

เครื่องจักรสร้างประโยคภาษาไทยอัตโนมัติด้วยตัวดำเนินการครอสโปรดัก ในฐานข้อมูลเชิงสัมพันธ์

Automatic Thai Sentences Generation Engine Using Cross-Product Operation in Relational Database

สุพัตรา แดงเจริญ¹ ขนิดา จรุงจิตต์¹ และชาวลิต ชันคำ²

Suphattrat Daengcharoen¹ Kanida Charungchit¹ and Chouvalit Khancome²

¹สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏราชชนครินทร์

¹Information Technology Department, Faculty of Science and Technology,

Rajabhat Rajanagarindra University

²ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง

²Computer Science, Department Faculty of Science, Ramkhamhaeng University

Received: June 19, 2023

Revised: August 22, 2023

Accepted: August 30, 2023

บทคัดย่อ

การประมวลผลภาษาธรรมชาติเป็นหลักการที่ทำให้คอมพิวเตอร์สามารถเข้าใจ ดีความ และใช้ภาษามนุษย์เพื่อการสื่อสารได้ โดยเฉพาะอย่างยิ่งการสร้างเท็กซ์เพื่อนำไปสู่สตอรี่เจนเนอเรชันอันเป็นหลักการทำให้คอมพิวเตอร์สร้างเรื่องราวอัตโนมัติ มีประโยชน์ในการสร้างเนื้อหาเท็กซ์ที่มีความซับซ้อนและเป็นเรื่องราวอย่างใดอย่างหนึ่งโดยเฉพาะ กลไกหลัก คือ การสร้างประโยคด้วยการนำคำชนิดต่าง ๆ วิธีหรือกลุ่มคำมาประกอบกันเป็นประโยคก่อนนำประโยคมาสร้างเนื้อความเชิงความหมายที่มนุษย์สามารถเข้าใจได้ งานวิจัยนี้พัฒนาและออกแบบเครื่องจักรซอฟต์แวร์สร้างประโยคภาษาไทยเพื่อเก็บไว้ในคลังประโยคภาษาไทยสำหรับนำไปใช้ในงานวิจัยสตอรี่เจนเนอเรชัน ออกแบบทั้งส่วนสถาปัตยกรรมเครื่องจักรและขั้นตอนวิธีแบ่งออกเป็นสองส่วนหลัก คือ ส่วนการสร้างพจนานุกรมชนิดของคำในภาษาไทยและส่วนการสร้างประโยคโดยใช้กลไกตัวดำเนินการครอสโปรดักของพีชคณิตเชิงสัมพันธ์ในฐานข้อมูลเป็นกฎควบคุมการสร้างประโยคตามรูปแบบไวยากรณ์ภาษาไทย ทดลองโดยสร้างเครื่องจักรซอฟต์แวร์ นำเข้าคำจากพจนานุกรมเล็กจิตรอนจำนวน 3×10^4 คำ สร้างประโยคภาษาไทยจำนวน 21 รูปแบบ ผลการทดลองพบว่า เครื่องจักรสามารถสร้างประโยคได้ปริมาณมากถึง 7.63926×10^{16} ประโยค วัดผลเชิงคุณภาพด้วยการพิจารณาว่า ประโยคที่สร้างขึ้นสามารถอ่านได้ความหมายถูกต้องหรือไม่ พบว่า ได้ประโยคที่สามารถอ่านได้ความหมายถูกต้องเฉลี่ยร้อยละ 36.70 น้อยที่สุดอยู่ที่ร้อยละ 13.33 มากที่สุดอยู่ที่ร้อยละ 64 พิจารณานำจำนวนคำที่นำมาสร้างประโยค ประโยคที่ประกอบด้วยคำ 2 คำ ประโยคที่สามารถอ่านได้ความหมายถูกต้องอยู่ระหว่างร้อยละ 44.00-64.00 คิดเป็นร้อยละเฉลี่ย 53.05 ประโยคที่ประกอบด้วยคำ 3 คำ ประโยคที่อ่านได้ความหมายถูกต้องอยู่ระหว่างร้อยละ 22.33-57.67 คิดเป็นร้อยละเฉลี่ยที่ 34.57 และประโยคที่ประกอบด้วย 4 คำ ได้ประโยคที่อ่านได้ความหมายถูกต้องอยู่ระหว่างร้อยละ 13.33-21.00 คิดเป็นร้อยละเฉลี่ยที่ 18.00

คำสำคัญ: การสร้างประโยคภาษาไทย คลังประโยคภาษาไทย เครื่องจักรสร้างคำไทย ครอสโปรดักรีเลชัน

Abstract

Natural Language Processing--NLP is the principle that enables computers to understand, interpret, and utilize human language for communication. Specifically, it involves generating text to create coherent narratives automatically. This is particularly useful for generating complex and narrative-rich textual content. The main mechanism involves constructing sentences by assembling various types of words, phrases, or groups of words into coherent sentences before creating meaningful content that humans can understand. This research develops and designs a software machine to generate Thai language sentences for storage in a Thai sentence repository for use in subsequent research on summarization. The design includes both the architecture of the machinery and the methods. It is divided into two main parts: generating a community dictionary of word types in the Thai language and generating sentences using the cross-productive operations of relational algebra on a database as control rules for generating sentences following Thai language syntax patterns. The experiment involves importing words from a small-sized city dictionary of 30,000 words and generating 21 Thai sentence patterns. The experimental results show that the machine can generate a large quantity of sentences, up to 7.63926×10^{16} sentences. The quality of the results is assessed by considering whether the generated sentences are readable and semantically correct. It is found that, on average, 36.70% of the sentences are readable and semantically correct, with a minimum of 13.33% and a maximum of 64%. Considering the number of words used to create sentences, sentences with two words have a readability and correctness rate between 44.00% and 64.00%, averaging 53.05%. For sentences with a length of 3 words, the readability and correctness rate range from 22.33% to 57.67%, with an average of 34.57%. Sentences with four words have a readability and correctness rate between 13.33% and 21.00%, averaging 18.00%.

