｌｅ
論文題目　Exploring and Evaluating the Reasoning Steps in Multi-hop Machine Reading Comprehension

本学位論文は、「Exploring and Evaluating the Reasoning Steps in Multi-hop Machine Reading Comprehension（多段階機械読解における推論ステップの分析と評価）」と題し、全 7 章で構成されている。

第 1 章では、研究の動機を述べている。言語モデルの文章読解能力の評価に用いられる機械読解タスクについて、表層的な手がかりで回答に到達してしまうショートカット、および回答結果に対する説明の欠落、の２つの問題点を指摘し、これらの問題点に対する分析や評価を研究目標として設定している。

第 2 章では、多段階機械読解タスクの概要を説明し、推論ステップの定義を示すとともに、関連研究を俯瞰している。

第 3 章では、複数文書にまたがる推論ステップを明示的に正解として付与した多段階機械読解タスクのデータセット 2WikiMultiHopQA を設計するとともに、大規模なデータセットを構築して既存の言語モデルを分析して、多段階推論における問題点を明らかにしている。

第 4 章では、与えられた文章からトリプル形式で知識を抽出して推論に用いるためのデータセットを構築している。このトリプル形式は、２つのエンティティとその関係で知識を表現する一般的な評点形式である。また新たに多段階推論モデルを設計し、読解質問に対する正答率、推論におけるショートカットの有無、頑健性の 3 つの観点から分析を行っている。実験結果から、トリプル形式で表現された推論ステップの学習が、モデルの正答率の向上と推論ショートカットの削減に有効であること、一方で敵対的な質問に対するモデルの頑健性を高めることにはつながらないことを結論付けている。

第 5 章では、多段階機械読解を必要とする質問形式として一般的な、「比較質問」に焦点をあてて、与えられた比較質問を分解して得られる複数のサブ質問を、推論ステップとして明示的に付与したデータセットを設計している。構築したデータセットを用いて、これまでの実験でもっともすぐれた性能を示した既存モデルを分析し、サブ質問形式による推論ステップによる学習が正答率の向上に寄与することを示している。また、2 つの日付の差を求めるなど特殊な演算能力を要するサブ質問について、モデルの能力に限界があることを示している。

第 6 章では、多段階機械読解を、「設問の部分問題への分解」、「部分問題への回答」、「回答の統合」の 3 つのプロセスに分けて、それぞれのプロセスにおける大規模言語モデルの動作を検証している。実験では、プロンプトとして異なる条件を設定して詳細な分析を実施した上で、現在の大規模言語モデルは特に設問の部分問題への分解に課題があると結論付けている。

第 7 章では、研究全体を通して提案した多段階機械読解タスクの分析手法、構築したデータセットの有効性、モデルの分析結果と限界などを俯瞰し、さらに今後の可能性について考察している。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究が、言語モデルの文章読解能力の評価に用いられる機械読解タスクについて、特に複数文書にまたがる推論を必要とする多段階推論に焦点を当てて、新しい着想に基づくデータセットの構築を行うことで、モデルの能力を評価および解明するものであることが評価された。

以上を要するに本学位論文は、多段階機械読解における推論ステップの究明と評価に貢献するものであり、研究分野の発展に貢献しているという点で学術的価値が大きい。また、本学位論文の成果は、情報学コースが要件として定めるトップ国際会議を含む査読付き国際会議論文３件として発表され、社会的な評価も得ている。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。