การสร้างกราฟความรู้เชิงปฏิบัติโดยใช้แบบจำลองภาษาขนาดใหญ่

Constructing Practical Knowledge Graphs by Large Language Model

นาย ภคิน ขวัญสุวรรณ
คณะเทคโนโลยีสารสนเทศ
สาขาวิทยาศาสตรบัณฑิต
(เทคโนโลยีสารสนเทศ)
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
2463110094@tni.ac.th

อัดนา เซนโต๊ะ
คณะวิศวกรรมศาสตร์ | วศ.บ. วิศวกรรมคอมพิวเตอร์และ
ปัญญาประดิษฐ์ (CE)
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
*Corresponding author: adna@tni.ac.th

บทคัดย่อ **—คุณภาพของโครงสร้างกราฟเชิงความรู้ (Knowledge Graph) เป็นส่วนสำคัญในแอปพลิเคชันที่ขับเคลื่อนด้วยข้อมูล และโดยทั่วไปแล้วชุดข้อมูลมักจะต้องผ่านการประเมินคุณภาพก่อนนำไปใช้งาน แต่อย่างไรก็ตามการประเมินคุณภาพของคลังชุดข้อมูลขนาดใหญ่มักพบปัญหา เช่น ความซับซ้อนของข้อมูล เวลาของการประเมิน การค้นหาข้อมูล และความล่าช้าในการติดตามข้อมูลที่เปลี่ยนแปลงตลอดเวลา ซึ่งส่งผลกระทบโดยตรงต่อประสิทธิภาพของระบบ ดังนั้นบทความวิจัยนี้จึงนำเสนอการสร้างข้อมูลเชิงปฏิบัติสำหรับโครงสร้างกราฟเชิงความรู้ด้วยวิธีการจัดรูปแบบโดยโมเดลภาษาขนาดใหญ่ (LLM) ซึ่งประกอบไปด้วย การออกแบบโครงสร้างข้อมูล (ontology) การสกัดข้อมูลเพื่อสร้างเอนทิตี้ตามโครงสร้าง และสุดท้ายเป็นการเชื่อมโยงแต่ละเอนทิตี้เข้าหากันด้วยความสัมพันธ์ระหว่างเอนทิตี้โดย node2Vec จากนั้นจะเป็นขั้นตอนการประเมินผ่านเมตริกสำคัญ 3 ด้าน ได้แก่ ความถูกต้อง ความสมบูรณ์ และความสอดคล้อง โดยทดสอบกับชุดข้อมูลกราฟความรู้มาตรฐานและเปรียบเทียบกับวิธีการทำกราฟความรู้แบบอื่น ๆ จากผลการทดลองพบว่าแนวทางที่นำเสนอนี้ช่วยให้สามารถกรองข้อมูลจากคลังความรู้ได้ง่ายขึ้น ซึ่งจากผลการทดลองพบว่าโมเดลที่นำเสนอนี้มีคุณภาพของชุดมูลสูงถึงเฉลี่ยร้อยละ 86**

คำสำคัญ — กราฟความรู้, การประเมินคุณภาพข้อมูล , โมเดลภาษาขนาดใหญ่, การประเมินด้วยโมเดล , ความสอดคล้องของข้อมูล

ABSTRACT **— The quality of a knowledge graph (KG) is a crucial aspect of data-driven applications, and datasets typically need to undergo quality assessment before deployment. However, evaluating the quality of large datasets often encounters problems such as data complexity, evaluation time, data searching, and delays in tracking constantly changing data, directly impacting system performance. Therefore, this research presents a practical approach for constructing knowledge graphs using a Large Language Model (LLM) formatting method. This involves designing the ontology, extracting data to create entities based on the structure, and finally linking these entities using node2Vec relationships. Evaluation is then conducted using three key metrics: accuracy, completeness, and consistency. The results are tested on a standard knowledge graph dataset compared to other knowledge graph methods. The experimental results show that the proposed approach simplifies data filtering from the knowledge repository, achieving an average dataset quality of 86%.**

Keywords — Knowledge graphs, data quality assessment, large language model, models evaluation, data consistency.