Keywords: Thai sentence generation, Thai sentence corpus, Thai sentence engine, Cross-Product Relation



บทนำ

การประมวลผลภาษาธรรมชาติ (Natural Language Processing--NLP) เป็นหลักการที่ทำให้คอมพิวเตอร์สามารถเข้าใจ ตีความ และใช้ภาษาของมนุษย์เพื่อการสื่อสารโดยทั่วไปแล้ว งานวิจัยด้านนี้ ประกอบด้วยการจัดการโครงสร้างของคำ การวิเคราะห์ความสัมพันธ์ทางไวยากรณ์ ความหมาย และบริบท ซึ่งเกี่ยวข้องกับการตีความความหมายของภาษาและความหมายตามบริบท ตามลำดับ แต่ในปัจจุบันพบว่า หนึ่งในงานประมวลผลภาษาธรรมชาติที่มีความสำคัญ คือ งานด้านสตอรีเจเนอเรชัน (Story Generation Principle) ซึ่งเป็นหลักการที่ทำให้คอมพิวเตอร์สามารถ

สร้างเรื่องราวต่าง ๆ ตามความต้องการของมนุษย์ได้โดยอัตโนมัติ ถูกนำมาประยุกต์ใช้เพื่อสร้างเรื่องเล่า นิทาน นวนิยาย วรรณกรรม สารคดี บทความวิชาการ และงานทางด้านภาษาศาสตร์อื่น ๆ กลไกสำคัญที่ทำให้เกิดการสร้างเรื่องราวขึ้นได้ คือ การทำให้คอมพิวเตอร์สามารถนำคำชนิดต่าง ๆ (parts of speech) วลี (phrases) กลุ่มคำ (clauses) ที่มีใช้ในภาษาธรรมชาติมาสร้างประโยคก่อน จากนั้นจึงนำประโยคประกอบกันเพื่อให้ได้เนื้อความเชิงความหมายที่มนุษย์สามารถเข้าใจได้

แม้ว่าในปัจจุบันการสร้างเรื่องราวสามารถดำเนินการได้ด้วยปัญญาประดิษฐ์ (generative AI) เช่น ChatGPT

หรือ Bard AI ที่อาศัยการเรียนรู้ของเอไอที่จดจำเรื่องราวจากข้อมูลการเรียนรู้ภาษาธรรมชาติเอาไว้แล้วนำสิ่งที่จดจำไว้นั้นมาประกอบกันทั้งคิดแบบคำต่อคำและคิดแบบโครงสร้างเนื้อเรื่องที่เรียนรู้เอาไว้ แต่หากมีใช้สิ่งที่เคยเรียนรู้ไว้ก่อน การสร้างเรื่องราวจะไม่สามารถทำได้ อีกทั้งการสร้างเนื้อหา ยังไม่ได้ประสิทธิภาพเท่าที่ควร ดังนั้นการสร้างคลังคำ การสร้างคำเก็บในรูปแบบพจนานุกรม หรือหากจะให้ดีกว่านั้น คือ การสร้างคลังประโยคเอาไว้เพื่อการสร้างเรื่องราวจึงเป็นสิ่งที่สำคัญต่อการพัฒนาเรื่องราวได้รวดเร็วและมีประสิทธิภาพมากขึ้น อีกทั้งยังเป็นประโยชน์สำหรับให้ปัญญาประดิษฐ์นำไปใช้สำหรับตัดสินใจ หรือสร้างเรื่องราวใหม่ ๆ ได้ต่อไป คณะผู้วิจัยจึงต้องการสร้างเครื่องจักรเพื่อสร้างประโยคเป็นแนวคิดเบื้องต้นเพื่อจะใช้เป็นข้อมูลพื้นฐานสำหรับทำวิจัยต่อยอดในลำดับถัดไป

จากการศึกษาพบว่ามีงานวิจัยสร้างประโยคภาษาไทยค่อนข้างน้อย เช่น การออกแบบและทดลองเกี่ยวกับการสร้างคลังของเท็กซ์ภาษาไทยเพื่อใช้ในการสร้างเรื่องเล่า งานวิจัย (Limpanadudadee, Punyabukkanna & Poobrasert, 2014) และการสร้างเครื่องจักรสำหรับการสร้างประโยคภาษาไทยด้วยการกำหนดรูปแบบตายตัว (Krukaset, Krukaset & Khancome, 2017) งานวิจัยดังกล่าวถือเป็นกลไกการสร้างประโยคอันเป็นจุดเริ่มต้นของงานวิจัยสตอรีเจนเนอเรชันภาษาไทย ในขณะที่คณะผู้วิจัยได้พัฒนางานวิจัยเครื่องจักรสร้างประโยคภาษาไทยสำหรับคลังประโยคภาษาไทย (Daengcharoen, Charungchit & Khancome, 2020) ซึ่งเป็นเครื่องจักรสำหรับสร้างประโยคภาษาไทยเบื้องต้น สามารถสร้างประโยคภาษาไทยจำนวน 14 รูป ที่ความแม่นยำอยู่ที่ร้อยละ 5-60 และงานวิจัยเครื่องจักรสร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสาร (Khancome, Daengcharoen & Charungchit, 2022) ที่สร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสารได้ กระนั้นก็ตามรูปแบบและการดำเนินการยังมีข้อจำกัดอยู่ที่จำนวนรูปแบบ 14 รูปแบบที่มีข้อจำกัดการสร้างรูปแบบสื่อสารได้เพียง 8 รูปแบบที่ความแม่นยำต่ำ ไม่ยืดหยุ่นในการรองรับการขยายรูปแบบที่เพิ่มขึ้น การสร้างประโยคยังไม่มีความเป็นอิสระ

ดังนั้นงานวิจัยนี้จึงนำเสนอเครื่องจักรทางซอฟต์แวร์สำหรับสร้างประโยคภาษาไทยอัตโนมัติเพื่อเก็บไว้ในคลังประโยค (sentence corpus) อาศัยกลไกตัวดำเนินการ

ครอสโปรดัก (Cross-Product Relation) ของฐานข้อมูลเชิงสัมพันธ์ (relational database) พัฒนาส่วนควบคุมการสร้างประโยคด้วยฟังก์ชันทางพีชคณิตเชิงสัมพันธ์ (relational algebra) เป็นกฎในการสร้างประโยค ทำให้สามารถขยายขีดความสามารถของเครื่องจักรสร้างประโยคได้มากถึง 21 รูปแบบ ขั้นตอนวิธีมีความเป็นอิสระและยืดหยุ่นอันเป็นการแก้ไขข้อจำกัดของงานวิจัยที่ได้ดำเนินการมาก่อนหน้า การวิจัยในครั้งนี้พัฒนาเครื่องจักรเพื่อสร้างประโยคเป็นคลังประโยคภาษาไทยเก็บไว้ทั้งหมดก่อนจึงทำให้ได้ประโยคแบบไม่พึ่งบริบท จากนั้นพิจารณาเลือกเฉพาะประโยคที่มีคุณภาพที่สามารถอ่านได้ถูกต้องและมีความหมาย (ประโยคแบบพึ่งบริบท) ไปใช้งาน

