| | | | |
|---|---|---|---|
| 氏名 | Le Khang Nguyen | | |
| 学位の種類 | 博士（情報科学） | | |
| 学位記番号 | 博情第 575 号 | | |
| 学位授与年月日 | 令和 8 年 3 月 25 日 | | |
| 論文題目 | Towards Quality, Reliability, and Efficiency in Large Multilingual and Multimodal Models via Answer-Prefix Generation and Efficient Inference | | |
| 論文審査委員 | Nguyen Le Minh | JAIST | Professor |
| | SHIRAI Kiyoaki | JAIST | Professor |
| | HASEGAWA Shinobu | JAIST | Professor |
| | Francesca Toni | Imperial College London | Professor |
| | Naoya Inoue | JAIST | Associate Professor |

論文の内容の要旨

Large Language Models (LLMs) have achieved remarkable success in natural language processing tasks, particularly in question-answering (QA). However, their integration into real-world systems is often hindered by two key challenges: first, the need for structured, concise answers and reliable confidence estimation to support aggregation across multiple reasoning paths; and second, the large model size and computational cost, which limit deployment efficiency.

To address these challenges, this thesis presents two complementary contributions. First, we introduce **Anspre**, a structured answer generation framework that guides LLMs to produce concise answers with reliable confidence scores. Anspre improves aggregation across retrievals and reasoning chains, enhances answer quality, and generalizes across multilingual and vision-language QA tasks. Extensive experiments on open-domain, multilingual, and visual QA benchmarks demonstrate that Anspre significantly improves Exact Match (EM) and F1 scores while providing well-calibrated confidence estimates.

Second, we investigate the behavior of pruning techniques across varying sparsity levels and identify key findings that inform optimal pruning strategies. Building on these insights, we propose **OptiPrune**, a method that dynamically selects the most suitable pruning approach for each sparsity regime. Empirical evaluation shows that OptiPrune consistently outperforms state-of-the-art pruning methods across multiple architectures, benchmarks, and

language-specific calibrations, enabling efficient deployment without significant performance degradation.

Together, Anspre and OptiPrune advance LLMs toward high-quality, reliable, and deployment-efficient systems, addressing critical gaps in structured reasoning, confidence estimation, and model compression.

**Keywords:** Model Reliability, Model Efficiency, Retrieval-Augmented Generation, Model Pruning, Question-Answering, Large-Language-Model

**論文審査の結果の要旨**

This **dissertation** addresses the critical challenges in the deployment of large language models for question answering: reliable structured answer generation with confidence estimation, and computational efficiency under resource constraints. The problem formulation is timely and well motivated, particularly in retrieval-augmented and multi-path reasoning settings. **The first** contribution, Anspre, introduces a structured answer generation framework that produces concise answers accompanied by well-calibrated confidence scores. The proposed method effectively improves aggregation across multiple retrievals and reasoning chains, and its strong generalization across multilingual and vision–language question-answering tasks is convincingly demonstrated through extensive experiments, yielding consistent improvements in exact match (EM) and F1 scores. **The second** contribution presents a systematic analysis of pruning behavior across different sparsity levels and proposes OptiPrune, a dynamic pruning strategy that adapts to varying sparsity regimes. Empirical results show that OptiPrune consistently outperforms state-of-the-art pruning methods across model architectures, benchmarks, and language settings, enabling more efficient deployment with minimal performance degradation. **The third contribution** investigates methods for improving inference speed in large language models via speculative decoding, exploiting statistical measures such as perplexity to guide decoding decisions. The contributions are novel and impactful, and the work is of a quality suitable for publication at top-tier conferences in natural language processing and machine learning, including ACL, ICML, ECAI, and COLING. In addition, the candidate received the Best Paper Award at JSAI and achieved Best System performance in the COLIEE 2025 competition.

Overall, the thesis is technically sound, well-executed, and experimentally rigorous. This is an excellent dissertation and we approve awarding a doctoral degree to Le Khang Nguyen.