I. บทนำ

ในปัจจุบันกราฟความรู้ (Knowledge Graph - KG) ได้กลายเป็นเครื่องมือสำคัญในการจัดเก็บและนำเสนอข้อมูลโดยการเชื่อมโยงโหนดหรือเอนทิตี (Entities) ด้วยความสัมพันธ์ (Relationships) ของข้อมูลนั้นเข้าด้วยกัน เทคโนโลยีนี้ถูกนำไปใช้งานอย่างแพร่หลายในระบบสืบค้นข้อมูล ระบบตอบคำถาม และระบบแนะนำในหลากหลายโดเมน, โดยเฉพาะในด้านการแพทย์และสุขภาพเนื่องจากกราฟความรู้ช่วยในการสร้างฐานข้อมูลที่เชื่อมโยงระหว่างโรค อาการของโรค ยา และข้อมูลที่ถูกต้องจากผู้เชี่ยวชาญได้อย่างมีประสิทธิภาพ แม้ว่า โมเดลภาษาขนาดใหญ่ (Large Language Models - LLMs) จะมีความสามารถในการประมวลผลภาษาธรรมชาติหรือภาษามนุษย์ได้อย่างยอดเยี่ยม แต่ยังคงประสบปัญหาสำคัญคือการสร้างข้อมูลที่ไม่มีอยู่จริงหรืออาการหลอน (Hallucinations) เนื่องจาการทำงานของโมเดลภาษาขนาดใหญ่นั้นคือการเดาสุ่มจากโทเคนความน่าจะเป็น ดังนั้นจึงไม่สามารถใช้โมเดลภาษาขนาดใหญ่เพียงอย่างเดียวกับทุกงานได้โดยเฉพาะในงานเฉพาะทางที่ต้องการความแม่นยำสูง เช่น การแพทย์แผนจีน (TCM) หรือโภชนาการ [1], [2] การนำกราฟความรู้ไปประยุกต์ใช้เข้ากับเทคโนโลยี Retrieval-Augmented Generation (RAG) จึงกลายเป็นแนวทางที่ได้รับความสนใจอย่างมาก เพราะช่วยให้โมเดลสามารถอ้างอิงข้อมูลจากฐานความรู้ภายนอกที่เชื่อถือได้โดยที่ไม่ต้องเสียทรัพยากรและเวลาในการฝึกฝนโมเดลใหม่ ลดข้อผิดพลาด และเพิ่มความสามารถในการอธิบายเหตุผลของคำตอบ นอกจากนี้ ระบบที่ใช้ KG+RAG ยังมีความยืดหยุ่นสูง โดยสามารถอัปเดตข้อมูลให้เป็นปัจจุบันได้ทันทีผ่านฐานข้อมูลกราฟ (เช่น Neo4j) โดยไม่จำเป็นต้องทำการฝึกฝนโมเดลใหม่ทั้งหมด อย่างไรก็ตามประสิทธิภาพของระบบที่ใช้วิธีการประยุกต์ KG+RAG จะขึ้นอยู่กับคุณภาพของข้อมูลที่มีอยู่ใน

ฐานข้อมูลกราฟซึ่งรวมไปถึงโครงสร้างของข้อมูลเช่น Label หรือ Entity ที่มีความหมายใกล้เคียงกันหรือทับซ้อนกันในบางบริบท (Semantic Overlap) มีผลทำให้โมเดลสับสนและสร้างข้อมูลที่ไม่มีอยู่จริง ดังนั้นการจัดการคุณภาพของกราฟความรู้จึงมีความสำคัญที่ทำให้ระบบมีความน่าเชื่อถือมากขึ้น ความน่าเชื่อถือไม่ได้ขึ้นอยู่กับการวัดผลทางสถิติเพียงอย่างเดียวแต่ยังต้องคำนึงถึงความสมเหตุสมผลและความทันสมัยของคำตอบโดยให้มนุษย์เป็นผู้ตรวจสอบ (Human in the loop) เพื่อให้เหมาะสมกับการใช้งานจริง จากงานวิจัย[3] ได้กล่าวว่าการประเมินกราฟความรู้แบบเดิมจะให้ความสำคัญกับทุกเอสารเท่ากันแต่ในความเป็นจริงผู้ใช้สนใจเพียงแค่บางส่วนของกราฟเท่านั้นและถึงแม้ว่าข้อมูลที่ถูกดึงออกมาจะมีความถูกต้องที่ต่ำเมื่อเทียบกับชุดคำตอบ ดังนั้นหากออกแบบโครงสร้างของกราฟให้ชัดเจนโดยการกำหนด ontology ที่แบ่งแยกอย่างชัดเจนก็จะสามารถกำหนดขอบเขตของการดึงข้อมูลจากกราฟได้

ด้วยเหตุผลที่ได้กล่าวมาข้างต้นนี้ทำให้งานวิจัยนี้มุ่งเน้นไปที่การสกัดข้อมูลจากเอกสาร pdf ซึ่งเป็นประเภทเอกสารที่นิยมใช้กันทั่วไป เช่น คู่มือ รายงาน เอกสารเฉพาะทาง อย่างไรก็ตามแม้ว่า pdf จะถูกใช้และเหมาะกับการเผยแพร่แต่มันไม่เหมาะกับการใช้เชิงคำนวณเพราะโครงสร้างของมันถูกออกแบบเพื่อการอ่านของมนุษย์ส่งผลให้เกิดปัญหาเรื่องข้อมูลกระจัดกระจาย ความไม่สม่ำเสมอของรูปแบบหัวข้อ ดังนั้นเพื่อแก้ปัญหาดังกล่าวงานวิจัยนี้จึงให้ความสำคัญกับการจัดรูปแบบเอกสารและการเตรียมข้อมูลก่อนการสกัด

II. วัตถุประสงค์ของการวิจัย

1) ออกแบบแนวทางการเตรียมข้อมูลแบบโครงสร้างกราฟเชิงความรู้ (Knowledge Graph) โดยโมเดลภาษาขนาดใหญ่ (LLM) ที่ใช้ผ่าน Neo4j เพื่อนำไปใช้กับ RAG