โดยการทดลองนี้ ทดลองด้วยคำจากพจนานุกรม LEXITRON (LEXITRON Data, 2018) จำนวน 3×10^4 คำ ผลการทดลองได้ประโยคตามรูปแบบไวยากรณ์โดยไม่พึ่งบริบท (context free sentence) ได้ปริมาณมากถึง 7.63926×10^{16} ประโยค นอกจากนั้นหากพิจารณาด้านคุณภาพของประโยคที่สร้างขึ้นซึ่งสามารถอ่านได้ความหมายถูกต้องในเชิงความหมายแบบพึ่งบริบท (contextual sentence) สามารถอ่านได้ความหมายถูกต้องอยู่ระหว่างร้อยละ 13.33-64 อีกด้วย

วัตถุประสงค์การวิจัย

เพื่อพัฒนาเครื่องจักรทางซอฟต์แวร์สำหรับสร้างประโยคภาษาไทยอัตโนมัติเก็บไว้ในคลังประโยค โดยอาศัยกลไกตัวดำเนินการครอสโปรดักของฐานข้อมูลเชิงสัมพันธ์ควบคุมกฎในการสร้างประโยคตามรูปแบบประโยคภาษาไทย

แนวคิดทฤษฎีที่เกี่ยวข้อง

สตอรีเจนเนอเรชัน

สตอรีเจนเนอเรชัน คือ หลักการที่ทำให้คอมพิวเตอร์สามารถสร้างเรื่องราวต่าง ๆ ได้เองโดยอาศัยขั้นตอนวิธีทางด้านปัญญาประดิษฐ์ จากหลักฐานใน (Story Generator Algorithms, 2021) ระบุว่า เชลดอน ไคลน์ (Klein และ Sheldon) ได้นำเสนอแนวคิดนี้ไว้ตั้งแต่ปี ค.ศ.1973 จากนั้นงานวิจัยด้านนี้ได้มีการพัฒนาอย่าง

ต่อเนื่อง ซึ่งมีงานวิจัยที่น่าสนใจและมีชื่อเสียงเป็นที่กล่าวถึงกัน คือ มีฮานนำเสนอหลักการสร้างเรื่องราวที่เรียกว่าเทลสปิน (Tale-Spin) เป็นเครื่องจักรสำหรับสร้างเรื่องราวจนเป็นที่อ้างอิงและกล่าวถึงกันต่อ ๆ มา (Meehan, 1977) นอกจากนั้นมีฮานได้นำเสนอการพัฒนาโปรแกรมเพื่อสร้างเรื่องเล่าภาษาอังกฤษ ตามรูปแบบของคติสอนใจ (Moral) ที่กำหนด อีกทั้งยังสามารถเพิ่มตัวละครใหม่ได้ (Meehan, 1976) งานวิจัยออนทานอนและซู นำเสนอขั้นตอนวิธีที่เรียกว่า SAM--Story Analogies through Mapping ซึ่งช่วยให้คอมพิวเตอร์สามารถสร้างเรื่องเล่าที่มีความซับซ้อนมากขึ้น (Ontanon & Zhu, 2011) ลีและคณะ สร้างเรื่องราวโดยอาศัยพล็อตกราฟ (Meehan J. R., 1977) บายและคณะ สร้างเรื่องราวโดยอาศัยไวยากรณ์เพื่อสร้างพล็อตอัตโนมัติ (Bui, Abbbass & Bender, 2010) งานวิจัยสร้างประโยคโดยอาศัยออนโทโลยีเป็นกลไกสำคัญในการสร้างประโยคที่มีความหมาย เป็นต้น นอกจากนั้นยังมีงานวิจัยที่น่าสนใจในปัจจุบันอีกเช่น ในซู อิกิโมโต และ Riedl สามารถศึกษาเพิ่มเติมจากบทความสตอรีเจเนอเรชันของเบนและราฟเอล (Kybartas & Bidarra, 2010)

ประโยคภาษาไทย

ประโยค คือ หน่วยสมบูรณ์ทางไวยากรณ์ สามารถจำแนกได้ตามสถานการณ์การใช้งานและจำแนกตามโครงสร้างได้ 4 แบบ ได้แก่ ประโยคสามัญ ประโยคซับซ้อน ประโยคประสม และประโยคเชื่อม (Dekpituksirikul, 2008) แต่หากพิจารณาโครงสร้างประโยคตามแนวทางการใช้งานภาษาธรรมชาติที่กล่าวไว้ในโครงสร้างและวากยสัมพันธ์ของนามวลีภาษาไทยตามการจัดเรียงของคำในภาษาโดยเฉพาะภาษาไทยเป็นแบบ SVO โดยที่ S คือ ประธาน (subject) V คือ กริยา (verb) และ O คือ กรรม (object) ซึ่งในงานวิจัยเครื่องจักรสร้างประโยคภาษาไทยสำหรับคลังประโยคภาษาไทย (Daengcharoen, Charungchit & Khancome, 2020) ได้ศึกษารูปแบบภาษาไทยเทียบกับภาษาอังกฤษ (Thongkaol, 2007) พบว่า มีความคล่องจองกัน 4 รูปแบบ ได้แก่

- 1) S->NP VP
- 2) NP->N(Adj) (Class) (Det)

3) VP-> V (NP) (PP) และ

4) PP->Prep NP

โดยที่ S คือ ประโยค (Sentence) NP คือ นามวลี (Noun Phrase) VP คือ กริยาวลี (Verb Phrase) PP คือ บุพบทวลี (Preposition Phrase) N คือ คำนาม V คือ คำกริยา (Verb) ADJ คือ คำคุณศัพท์ (Adjective) Det คือ คำประกอบคำนามเพื่อกำหนดความหมายแบบเจาะจง (Determiner) Class คือ ลักษณะนาม (Classifier) และ Prep คือ คำบุพบท (Preposition)

