การพัฒนาระบบสืบค้นและตอบคำถามอัตโนมัติสำหรับเอกสารภาษามือไทยด้วยเทคโนโลยี Local RAG (Retrieval-Augmented Generation)

กรธวัช สะสม[1],ฉัตรธำรงณ์ ฮั่นพงษ์กุล[1], อลงกต กองมณี[1], ปวีณ เขื่อนแก้ว[1] และ สมนึก สินธุปวน[1,*]
Korntawat Sasom[1],Chatamrong Hanponggul[1],Alongkot Gongmanee[1], Paweem Khoenkaw[1] and Somnuek Sinthupuan[1,*]
[1]*สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้ 50290 ประเทศไทย*
*ผู้นิพนธ์ประสานงาน: สมนึก สินธุปวน อีเมล: somnuk@mju.ac.th

บทคัดย่อ:

การวิจัยนี้มีวัตถุประสงค์เพื่อ (1) พัฒนา ระบบสืบค้นและตอบคำถามอัตโนมัติแบบออฟไลน์ด้วยแนวคิด Retrieval-Augmented Generation (Local RAG) สำหรับเอกสาร PDF ภาษาไทยเกี่ยวกับภาษามือไทยและผู้บกพร่องทางการได้ยิน โดยเน้นความเป็นส่วนตัวผ่านการประมวลผลบนเครื่องผู้ใช้ (2) จัดเตรียมคลังความรู้ภายในเครื่องจากเอกสารดังกล่าวด้วยการสกัดและทำความสะอาดข้อความจาก PDF และจัดเก็บพร้อมเมทาดาทาเพื่อการอ้างอิง และ (3) ประเมินประสิทธิภาพของระบบด้วยชุดคำถามอ้างอิง

ระบบพัฒนาด้วย Python และ Ollama ใช้ gemma3:4b สำหรับสร้างคำตอบ และ mxbai-embed-large สำหรับสร้าง embeddings เพื่อค้นคืนบริบทจากคลังความรู้ การประเมินใช้ชุดคำถาม ground truth 500 ข้อ ให้คะแนนความถูกต้องด้วยวิธี LLM-as-a-Judge ระดับ 1–5 คะแนน ผลการทดลองพบว่าระบบได้คะแนนเฉลี่ย 2.62/5.00 โดยตอบคำถามข้อเท็จจริงพื้นฐานได้ระดับปานกลาง แต่ยังมีข้อจำกัดกับคำศัพท์เฉพาะทางภาษามือไทยและคำถามซับซ้อน ทำให้บางกรณีค้นคืนบริบทไม่ตรงหรือเกิดการปะปนของเนื้อหา ข้อเสนอแนะคือควรปรับปรุงการทำความสะอาดข้อความจาก PDF และปรับกลยุทธ์การแบ่งส่วนข้อความ (chunking) และการตั้งค่า embeddings/การจัดอันดับผลสืบค้นให้เหมาะกับภาษาไทยและโครงสร้างเอกสาร เพื่อเพิ่มความแม่นยำของระบบ

คำสำคัญ : Local RAG ระบบตอบคำถามอัตโนมัติ การสืบค้นข้อมูล เอ็มเบดดิงส์ ภาษามือไทย เอกสาร PDF ภาษาไทย

Abstract:

This study aims to (1) develop a privacy-preserving offline Retrieval-Augmented Generation (Local RAG) system for Thai-language PDF documents about Thai Sign Language and individuals with hearing impairments, (2) construct a local knowledge base by extracting and cleaning PDF text and storing it with traceable metadata, and (3) evaluate system performance using a ground-truth QA set.

The system was implemented in Python with Ollama, using gemma3:4b for answer generation and mxbai-embed-large for embedding-based contextual retrieval. Evaluation on 500 ground-truth questions employed an LLM-as-a-Judge protocol with a 1–5 rating scale, yielding an average score of 2.62/5.00. The system performs moderately on basic factual queries but struggles with domain-specific terminology and complex questions, sometimes retrieving partially relevant context or producing mixed-context answers. Future work should improve PDF text cleaning and optimize chunking and embedding/ranking configurations to better match Thai linguistic characteristics and document structure.

Keywords: Local RAG, question answering, information retrieval, embeddings, Thai Sign Language, Thai PDF documents