2) ประเมินโมเดลที่ได้ออกแบบและเปรียบเทียบประสิทธิภาพเพื่อการนำไปใช้งาน

III. วิธีดำเนินการวิจัย

การเตรียมข้อมูลถือเป็นขั้นตอนที่มีสำคัญอย่างยิ่งสำหรับการออกแบบโครงสร้างกราฟเชิงความรู้ (Knowledge Graph) เนื่องจากคุณภาพของกราฟเชิงความรู้ส่งผลต่อการวิเคราะห์และการเข้าถึงข้อมูล ซึ่งงานวิจัยนี้ได้พัฒนาการออกแบบระบบข้อมูลที่มีโครงสร้างเพื่อให้การเข้าถึงและวิเคราะห์ข้อมูลให้มีประสิทธิภาพจากงานวิจัย [4]

กระบวนการของงานวิจัยเริ่มต้นจากการรวบรวมและประมวลผลข้อมูลออนไลน์ซึ่งได้คัดกรองข้อมูลสูตรอาหารจากเว็บไซต์ของ “วงใน” ที่รวบรวมสูตรอาหารและวิธีการทำต่างๆ จากข้อมูลที่ได้มาจากเพจของวงในระบุว่ามีมากกว่าแปดพันสูตรในปี 2018 มา 5 เมนูเพื่อให้ง่ายต่อการตรวจสอบ เมื่อรวบรวมสูตรอาหารเรียบร้อยแล้ว ข้อมูลจะถูกจัดเก็บรูปแบบไฟล์ PDF จากนั้นนำข้อมูลไปทำความสะอาดเพื่อให้พร้อมสำหรับขั้นต่อไป ซึ่งมีรายละเอียดดังต่อไปนี้

I.I การออกแบบ Ontology

Ontology คือแผนผังโครงสร้างความสัมพันธ์ของข้อมูลโดยกำหนดว่า มีเอนทิตี (Entities) และความสัมพันธ์ (relation) โดยกำหนดให้เอนทิตี้คือแนวคิดในเชิงนามธรรมที่ถูกนำมาจัดเก็บในรูปแบบโครงสร้างกราฟความรู้โดยจะแทนเอนทิตี้ด้วยโหนด (Node) และมีคุณสมบัติ (properties) ซึ่งเป็นข้อมูลเพิ่มเติมของแต่ละเอนทิตี้นั้นๆ ในฐานข้อมูลกราฟนั้นจะเชื่อมแต่ละเอนทิตี้ด้วย ความสัมพันธ์ (relation) ที่ทำหน้าที่เป็นเส้นเชื่อมระหว่างเอนทิตี้เพื่อให้ข้อมูลเหล่านั้นมีความหมายหรือมีความเกี่ยวข้องกันอย่างไรในเชิงโครงสร้าง

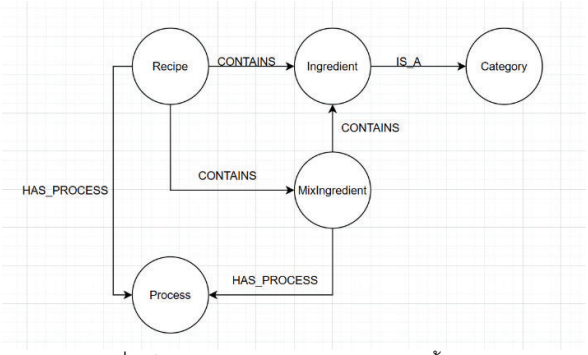
เพื่อให้ครอบคลุมกับระบบถามตอบในอนาคตจึงได้ออกแบบ ontology ที่ครอบคลุมคำถามทั่วไปและคำถามที่ลึกมากขึ้นในงานนี้เอนทิตี้หลักประกอบไปด้วย

- Recipe คือสูตรของอาหารที่จะต้องประกอบไปด้วยวัตถุดิบและวิธีการทำ
- Ingredient คือวัตถุดิบและส่วนผสมที่เป็นส่วนประกอบของเมนูต่างๆ
- MixIngredient คือวัตถุดิบที่เกิดจากการนำ Ingredient มาผ่านกระบวนการบางอย่างเพื่อให้กลายเป็นวัตถุดิบใหม่ กล่าวคือเอนทิตี้นี้เปรียบเสมือนกับสูตรอาหารย่อยที่จะนำไปทำสูตรอาหารอื่นอีกที
- Process คือเอนทิตี้ที่เก็บข้อมูลกระบวนการและขั้นตอนการทำอาหารเอาไว้
- Category คือเอนทิตี้ที่ใช้เพื่อรวมเอนทิตี้อื่นๆ เอาไว้ให้เป็นหมวดหมู่เดียวกันเช่น จัดหมวดหมู่ของเนื้อแดงให้กับ Ingredient หรือจัดหมวดหมู่ประเทศให้กับ Recipe