งานวิจัยเครื่องจักรสร้างประโยคภาษาไทยสำหรับคลังประโยคภาษาไทย (Daengcharoen, Charungchit & Khancome, 2020) ซึ่งเป็นเครื่องจักรสำหรับสร้างประโยคภาษาไทยเบื้องต้น สามารถสร้างประโยคภาษาไทยจำนวน 14 รูป และงานวิจัยเครื่องจักรสร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสาร (Khancome, Daengcharoen & Charungchit, 2022) สามารถสร้างรูปแบบประโยคเพื่อทดลอง 14 รูปแบบ แบ่งเป็น 2 กลุ่ม แสดงดังตาราง 1

คณะผู้วิจัยได้ศึกษาเพิ่มเติมจากงานวิจัยเรื่อง ความซับซ้อนของนามวลีแปลงดัชนีบ่งชี้ความยากง่ายของตัวบท: กรณีศึกษาหนังสือเรียนวิชาภาษาไทย ระดับชั้นประถมศึกษาปีที่ 1-3 โดย ธารทอง แจ่มไพบูลย์ และวิโรจน์ อรุณมานะกุล (Jampaibool & Arunmanakul, 2016) ที่มีความละเอียดสำหรับสร้างรูปแบบภาษาไทยมากกว่างานวิจัยเดิมของคณะผู้วิจัย ผนวกกับการออกแบบกับงานเดิมจึงทำให้ได้รูปแบบของภาษาไทยที่นำมาใช้ในงานวิจัยนี้เพิ่มขึ้นถึง 21 รูปแบบ พร้อมส่วนการใช้งานโปรเจกชันครอสโปรดัก ดังตาราง 2 (นำเสนอในหัวข้อพีชคณิตเชิงสัมพันธ์)

พีชคณิตเชิงสัมพันธ์

พีชคณิตฐานข้อมูล (database algebra) หรือ พีชคณิตเชิงสัมพันธ์หรือฐานข้อมูลเชิงสัมพันธ์ถูกเขียนขึ้นเพื่อให้สามารถมองเห็นถึงความสัมพันธ์ที่เกิดขึ้นของข้อมูล โดยใช้พีชคณิตอธิบายความสัมพันธ์ของตารางต่าง ๆ ในฐานข้อมูลสามารถประยุกต์เข้ากับงานวิจัยทางด้านคอมพิวเตอร์ได้อย่างหลากหลาย ตัวดำเนินการพื้นฐาน (basic operations) ของพีชคณิตเชิงสัมพันธ์มี 6 ประเภท ดังนี้

1. ซีเล็คชัน (Select Operation-- σ (ซิกมา)) ทำหน้าที่เป็นตัวเลือกแถวข้อมูลจากเงื่อนไขที่กำหนดรูปแบบ คือ σ condition (Table)

2. โปรเจคชัน (Project Operation-- π (พาย)) ทำหน้าที่เป็นตัวเลือกคอลัมน์ข้อมูลจากเงื่อนไขที่กำหนดรูปแบบ π attribute1, attribute2,..., attribute n (Table)

3. ครอสโปรดัค (Cross-Product-- $\times$) หรือ คาร์ทีเซียนโปรดัค (Cartesian Product) ทำหน้าที่แสดงความสัมพันธ์ด้วยการคูณค่าที่เขียน (Cartesian) ระหว่าง 2 รีเลชัน รูปแบบ คือ $R \times S$

4. เซตดิฟเฟอเรน (Set-Difference-- $-$) ทำหน้าที่หาความแตกต่างระหว่าง 2 รีเลชันเพื่อแสดงข้อมูลของแถวที่ต่างจากรีเลชันแรก รูปแบบ คือ $R - S$

5. ยูเนียน (Union-- $\cup$) ทำหน้าที่นำข้อมูลจาก 2 รีเลชันมารวมกัน และหากมีข้อมูลซ้ำจะแสดงเพียงแถวเดียว รูปแบบ คือ $R \cup S$

6. อินเตอร์เซกชัน (Intersection-- $\cap$) ทำหน้าที่นำข้อมูลจาก 2 รีเลชันมารวมกัน และแสดงเฉพาะข้อมูล

ที่เหมือนกันจากทั้ง 2 รีเลชันเท่านั้น รูปแบบ คือ $R \cap S$

ตัวอย่างงานวิจัยที่ใช้ตัวดำเนินการทางด้านพีชคณิตฐานข้อมูล เช่น ดำเนินการแรงค์เอสคิวแอล (RankSQL) (Grandi, 2017) โดยกำหนดการใช้พีชคณิตเชิงสัมพันธ์เพื่อการแก้ปัญหา (Li, 2005) ส่วนวิจัยเกี่ยวกับการจัดการสายฐานข้อมูล (streaming database) โดยอาศัยความรู้เกี่ยวกับพีชคณิตเชิงสัมพันธ์ (Elmasri & Navathe, 2010) เป็นต้น

งานวิจัยนี้ใช้ตัวดำเนินการโปรเจคชันร่วมกับตัวดำเนินการค่าที่เขียนโปรดัคหรือครอสโปรดัคมากำหนดเป็นกฎในการสร้างประโยค เพื่อควบคุมการเลือกตารางและค่าเพื่อสร้างประโยคตามรูปแบบไวยากรณ์ คณะผู้วิจัยวิเคราะห์รูปแบบประโยคภาษาไทยผนวกกับรูปแบบงานวิจัยเดิม (Daengcharoen, Charungchit & Khancome, 2020) และงานวิจัยเครื่องจักรสร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสาร (Khancome, Daengcharoen & Charungchit, 2022) ได้นี้เพิ่มขึ้นถึง 21 รูปแบบ

ตาราง 1

รูปแบบประโยคจากงานวิจัยที่ใช้เป็นต้นแบบงานวิจัยนี้

กลุ่มที่ 1	กลุ่มที่ 2
(1) N + ADJ	(1) NP1 + ADJ
(2) N + BE + ADV	(2) NP1 + BE + ADV
(3) N + BE + NP1	(3) NP1 + BE + NP2
(4) N + V	(4) NP1 + V
(5) N + V + PP	(5) NP1 + V + PP
(6) N + V + NP1	(6) NP1 + V + NP2
(7) N + V + NP1 + NP2	(7) NP1 + V + NP2 + NP3

