

論文内容の要旨

申請者氏名 坂井 優介

Large language models (LLMs) have become foundational resources in the natural language processing field. The capabilities of LLMs in inference, natural language understanding (NLU), and their pre-trained knowledge are widely utilized in various applications. To fully leverage these capabilities, it is essential to properly evaluate them through appropriate benchmarks. However, it is very challenging to evaluate LLMs accurately. For example, when evaluating an LLM's inference capabilities to infer unseen facts from seen ones, it may rely on memorized knowledge from pre-training rather than pure inference, due to the pre-training datasets being curated from web data, which can lead to contamination with related knowledge. Therefore, traditional evaluation frameworks for inference capabilities struggle to capture the pure inference capability of LLMs, independent of their pre-trained knowledge. Furthermore, when evaluating the NLU capabilities of LLMs, we should take into account the stability in outputs for various input prompts. Finally, constructing evaluation datasets involves significant costs of financial, time, and human labor. Especially when focusing on multilingual datasets, there is a limitation in the availability of annotators, so most datasets are machine-translated from English. However, due to cultural differences and noise introduced by translation, these evaluations are unsuitable for language-specific evaluation. This dissertation tackles these challenges. First, we evaluate the pure inference capability of LLMs using the knowledge graph completion task. We create synthetic datasets by considering the structure of knowledge graphs to avoid data contamination from pre-trained knowledge. Next, we reveal that the robust evaluation of LLMs should consider the variance of outputs for various input prompts. We propose evaluation datasets and methods that take this variance into account. Lastly, we utilize LLMs to replace human labor in the evaluation dataset creation process, proposing efficient methods for constructing multilingual evaluation datasets for LLMs. This dissertation discusses efficient, effective, and sustainable evaluation methods for LLMs.

論文審査結果の要旨

申請者氏名 坂井 優介

Large language models (LLMs) have become foundational resources in the natural language processing field. The capabilities of LLMs in inference, natural language understanding (NLU), and their pre-trained knowledge are widely utilized in various applications. To fully leverage these capabilities, it is essential to properly evaluate them through appropriate benchmarks. This dissertation tackles the challenges to evaluate LLMs accurately; The first method evaluates the pure inference capability of LLMs using the knowledge graph completion task. In this method, synthetic datasets are created by considering the structure of knowledge graphs to avoid data contamination from pre-trained knowledge. The second method reveals that the robust evaluation of LLMs should consider the variance of outputs for various input prompts. In this work, evaluation datasets and methods are proposed that take the variance into account. Lastly, the third method utilizes LLMs to replace human labor in the evaluation dataset creation process, proposing efficient methods for constructing multilingual evaluation datasets for LLMs. This dissertation discusses efficient, effective, and sustainable evaluation methods for LLMs.

The three studies are published as one high quality peer-reviewed journal paper, two peer-reviewed international conference papers and one report at a domestic conference. The research would have an impact not only to natural language processing, in particular, in the research on large language models and language understanding, but to the relevant fields of machine learning, e.g., text generation. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.