ความสัมพันธ์ที่ใช้เชื่อมโยงเอนทิตี้เหล่านี้ได้แก่

- CONTAINS ใช้ระบุความเป็นองค์ประกอบเช่น เมนูประกอบด้วยวัตถุดิบ
- IS_A ใช้ในการระบุความเป็นหมวดหมู่
- HAS_PROCESS ใช้ระบุความเกี่ยวข้องของขันตอนและวิธีการทำ

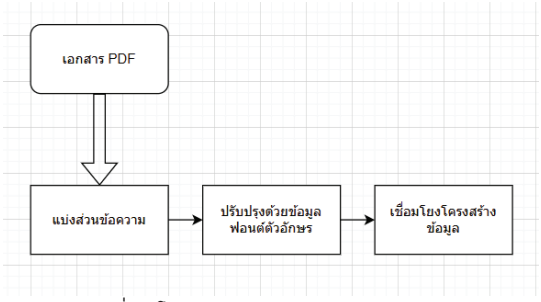
เมื่อกำหนด ontology ให้กับกราฟความรู้ได้แล้ว จากนั้นจะนำข้อมูลที่เตรียมไว้ไปสร้างกราฟความรู้และเก็บเอาไว้ใน neo4j ซึ่งเป็นฐานข้อมูลที่เก็บข้อมูลไว้ในรูปแบบกราฟ โดยมีโครงสร้างดังนี้



รูปที่ I. โครงสร้างความสัมพันธ์ของเอนทิตี้

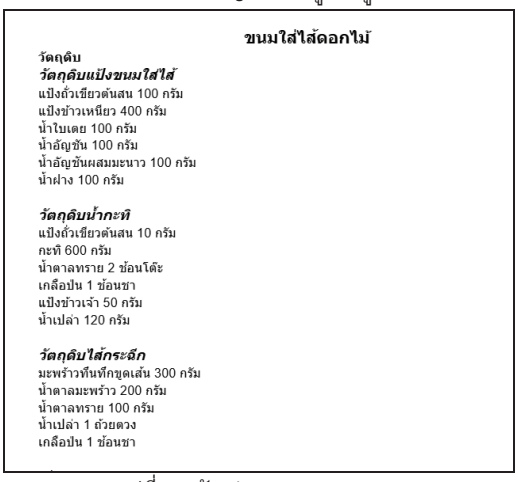
I.II กระบวนการสกัดข้อมูล

เมื่อกำหนด ontology ให้กับกราฟความรู้ได้แล้ว จากนั้นจะนำข้อมูลที่เตรียมไว้ไปสร้างกราฟความรู้และเก็บเอาไว้ใน neo4j ซึ่งเป็นฐานข้อมูลที่เก็บข้อมูลไว้ในรูปแบบกราฟ โดยมีโครงสร้างดังนี้



รูปที่ II. โครงสร้างการทำงานของระบบ [5]

เอกสารที่ถูกส่งเข้ามานั้นจะถูกกำหนดกรอบของหัวข้อด้วยแท็กตัวอักษรหนา “ ” และสิ่งที่อยู่ภายใต้แท็กจะถือว่าเป็นรายละเอียดของหัวข้อนั้นๆ จนกว่าจะเจอแท็กใหม่ในระดับเดียวกัน เมื่อทำการแบ่งส่วนข้อความออกเป็นกล่องข้อความได้แล้วให้พิจารณาแต่ละกล่องข้อความว่าอาจจะเป็นชื่อของเอนทิตี้ใดสักประเภทหนึ่งแต่นั่นก็ยังไม่ชัดเจนพอสำหรับให้โมเดลแยกแยะตาม ontology ที่กำหนดไว้ได้อย่างมีประสิทธิภาพจึงเพิ่มเงื่อนไขอีกการใส่แท็กตัวอักษรเอียง “ ” ไปที่หัวข้อที่บอกถึงวัตถุดิบและส่วนผสม แต่การสกัดเอนทิตี้ MixIngredient นั้นยังทำได้ไม่ดีนักเนื่องจากความไม่ชัดเจนเพราะ MixIngredient เกิดจากการนำวัตถุดิบต่างๆ มาประกอบเป็นบางอย่างขึ้นซึ่งมีลักษณะใกล้เคียงกับเอนทิตี้ recipe หรือก็คือ MixIngredient คือ sub Recipe ดังนั้นเพื่อให้สามารถสกัดเอนทิตี้นี้ได้ดีขึ้นจึงเพิ่มเงื่อนไขว่าถ้าหากเจอหัวข้อที่บ่งบอกถึงวัตถุดิบหลายหัวข้อในเอกสารเดียวให้พิจารณาว่ามีโอกาสที่จะเป็น MixIngredient สูง ดังรูปที่ III



ขนมใส่ไส้ดอกไม้

วัตถุดิบ

วัตถุดิบแป้งขนมใส่ไส้
แป้งถั่วเขียวต้นสน 100 กรัม
แป้งข้าวเหนียว 400 กรัม
น้ำใบเตย 100 กรัม
น้ำอัญชัน 100 กรัม
น้ำอัญชันผสมมะนาว 100 กรัม
น้ำฝาง 100 กรัม