ตาราง 2

รูปแบบประโยคที่ใช้ในงานวิจัย

รูปแบบ	โปรเจกชัน-ครอสโปรดัก	รูปแบบประโยค
T_1	$\pi(N^{(noun)}ADJ^{(adj)})$	N x ADJ
T_2	$\pi(N^{(noun)}BE^{(be)}ADV^{(adv)})$	N x BE x ADV
T_3	$\pi(N^{(noun)}BE^{(be)}NP1^{(np1)})$	N x BE x NP1
T_4	$\pi(N^{(noun)}V^{(verb)})$	N x V
T_5	$\pi(N^{(noun)}V^{(verb)}PP^{(pp)})$	N x V x PP
T_6	$\pi(N^{(noun)}V^{(verb)}NP1^{(np1)})$	N x V x NP1
T_7	$\pi(N^{(noun)}V^{(verb)}NP1^{(np1)}NP2^{(np2)})$	N x V x NP1 x NP2
T_8	$\pi(NP1^{(np1)}ADJ^{(adj)})$	NP1 x ADJ
T_9	$\pi(NP1^{(np1)}BE^{(be)}ADV^{(adv)})$	NP1 x BE x ADV
T_{10}	$\pi(NP1^{(np1)}BE^{(be)}NP2^{(np2)})$	NP1 x BE x NP2
T_{11}	$\pi(NP1^{(np1)}V^{(verb)})$	NP1 x V
T_{12}	$\pi(NP1^{(np1)}V^{(verb)}PP^{(pp)})$	NP1 x V x PP
T_{13}	$\pi(NP1^{(np1)}V^{(verb)}NP2^{(np2)})$	NP1 x V x NP2
T_{14}	$\pi(NP1^{(np1)}V^{(verb)}NP2^{(np2)}NP3^{(np3)})$	NP1 x V x NP2 x NP3
T_{15}	$\pi(V^{(verb)}NP1^{(np1)})$	V x NP1
T_{16}	$\pi(V^{(verb)}NP1^{(np1)}N^{(noun)})$	V x NP1 x N
T_{17}	$\pi(NP1^{(np1)}NP2^{(np2)}V^{(verb)})$	NP1 x NP2 x V
T_{18}	$\pi(V^{(verb)}NP1^{(np1)}NP2^{(np2)})$	VxNP1 x NP2
T_{19}	$\pi(V^{(verb)}NP1^{(np1)}NP2^{(np2)}NP3^{(np3)})$	VxNP1 x NP2 x NP3
T_{20}	$\pi(NP1^{(np1)}NP2^{(np2)}V^{(verb)}NP3^{(np3)})$	NP1 x NP2 x V x NP3
T_{21}	$\pi(NP1^{(np1)}NP2^{(np2)})$	NP1 x NP2

เครื่องจักรสร้างประโยคภาษาไทยที่เกี่ยวข้อง

งานวิจัยเครื่องจักรสร้างประโยคภาษาไทยสำหรับคลังประโยคภาษาไทย Daengcharoen, Charungchit and Khancome (2020) ซึ่งถือเป็นเครื่องจักรซอฟต์แวร์ต้นแบบสำหรับสร้างประโยคภาษาไทยเก็บไว้ในคลังประโยคภาษาไทยอาศัยกลไกโดยอาศัยฐานข้อมูลเชิงสัมพันธ์ ใช้ฟังก์ชันทางพีชคณิตเชิงสัมพันธ์เป็นกฎควบคุมการสร้างประโยคตามรูปแบบไวยากรณ์ภาษาไทย แสดงสถาปัตยกรรมและการออกแบบขั้นตอนวิธี แสดงดังภาพ 1

เครื่องจักรและขั้นตอนวิธีดังกล่าวนี้สามารถสร้างประโยคได้เพียง 14 รูปแบบเท่านั้น และยังเป็นประโยคความเดียวที่เน้นเป็นประโยคบอกเล่า ยิ่งเมื่อพิจารณาประโยคแบบพหุบริบท พบว่า ประโยคแบบพหุบริบทที่สามารถเข้าใจตามความหมายอยู่ที่ประมาณร้อยละ 5-60 ของประโยคที่ได้จากการสุ่มเท่านั้น ซึ่งเมื่อพิจารณาละเอียดลงโดยจำแนกประโยคที่เกิดจากรูปแบบของคำ 2 3 และ 4 คำ สามารถเข้าใจความหมายได้มากที่สุดร้อยละ 60 45 และ 5 เครื่องจักรยังไม่ตอบสนองต่อประโยคที่แบ่งเจตนาการสื่อสารหรือรูปแบบประโยคมากกว่า 14 รูปแบบได้ อีก

ทั้งเมื่อนำไปพัฒนาเพื่อสร้างประโยคยังใช้เวลาสร้างมาก ไม่เป็นอิสระในการสร้างประโยคเดียว ไม่รองรับการขยาย รูปแบบประโยคอีกด้วย

ดังนั้นในงานวิจัยเครื่องจักรสร้างประโยคภาษาไทย ตามวัตถุประสงค์ของการสื่อสาร Khancome, Daengcharoen and Charungchit (2022) จึงพยายามสร้างประโยคภาษาไทยให้ได้ตามวัตถุประสงค์ของการสื่อสาร โดยปรับปรุง และแบ่งรูปแบบ 14 รูปแบบของประโยคออกเป็นตาม หน้าที่ของประโยคได้แก่ สร้างประโยคความรวมภาษาไทย ที่มีเนื้อความคล้ายตามกัน เนื้อความขัดแย้งกัน เนื้อความ เลือกเอาอย่างใดอย่างหนึ่ง และเนื้อความเป็นเหตุเป็นผล กัน ขั้นตอนวิธีที่พัฒนาขึ้นเพื่อรองรับการสร้างประโยคตาม วัตถุประสงค์เหล่านั้น ยังมีข้อจำกัดการสร้างประโยคที่มากกว่า 14 รูปแบบและการสร้างแต่ละรูปแบบไม่เป็นอิสระไม่ เหมาะสำหรับการขยายรูปแบบประโยค

วิธีดำเนินการวิจัย

กระบวนการดำเนินการวิจัย ดำเนินการตามลำดับ ดังนี้

วิเคราะห์รูปแบบประโยคจากงานวิจัยเดิมของ คณะวิจัยที่ได้นำเสนอข้างต้น ร่วมกับการพัฒนารูปแบบ ประโยคใหม่ให้ได้จำนวน 21 รูปแบบ

