氏　　名	Nguyen Dieu Hien		
学位の種類	博士（情報科学）		
学位記番号	博情第 569 号		
学位授与年月日	令和 8 年 3 月 25 日		
論文題目	Language-Specific Compact Large Language Models with Precise Knowledge through Language-aware Pruning and Low-resource Retrieval		
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 advances in reasoning, understanding, and generation across a wide range of natural language processing (NLP) tasks. However, their large scale imposes substantial computational and memory costs, posing challenges for efficient deployment, particularly in low-resource settings. Simultaneously, while retrieval-augmented methods have enhanced the factual grounding and interpretability of question answering (QA) systems, their effectiveness in low-resource languages remains limited by data scarcity and the absence of high-quality retrieval frameworks.

This thesis addresses these challenges through two complementary research directions. First, it introduces **LangCompress**, a language-aware model compression framework that integrates self-supervised instruction generation and vocabulary optimization to adapt LLMs for specific languages. LangCompress improves efficiency while preserving linguistic competence and can be seamlessly combined with existing pruning and quantization methods. Second, the thesis proposes a graph-based retrieval framework for multi-hop question answering that leverages Wikipedia's hyperlink structure to identify semantically connected evidence across documents. This framework enables efficient retrieval and reasoning in low-resource languages. To support this line of research, a generalizable dataset construction framework is developed, resulting in *VIMQA*, a Vietnamese multi-hop QA dataset designed to evaluate explainable and complex reasoning grounded in Wikipedia evidence.

Together, these contributions advance the development of retrieval- and compression-aware LLM systems that are efficient, interpretable, and inclusive across diverse linguistic environments.

Keywords: low-resource language, quantization, pruning, large language model, retrieval, question-answering

論文審査の結果の要旨

This **dissertation** addresses two important and timely challenges in the deployment of large language models: computational efficiency under resource constraints and effective retrieval-augmented reasoning for low-resource languages. The problem formulation is well motivated, particularly in highlighting the gap between recent advances in large language model capabilities and their limited accessibility across diverse linguistic settings. The first contribution, **LangCompress**, presents a language-aware compression framework that combines self-supervised instruction generation with vocabulary optimization. The proposed approach is technically sound and practically relevant, demonstrating that efficiency gains can be achieved while preserving language-specific competence, and that the method integrates effectively with existing pruning and quantization techniques. The second contribution introduces a graph-based multi-hop retrieval framework that leverages Wikipedia's hyperlink structure to support efficient evidence discovery and reasoning in low-resource languages. The accompanying dataset construction framework, together with the resulting **VIMQA** dataset, constitutes a valuable resource that enables systematic evaluation of explainable multi-hop reasoning grounded in real-world evidence. The **dissertation** is well structured and methodologically rigorous, addressing an underexplored yet critical area of large language model research. The contributions are novel and impactful, and are of a quality suitable for publication in a leading journal (i.e., *Information Processing & Management*) and reputable international conferences in natural language processing (e.g., AACL and PRICAI), particularly those focused on multilinguality, model efficiency, and retrieval-augmented reasoning.
Overall, the candidate shows an excellent dissertation, and we approve awarding a doctoral degree to **Nguyen Dieu Hien**.