วัตถุดิบน้ำกะทิ
แป้งถั่วเขียวต้นสน 10 กรัม
กะทิ 600 กรัม
น้ำตาลทราย 2 ช้อนโต๊ะ
เกลือป่น 1 ช้อนชา
แป้งข้าวเจ้า 50 กรัม
น้ำเปล่า 120 กรัม

วัตถุดิบไส้กระอิก
มะพร้าวทึนทึกขูดเส้น 300 กรัม
น้ำตาลมะพร้าว 200 กรัม
น้ำตาลทราย 100 กรัม
น้ำเปล่า 1 ถ้วยตวง
เกลือป่น 1 ช้อนชา

รูปที่ III. ตัวอย่างเอกสาร PDF

I.III การสร้างกราฟเชิงความรู้ (Knowledge Graph)

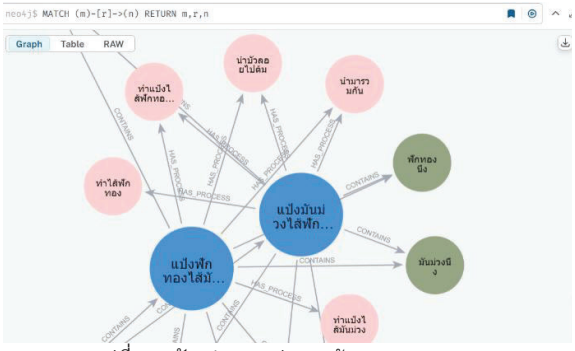
กราฟความรู้คือโครงสร้างข้อมูลที่จัดเก็บความรู้ในรูปแบบของกราฟที่ประกอบด้วยความสัมพันธ์ที่หลากหลาย โดยมีองค์ประกอบสำคัญได้แก่เอนทิตี (Entities) และความสัมพันธ์ (Relationships) ที่ถูกกำหนดนิยามหรือชนิดของเอนทิตี้กับความสัมพันธ์ด้วย ontology และเก็บข้อมูลคุณสมบัติต่างๆ (Property) เอาไว้ทั้งในแต่ละเอนทิตี้และในเส้นความสัมพันธ์เพื่อให้สามารถบ่งบอกความเกี่ยวข้องของข้อมูลเหล่านั้นได้

คุณภาพของกราฟนั้นไม่ได้ขึ้นอยู่กับปริมาณข้อมูลเพียงอย่างเดียวแต่ขึ้นอยู่กับความละเอียดในการนิยามประเภทของเอนทิตี้และความสัมพันธ์ ในปัจจุบันกราฟความรู้ถูกนำมาใช้เป็นฐานข้อมูลภายนอกให้กับโมเดลภาษาขนาดใหญ่ (LLM) เพื่อช่วยลดข้อผิดพลาดและเพิ่มความแม่นยำในการให้เหตุผล การสร้างกราฟความรู้นั้นสามารถทำได้หลักๆ 2 วิธีคือการใช้แรงงานคนกับใช้ระบบอัตโนมัติ *[6], [7], [8]* โดยที่แนวทางแรกจะให้ผลลัพธ์ที่แม่นยำกว่าแต่ก็ใช้ทรัพยากรมากกว่าเช่นกันสำหรับแนวทางที่ 2 จะเป็นการอ่านข้อมูลจากเอกสาร PDF โดยทั่วไปมักอาศัยเทคนิคการสร้าง bounding box จาก pixel position ในลักษณะเดียวกับ OCR *[9], [10]*ซึ่งมีข้อจำกัดสำคัญคือการพึ่งพาโครงสร้างเชิงภาพ (layout) ของเอกสาร หากรูปแบบเอกสารมีการเปลี่ยนแปลง ความแม่นยำของระบบจะลดลงอย่างมีนัยสำคัญ ด้วยเหตุนี้ งานวิจัยนี้จึงเสนอแนวทางที่แตกต่าง โดยมุ่งเน้นการใช้ ข้อมูลเชิงตัวอักษร (text-based structure) ที่มีอยู่ในเอกสารมาทำการจำแนกระดับหัวข้อ (hierarchical structure) พร้อมทั้งทำการ mark และ tagging เพื่อสร้างโครงสร้างเชิงความหมาย (semantic structure) ซึ่งช่วยลดการพึ่งพาตำแหน่งเชิงภาพ และเพิ่มความสามารถในการรองรับเอกสารที่มีรูปแบบหลากหลายได้ดียิ่งขึ้น แม้ว่างานวิจัยก่อนหน้า*[11]*จะมีการใช้โครงสร้างข้อความร่วมกับโมเดลแบบดั้งเดิม เช่น SVM หรือ TextCNN แต่โมเดลเหล่านั้นจำเป็นต้องผ่านการฝึกฝนและมีข้อจำกัดด้าน scalability ในเชิงวิธีการ