นำคำศัพท์จากพจนานุกรม LEXITRON มา วิเคราะห์คำชนิดต่าง ๆ ว่าเป็นคำประเภทใดบ้าง สามารถ ที่จะนำไปสร้างเป็นรหัสชั้นในระบบจัดการฐานข้อมูลเชิง สัมพันธ์ได้อย่างไรบ้าง

ออกแบบพจนานุกรมใหม่เรียกว่า Thai POS Dictionary เพื่อให้เหมาะที่จะจัดเก็บรหัสชั้น (ตาราง ประเภทของคำ) ได้แก่ คำนาม คำสรรพนาม คำกริยา คำ วิเศษณ์ คำบุพบท คำสันธาน คำอุทาน พร้อมทั้งออกแบบ ขั้นตอนวิธีสำหรับแปลงข้อมูลจากพจนานุกรม LEXITRON ให้เป็น Thai POS Dictionary และเขียนโปรแกรมตามขั้น ตอนวิธีแปลงข้อมูลเก็บไว้ในพจนานุกรมเพื่อใช้ทดลองต่อไป

ออกแบบเครื่องจักรและขั้นตอนวิธีเพื่อสร้างประโยค ตามวัตถุประสงค์การวิจัย และเขียนโปรแกรมเพื่อทดลอง ประเมินผล และเขียนรายงานวิจัย

ต่อไปนี้จะนำเสนอเครื่องจักรสร้างพจนานุกรม Thai POS ขั้นตอนวิธีสำหรับสร้างเครื่องจักร และส่วนของ เครื่องจักรสร้างประโยคพร้อมขั้นตอนวิธีประกอบ รวมถึง

นำเสนอานิยามต่าง ๆ ที่ใช้ประกอบการอธิบายขั้นตอนวิธีที่ สร้างขึ้นตามลำดับ

เครื่องจักรสร้างพจนานุกรม

แนวคิดการออกแบบเครื่องจักรสำหรับสร้าง พจนานุกรม Thai POS แสดงดังภาพที่ 2 ดังนี้

ขั้นตอนวิธีแบ่งเป็นส่วนนำเข้า (input) ส่วนประมวล ผล (processing) และส่วนนำผลออก (result output) ก่อนการนำเสนอขั้นตอนวิธีจำเป็นต้องนำเสนอานิยามที่ เกี่ยวข้อง ดังนี้

นิยามที่ 1 กำหนดให้ L คือ พจนานุกรม LEXITRON ที่ประกอบด้วยคำไทย $w_1, w_2, w_3, \dots, w_n$ เขียนแทนด้วย $L = \{ w_1, w_2, w_3, \dots, w_n \}$

นิยามที่ 2 กำหนดให้ WT คือ ชนิดของคำตาม ชนิดของคำ (Part of Speech) บรรจุชนิดของคำ $wt_1, wt_2, wt_3, \dots, wt_k$ เขียนแทนด้วย $WT = \{wt_1, wt_2, wt_3, \dots, wt_k\}$

นิยามที่ 3 w_b ใด ๆ ใน L ของนิยามที่ 1 สามารถ ระบุได้ว่า ชนิดของคำได้ว่าเป็น t_x ใดใน WT เขียนแทนได้ ด้วย w_b : wtx

ตัวอย่างที่ 1 หากกำหนดให้ $L = \{นก, กิน, ปลา, \dots\}$ กำหนด $WT = \{\text{คำนาม (N), กริยา (V), สรรพนาม (PRN), \dots}\}$

จากนิยามที่ 1 และ 2 สามารถอธิบายและนำเสนอ ให้เห็นภาพได้ดังตัวอย่างนี้ $w_1 = \text{นก}$, $w_2 = \text{กิน}$, $w_3 = \text{ปลา}$ เขียนแทนสัญลักษณ์ตามนิยามที่ 3 ได้ นก:N, กิน:V, ปลา:N เป็นต้น

นิยามที่ 4 กำหนดให้ TPOS คือ พจนานุกรม Thai POS Dictionary ที่จัดเก็บตาราง (ความสัมพันธ์) $R_1, R_2, R_3, \dots, R_j$ ที่แต่ละ R_i เก็บคำชนิด wt_i ดังนี้ยามที่ 2

จากนิยามที่ 4 สามารถกล่าวได้ว่า ตาราง คำนาม (Noun) สรรพนาม (Pronoun) คำกริยา (Verb) นามวลี (Noun Phrase) คำบุพบท (Preposition) หรืออื่น ๆ ที่เป็นไปตามนิยามที่ 1 สามารถแทนด้วยความสัมพันธ์ตาม นิยามที่ 4 ได้ แสดงขั้นตอนวิธีสำหรับสร้าง TPOS ดังนี้

Algorithm 1: Generate Thai POS Engine

Input: $L = \{w_1, w_2, w_3, \dots, w_n\}$,

$WT = \{wt_1, wt_2, wt_3, \dots, wt_k\}$, Empty TPOS

Output: $TPOS = (R_1, R_2, R_3, \dots, R_j)$

Create Empty $R_1, R_2, R_3, \dots, R_j$

For $i = 1$ To n Do

Get w_i form L

$T_x = wt_k$ of w_i

Do case t_x :

1: Store w_i to R_1 in TPOS

2: Store w_i to R_2 in TPOS

3: Store w_i to R_3 in TPOS, ...,

j : Store w_i to R_j

End of Do case

End of For

Return TPOS

นิยามที่ 5 กำหนดให้ TP แทน เซตของรูปแบบประโยคภาษาไทย (Thai Sentence Patterns)

นิยามที่ 6 กำหนดให้ $T_1, T_2, T_3, \dots, T_r \in TP$ โดยที่แต่ละ T ประกอบด้วยลำดับของความสัมพันธ์ในนิยามที่ 2 ตั้งแต่หนึ่งความสัมพันธ์ขึ้นไป โดยที่ $1 \leq q \leq r$

ตัวอย่างที่ 2 ตามนิยามที่ 6 อธิบายได้ว่า รูปแบบประโยค

T1-T21 ตามภาพ 1 เช่น $T1 = N \times ADJ$ หรือ $T2 = NP \times BE \times NP$ ต่างถือว่าเป็นรูปแบบหนึ่งของ T_q ใด ๆ

นิยามที่ 7 กำหนด TSC คือ คลังเก็บประโยคภาษาไทยที่ประกอบด้วยรีเลชันผลลัพธ์ของการสร้างประโยคภาษาไทยจากรูปแบบภาษาไทยตามนิยามที่ 6 เขียนแทนด้วย $TSC = \{tsp1, tsp2, tsp3, \dots, tspq\}$ โดยที่ $tspq$ ใด ๆ สอดคล้องกับตรงกันกับรูปแบบประโยค T_q

