Knowledge Editing of Large Language Models Unconstrained by Word Order

Ryoma Ishigaki[1] Jundai Suzuki[1] Masaki Shuzo[1] Eisaku Maeda[1]
[1]Tokyo Denki University
{24amj02@ms, 24amj20@ms, shuzo@mail, maeda.e@mail}dendai.ac.jp

掲載号の情報

32 巻 4 号 pp. 1062-1102.
doi: https://doi.org/10.5715/jnlp.32.1062

概要

Large Language Models (LLMs) possess potentially extensive knowledge; however, because their internal processing operates as a black box, directly editing the knowledge embedded within the LLMs is difficult. To address this issue, a method known as local-modification-based knowledge editing has been developed. This method identifies the “knowledge neurons” that encode the target knowledge and adjusts the parameters associated with these neurons to update the stored information. Knowledge neurons are identified by masking the object (o) from sentences representing relational triplets (s, r, o), with the LLM predicting the masked element, and observing its internal activation patterns during the prediction. When the architecture is decoder-based, the predicted object (o) must be located at the end of the sentence. Previous local-modification-based knowledge-editing methods for decoder-based models have assumed subject-verb-object languages and faced challenges when applied to subject-object-verb languages such as Japanese. In this study, we propose a knowledge-editing method that eliminates the need for word order constraints by converting the input used to identify knowledge neurons into a question, where object (o) is the answer. We conducted validation experiments using a known-facts dataset and confirmed that the proposed method is effective for Japanese language, which is a non- subject-verb-object language.