งานวิจัยนี้ได้เสนอวิธีการสกัดข้อมูลโดยการใช้ LLM ร่วมกับการใช้ข้อมูลรูปแบบของตัวอักษรที่มีอยู่แล้ว ซึ่งมีจุดเด่นอยู่ที่การผสานองค์ประกอบหลายส่วนเข้าด้วยกัน ได้แก่ การจัดรูปแบบเอกสาร การกำหนด ontology และการใช้ LLM เพื่อสกัดความรู้จากเอกสารกึ่งโครงสร้าง แนวทางดังกล่าวช่วยลดการพึ่งพาการทำ manual annotation จำนวนมาก และเอื้อต่อการประยุกต์ใช้ในลักษณะ training-free หรือ low-adaptation เมื่อเทียบกับแนวทางที่ต้องพึ่งการฝึกโมเดลเฉพาะงานหลังจากที่ได้ข้อมูลที่ถูกสกัดออกมาแล้วจะเข้าสู่กระบวนการการสร้างฐานข้อมูลบน neo4j ซึ่งเป็นเครื่องมือช่วยในการติดตามและตรวจสอบข้อมูลโดยจะใช้คำสั่ง cypher ในการสร้างและกำหนดสิ่งต่างๆ ภายใน neo4j ประกอบด้วยการกำหนดกฎที่บังคับการควบคุมคุณภาพข้อมูล (constraint) โดยจะกำหนดให้เอนทิตี้แต่ละประเภทมีชื่อไม่ซ้ำกัน (unique) ป้องกันข้อมูลขยะ เพื่อให้การสืบค้นข้อมูลทำได้ง่ายจึงสร้าง full text index ขึ้นมาเป็นตัวกรองในการค้นหาด้วยคุณสมบัติ (properties) ของแต่ละเอนทิตี้ เมื่อได้ข้อมูลที่สกัดด้วย LLM แล้วจะทำการอัพเดตข้อมูลที่แยกเอาไว้ตามชนิดของความสัมพันธ์เข้าไปใน neo4j ด้วย cypher query โดยการดึงข้อมูลจาก csv ทีละบรรทัดแล้ว map ลง query ที่เตรียมเอาไว้ให้ในแต่ละความสัมพันธ์

การพัฒนาระบบที่ต้องรองรับข้อมูลเชิงความหมาย (semantic) และความสัมพันธ์ที่ซับซ้อน เช่น “เมนู–วัตถุดิบ–ขั้นตอน–หมวดหมู่” มักเจอข้อจำกัดเมื่อจัดเก็บด้วยฐานข้อมูลทั่วไป เช่น ทำให้ขยายระบบยากเมื่อโดเมนเปลี่ยนหรือเพิ่มชนิดความสัมพันธ์ใหม่ๆ ดังนั้นงานวิจัยและงานพัฒนาระบบจำนวนมากจึงหันมาใช้ฐานข้อมูลกราฟ Neo4j ถูกออกแบบมาเพื่อจัดเก็บและสืบค้นข้อมูลในรูปแบบ “เอนทิตี้–ความสัมพันธ์” ได้โดยตรงภายใต้บริบทนี้ Neo4j ทำหน้าที่เป็นโครงสร้างพื้นฐานสำหรับจัดเก็บ KG ในรูปแบบ property graph ที่รองรับทั้งคุณสมบัติของเอนทิตี้/ความสัมพันธ์ และการสืบค้นเชิงความสัมพันธ์อย่างมีประสิทธิภาพส่งผลให้การพัฒนาระบบไม่เพียงเก็บข้อมูลแต่สามารถจัดเก็บความรู้ที่พร้อมนำไปใช้ต่อยอดในการค้นคืน

ข้อมูล การให้เหตุผลเชิงโครงสร้าง และการสร้างระบบตอบคำถามหรือระบบแนะนำบนโดเมนต่างๆ ได้อย่างเป็นระบบ



รูปที่ IV. ตัวอย่างกราฟความรู้จาก neo4j

IV. ผลการวิจัย

เพื่อประเมินประสิทธิภาพของระบบที่นำเสนอนี้ทางผู้วิจัยได้รวบรวมข้อมูลทำการวัดผลโดยเทียบกับเฉลย (Ground Truth) csv ที่เตรียมเอาไว้โดยนำแนวคิดของชุดข้อมูล FB15k-237 มาประยุกต์ในโดเมนอาหาร ทั้งนี้ FB15k-237 เป็นชุดข้อมูลสำหรับกราฟความรู้ที่จัดเก็บข้อมูลอยู่ในรูปแบบของ triples (head, relation, tail) ใช้สำหรับการประเมินคุณภาพของการทำนายความสัมพันธ์ดังนั้นในงานวิจัยนี้จึงได้นำแนวคิดนี้มาประยุกต์ใช้โดยการแยกชนิดความสัมพันธ์ออกเป็นชุดข้อมูลย่อยเพื่อให้สามารถประเมินได้ชัดเจนและเป็นระบบ ลดความสับสนของการปะปนหลายความสัมพันธ์จึงได้ทำการแยกไฟล์ csv ออกเป็น 5 ไฟล์ที่สำคัญตามความสัมพันธ์ในรูปที่ 1 เพื่อให้ได้ข้อมูลที่ครบถ้วนจึงได้เก็บคุณสมบัติ (properties) ของข้อมูลมาด้วยไม่ว่าจะเป็น ปริมาณ หน่วย ลำดับขั้นตอนและรายละเอียดของขั้นตอน