ตัวอย่างที่ 3 จากนิยามที่ 6 และ 7 อธิบายได้ว่า $tspq$ คือ ผลของการนำรูปแบบ T_q ใด ๆ มาสร้างประโยคภาษาไทย เช่น ถ้า T_q คือ $NP \ BE \ ADV$ ผลลัพธ์ที่เกิดจากการสร้างประโยคจากรูปแบบนี้ด้วยการนำตาราง $NP \ BE$ และ ADV มาเพื่อสร้างประโยคตามแนวทางของเครื่องจักรที่พัฒนาขึ้นในงานวิจัยนี้ (ดังแสดงในลำดับถัดไป) คือ รีเลชัน $tspq$ ที่บรรจุใน TSC

นิยามที่ 8 กำหนดให้ $PT[1...f]$ คือ พื้นที่ชั่วคราวที่ใช้เก็บชื่อรีเลชันที่ประกอบกันอยู่ในรูปแบบประโยค T_q ที่ประกอบด้วยลำดับของ wt ใด ๆ ที่ประกอบกันเป็นรูปแบบประโยค

ตัวอย่างที่ 4 ตามนิยามที่ 8 ถ้า $T_q = NP \times BE \times ADJ$ พื้นที่ชั่วคราวสำหรับเก็บรูปแบบดังกล่าวคือ $PT [1]$ เก็บ $NP \ PT [2]$ เก็บ BE และ $PT [3]$ เก็บ ADJ เป็นต้น

Input: $(R_1, R_2, R_3, \dots, R_l), (T_1, T_2, T_3, \dots, T_h)$ where $1 \leq l \leq n$ and $1 \leq h \leq m$

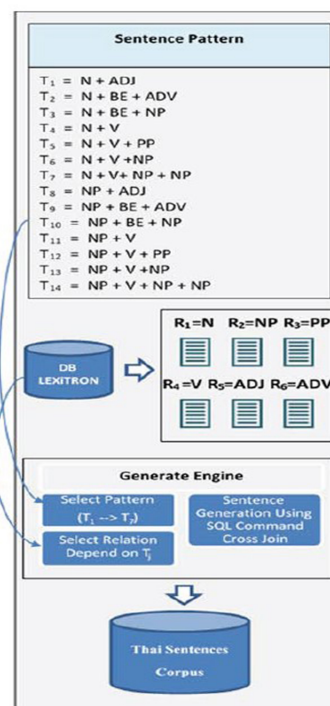
Processing:

(1) For $j=1$ To h Do

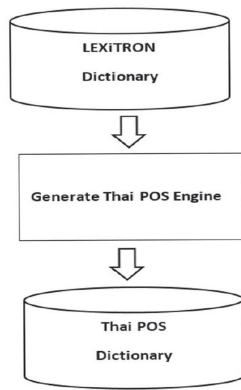
(2) Analyze Thai sentence pattern of T_j and Select relation(s) from R_1 to R_l followed by T_j

(3) $rs_j = \pi(R_1^{(field1)} \text{ to } R_l^{(fieldl)}) R_1 \times \dots \text{ to } R_l$

Result Output: $rs_1, rs_2, rs_3, \dots, rs_h$



ภาพ 1 สถาปัตยกรรมและขั้นตอนวิธีของเครื่องต้นแบบ



ภาพ 2 เครื่องจักรสร้าง Thai POS Dictionary

เครื่องจักรหลักและขั้นตอนวิธีสร้างประโยค

เบื้องต้นจำเป็นต้องกำหนดนิยามเพิ่มเติมจากส่วนของการสร้างพจนานุกรม TPOS ดังนี้

แนวคิดขั้นตอนวิธีหลักของการออกแบบเครื่องจักรแบ่งเป็นส่วนนำเข้า (input) เป็น TPOS และ รูปแบบประโยค T_q ส่วนประมวลผล (processing) คือ เครื่องจักรซอฟต์แวร์หลัก ATSGE--Automatic Thai Sentence Generation Engine และส่วนนำผลออก (result output) คือ Thai Sentence Corpus ซึ่ง คือ tsp_q ในนิยามที่ 7 แสดงดังเครื่องจักรภาพ 3 ดังนี้

แสดงขั้นตอนวิธีที่สร้างประโยคด้วยเครื่องจักร ATSGE ดังนี้

Algorithm 2: ATSGE

Input : TPOS= $(R_1, R_2, R_3, \dots, R_j)$, $(TP = T_1, T_2, T_3, \dots, T_r)$, T_q ,

PT[1...f], TSC={ $tsp_1, tsp_2, tsp_3, \dots, tsp_i$ }

Output: tsp_q Add to TSC

1. $T_c \leftarrow T_q \cap TP$
2. If $T_c \neq \emptyset$ Then
3. Read T_c form TP add PT [1...f]
4. $\pi(R_{PT[1]}^{(field1)} \text{ to } R_{PT[f]}^{(field1)})$
5. $Tsp_c \leftarrow \text{Cross Join } R_{PT[1]} \times R_{PT[2]} \times R_{PT[3]} \times \dots \times R_{PT[f]}$
6. Else
7. Return
8. End of If
9. $TSC \leftarrow tsp_c$
10. Return

ตัวอย่างที่ 5 จากแนวคิดของเครื่องจักรดังกล่าว แสดงการสร้างประโยคภาษาไทยได้ดังตัวอย่างต่อไปนี้ เมื่อ

ส่วนนำเข้า: $R_1=N$, $R_2=BE$, $R_3=ADV$ และรูปแบบ $T_q=N \ BE \ ADV$ โดยสมมุติจำนวนเรคคอร์ดของ $N=100$ $BE=10$ $ADV=100$

ส่วนการประมวลผล: เลือกรูปแบบ $N \ BE \ ADV$ และจึงไปเลือกตาราง $N \ BE \ ADV$ จากนั้นทำการ Cross join ด้วย $tsp_c = \pi((N^{(noun)}BE^{(be)})ADV^{(adb)})--(N \times B \times ADV)$ $(100 \times 10 \times 100)$

ส่วนนำผลออก: tsp_c ด้วยปริมาณเรคคอร์ดประโยคภาษาไทยจำนวน 100,000 เรคคอร์ด (ประโยค)

