

Title	Towards End-to-end Wikipedia-based Open-domain Question-Answering Systems for Single-hop and Multi-hop Questions in Low-resource Languages
Author(s)	Nguyen, Hien Dieu
Citation	
Issue Date	2023-03
Type	Thesis or Dissertation
Text version	author
URL	<a href="http://hdl.handle.net/10119/18307">http://hdl.handle.net/10119/18307</a>
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 修士(情報科学)

Towards End-to-end Wikipedia-based Open-domain Question-Answering Systems for Single-hop and Multi-hop Questions in Low-resource Languages

2110064 NGUYEN, Hien Dieu

Open-domain Question-Answering (QA) task involves using a large knowledge base, such as Wikipedia, to answer a given question. This is often done using a two-stage framework that includes a Retriever and a Reader. The performance of the QA system is greatly influenced by the effectiveness of the Retriever stage. Despite being the first language of roughly a hundred million people worldwide, Vietnamese remains a low-resource language with a scarcity of research on QA systems. No efficient Vietnamese Open-domain QA system for single and multi-hop questions has been studied. Although resource-rich languages like English witnessed many advancements in Open-domain QA, these methods often suffer from low data situations. The objective of this study is to design an efficient Open-domain QA system utilizing the Wikipedia knowledge base, which can handle both single and multi-hop questions. The proposed system is robust when applied to low-resource languages. This research was initially conducted in the Vietnamese language, but the methodology can be generalized to other low-resource languages. This study proposes ViWiQA, an efficient Vietnamese Open-domain QA system over the Wikipedia knowledge base, with two novel retriever methods for single-hop and multi-hop questions. ViWiQA can be effectively trained with low data and significantly outperforms Lucene-BM25 and Dense Passage Retrieval when adapted to Vietnamese datasets. ViWiQA demonstrates a significant improvement of 20% in single-hop retrieval accuracy compared to Lucene-BM25 and sets a new standard in single-hop and multi-hop Vietnamese Open-domain QA benchmarks.