ตารางที่ I. ตัวอย่าง csv ที่ใช้เก็บข้อมูลวัตถุดิบ

recipe_name	ingredient_name	qty	unit
บัวลอยไข่ดาว	น้ำตาลทราย	70	กรัม
บัวลอยไข่ดาว	ใบเตย	2	ใบ
บัวลอยไข่ดาว	เกลือป่น	1	ช้อนชา

ตารางที่ II. ตัวอย่าง csv ที่ใช้เก็บข้อมูลขั้นตอนต่างๆ

miix_name	process_name	step_no	detail
ไส้กระฉีก	ผัดไส้กระฉีก	1	ทำไส้กระฉีกโดยนำ.....
แป้งขนมใส่ไส้	ทำแป้ง	2	ทำแป้งห่อไส้....

และได้ใช้เมตริกการวัดที่แสดงให้เห็นถึงความแม่นยำในการสกัดและการจดจำเอนทิตี้ของโมเดลได้แก่ F1 Precision Recall

เมื่อทำการนำผลลัพธ์ในรูปแบบเดียวกับ ตารางที่ 1 ไปเทียบกับเฉลย (Ground Truth) ที่มีรูปแบบเดียวกันจะได้ 3 ค่าจากการทำนายคือ F1-score Precision และ Recall เพื่อความถูกต้องและครบถ้วนจึงใช้ทุกคอลัมน์ในการวัดยกเว้นคอลัมน์รายละเอียดของขั้นตอนการทำเมนูอาหารเพราะว่ามีลักษณะเป็นข้อความยาวซึ่งอาจจะมีตกหล่นไปบ้างจากการสกัดข้อมูลของโมเดล เพื่อแสดงให้เห็นถึงความแตกต่างจึงได้ทำการทดสอบด้วยวิธีการใช้ prompt เพียงอย่างเดียวกับใช้ prompt ร่วมกับการจัดรูปแบบข้อมูลดังนี้

ตารางที่ III. ผลลัพธ์ก่อนการจัดรูปแบบเอกสาร

Model	F1	Precision	Recall
qwen2.5:3b	0.11	0.09	0.14
qwen2.5:7b	0.56	0.64	0.52
gemma3:4b	0.36	0.29	0.51
gemma3:12b	0.61	0.56	0.69

ตารางที่ IV. ผลลัพธ์หลังจากจัดรูปแบบเอกสาร

Model	F1	Precision	Recall
qwen2.5:3b	0.44	0.41	0.56
qwen2.5:7b	0.83	0.96	0.77
gemma3:4b	0.54	0.53	0.60
gemma3:12b	0.86	0.86	0.87

จากตารางที่ III ผลการทดลองแสดงให้เห็นว่า ขนาดของโมเดลมีผลต่อประสิทธิภาพอย่างมีนัยสำคัญ โดยโมเดลขนาดใหญ่ เช่น qwen2.5-7b และ gemma3-12b ให้ค่า F1-score สูงกว่าโมเดลขนาดเล็กอย่างชัดเจน สะท้อนถึงความสามารถในการเข้าใจบริบทและอนุมานเชิงความหมายที่ดีกว่า โดยเฉพาะในงานที่เกี่ยวข้องกับโครงสร้างข้อมูลเชิงซับซ้อน เช่น Knowledge Graph

อย่างไรก็ตาม การทำ formatting มีผลต่อทุกโมเดลอย่างสม่ำเสมอและช่วยยกระดับประสิทธิภาพได้อย่างมาก โดยเฉพาะในโมเดลขนาดเล็ก เช่น qwen2.5-3b ที่มีการเพิ่มขึ้นของ F1 อย่างเด่นชัด แสดงให้เห็นว่าการจัดโครงสร้างข้อมูลสามารถช่วยชดเชยข้อจำกัดของโมเดลได้ในระดับหนึ่ง นอกจากนี้ แม้แต่ในโมเดลขนาดใหญ่ เช่น qwen2.5-7b การทำ formatting ยังช่วยผลักดันประสิทธิภาพจนเข้าใกล้ระดับสูงสุด (F1 $\approx$ 0.86)

ในเชิงเปรียบเทียบระหว่างโมเดล พบว่า Qwen มีแนวโน้มตอบสนองต่อทั้งการเพิ่มขนาดและการทำ formatting ได้ดีกว่า gemma โดยมีอัตราการเพิ่มของ F1 สูงกว่าในทุกเงื่อนไข ซึ่งอาจสะท้อนถึงความสามารถของสถาปัตยกรรมโมเดลในการจัดการข้อมูลเชิงโครงสร้าง

V. สรุปและอภิปรายผล

บทความวิจัยนี้นำเสนอโมเดลสำหรับการเตรียมข้อมูลเพื่อสร้างกราฟเชิงความรู้แบบใหม่ด้วยการสกัดคำด้วยโมเดลภาษาขนาดใหญ่ (LLM) และ สัญญาณจากการจัดรูปแบบเอกสาร เพื่อนำไปสร้างกราฟเก็บลงใน neo4j โดยโมเดลนี้ประกอบไปด้วยขั้นตอนการทำงานดังนี้ การออกแบบ Ontology การสร้างป้ายกำกับให้กับเอนทิตี้ การกำหนดค่าคงที่และการสร้างสัมพันธ์ของกราฟ และในส่วนของการงานวิจัยได้ทางผู้วิจัยได้นำชุดข้อมูลที่รวบรวมมาทดสอบและเปรียบเทียบกับเฉลย (Ground Truth) เพื่อประเมินประสิทธิของโมเดลที่นำเสนอนี้ในด้าน

ความถูกต้อง ความสมบูรณ์ และความสอดคล้อง ซึ่งจากผลการทดลองพบว่าเมื่อนำชุดข้อมูลมาใช้โดยไม่ผ่านโมเดลนี้หรือโมเดลอื่นๆ ทำให้ประสิทธิของการจัดรูปแบบเอกสารด้อยกว่าการเตรียมชุดข้อมูลโดยใช้โมเดลที่นำเสนอนี้ ซึ่งสามารถทำให้การจัดรูปแบบเอกสารมีคุณภาพของชุดมูลสูงถึงเฉลี่ยร้อยละ 86

อย่างไรก็ตามงานนี้ยังมีข้อจำกัดในด้านความหลากหลายของข้อมูลในการทดลองซึ่งมุ่งเน้นไปที่สูตรอาหารเป็นหลักดังนั้นงานในอนาคตควรมุ่งเน้นไปที่การขยายขอบเขตเอกสารในหลายหัวข้อเพื่อประเมินและประยุกต์ใช้วิธีการนี้กับเอกสาร PDF ทั่วไปได้กว้างขวางมากยิ่งขึ้น

บรรณานุกรม

[1] L. Abdur Rahman, I. Papathanail, and S. Mougiakakou, *Introducing the Swiss Food Knowledge Graph: AI for Context-Aware Nutrition Recommendation*. 2025. doi: 10.48550/arXiv.2507.10156.

[2] H. Sha, F. Gong, B. Liu, R. Liu, H. Wang, and T. Wu, “Leveraging Retrieval-Augmented Large Language Models for Dietary Recommendations With Traditional Chinese Medicine’s Medicine Food Homology: Algorithm Development and Validation,” *JMIR Med. Inform.*, vol. 13, pp. e75279–e75279, Aug. 2025, doi: 10.2196/75279.

[3] Y. Zhang and G. Xiao, “A novel customizing knowledge graph evaluation method for incorporating user needs,” *Sci. Rep.*, vol. 14, p. 9594, Apr. 2024, doi: 10.1038/s41598-024-60004-x.

[4] X. Wang *et al.*, “Knowledge Graph Quality Control: A Survey,” *Fundam. Res.*, vol. 1, Sep. 2021, doi: 10.1016/j.fmre.2021.09.003.

[5] H. Kim, J. Choi, S. Park, and Y. Jung, “Layout Aware Semantic Element Extraction for Sustainable Science & Technology Decision Support,” *Sustainability*, vol. 14, no. 5, Feb. 2022, doi: 10.3390/su14052802.

[6] Y. Liu, J. Chen, X. Lin, J. Song, and S. Chen, “Development of a large language model–based knowledge graph for chemotherapy-induced nausea and vomiting in breast cancer and its implications for nursing,” *Int. J. Nurs. Sci.*, vol. 12, no. 6, pp. 524–531, Nov. 2025, doi: 10.1016/j.ijnss.2025.10.010.

[7] T.-W. Lin, G. Fierro, H. Li, T. Hong, P. Nuzzo, and A. Sangiovanni-Vinentelli, “Systematic Evaluation of Knowledge Graph Repair with Large Language Models,” arXiv.org. Accessed: Feb. 03, 2026. [Online]. Available: https://arxiv.org/abs/2507.22419v1

[8] M. Schaeffer, “Towards efficient Knowledge Graph-based Retrieval Augmented Generation for conversational agents,” phdthesis, Normandie Université, 2025. Accessed: Sep. 30, 2025. [Online]. Available: https://theses.hal.science/tel-05063876

[9] Ł. Garncarek *et al.*, “LAMBERT: Layout-Aware Language Modeling for Information Extraction,” 2021, pp. 532–547. doi: 10.1007/978-3-030-86549-8_34.

[10] M. Lamott, Y.-N. Weweler, A. Ulges, F. Shafait, D. Krechel, and D. Obradovic, “LAPDoc: Layout-Aware Prompting for Documents,” in *Document Analysis and Recognition - ICDAR 2024*, E. H. Barney Smith, M. Liwicki, and L. Peng, Eds., Cham: Springer Nature Switzerland, 2024, pp. 142–159. doi: 10.1007/978-3-031-70546-5_9.

[11] T. Yue, Y. Li, and Z. Hu, “DWSA: An Intelligent Document Structural Analysis Model for Information Extraction and Data Mining,” *Electronics*, vol. 10, no. 19, p. 2443, Jan. 2021, doi: 10.3390/electronics10192443.