พิจารณาเกี่ยวกับความยืดหยุ่นของขั้นตอนวิธีที่นำเสนอ (algorithm 2) และขั้นตอนวิธีในภาพ 1 พบว่า ขั้นตอนวิธีใหม่ที่พัฒนาขึ้นมีความยืดหยุ่นมากกว่าโดยสามารถคำนวณและสร้างประโยคได้ทีละรูปแบบของประโยค T_c โดยไม่ผูกติดด้วยอุปกรณ์การวนซ้ำทุกรูปแบบจำนวน 14 รูปแบบ ดังที่นำเสนอไว้ในงานวิจัยก่อนหน้านี้ในขณะเดียวกัน ขั้นตอนวิธีใหม่นี้ยังสามารถดำเนินการกับประโยครูปแบบใหม่ ๆ และเพิ่มรูปแบบประโยคเป็นจำนวน 21 รูปแบบ

เครื่องมือที่ใช้ในการวิจัย

1. ข้อมูลคำจากพจนานุกรมอิเล็กทรอนิกส์ไทย-อังกฤษ (LEXITRON) (LEXITRON Data, 2018)
2. โครงสร้างจักรประโยคภาษาไทย
3. โปรแกรมบริหารจัดการฐานข้อมูล MySQL Version 5.2
4. โปรแกรมภาษา PHP
5. คอมพิวเตอร์โน้ตบุ๊ก ASUS ซีพียู Core i3 1.8 GHz แรม 4 กิกะไบต์ ฮาร์ดดิสก์ 1 เทราไบต์

สถิติที่ใช้ในการวิเคราะห์ข้อมูล

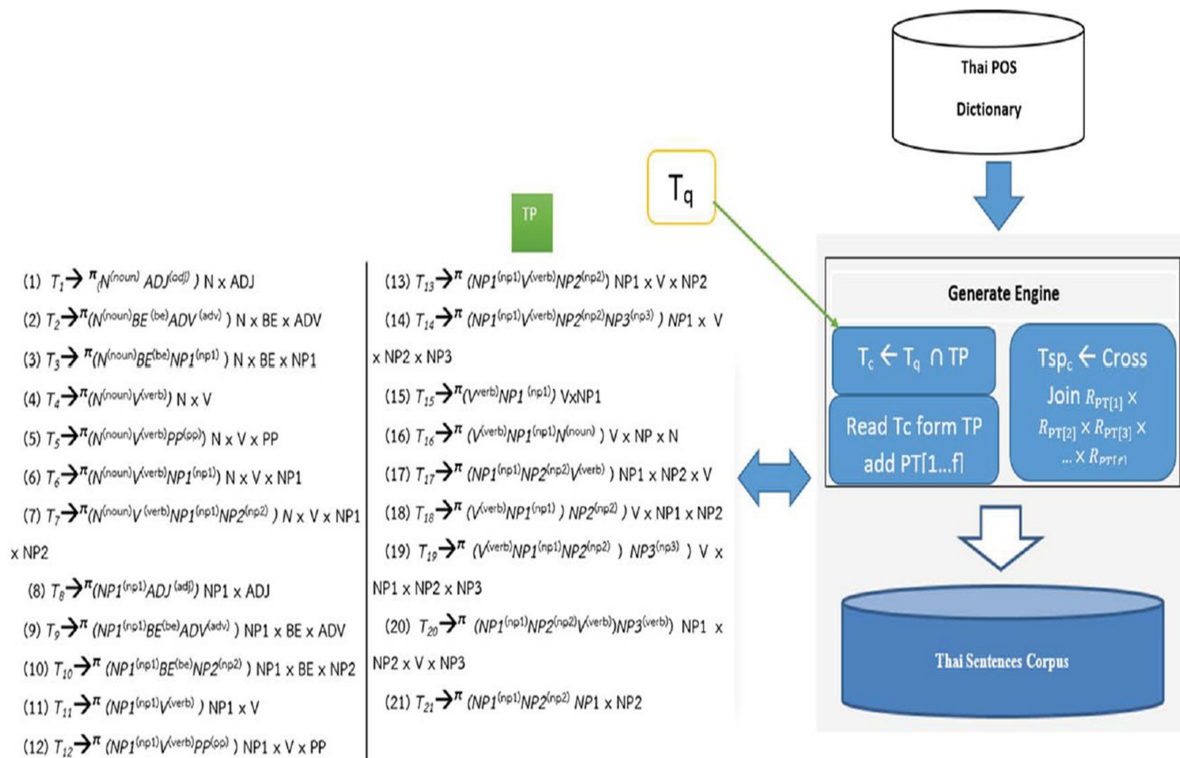
สถิติที่ใช้ในการวิจัย ได้แก่ ร้อยละ (%) ค่าเฉลี่ย ($\bar{X}$) ส่วนเบี่ยงเบนมาตรฐาน (SD)

การทดลอง

1. นำเข้าข้อมูลจาก LEXiTRON จำนวน 40,854 เรคคอร์ด แสดงตัวอย่างดังภาพ 4
2. พัฒนาโปรแกรมด้วย PHP สร้างรีเลชันตามรูปแบบ แต่ละ R_q เพื่อสร้างเป็น TPOS ตามขั้นตอนวิธี 1
3. พัฒนาโปรแกรมด้วย PHP ตามขั้นตอนวิธี 2 โดยนำเข้ารูปแบบประโยค TP และ TPOS สร้างประโยคภาษาไทย

4. เขียนโปรแกรมด้วย PHP นับจำนวนรวบรวมจำนวนเรคคอร์ดที่ได้จากการสร้างประโยคทั้งหมด เพื่อวัดผลเชิงปริมาณ

5. เขียนโปรแกรมสุ่มประโยคอย่างอิสระจากผลการสร้างประโยคตามรูปแบบประโยค 21 รูปแบบ จำนวน 12 ชุด ชุดละ 25 ประโยค (300 ประโยคต่อชุด) รวมทั้งสิ้น 6,300 ประโยค เพื่อวัดผลเชิงคุณภาพด้วยการพิจารณาการอ่านว่าได้ความหมายถูกต้องหรือไม่



ภาพ 3 เครื่องจักรซอฟต์แวร์ ATSGE

word_id	words_word	words_eng	words_cat	words_syn	words_sample
1	ดังกล่าวข้างต้น	abovementioned	PRON	ดังกล่าว	หน่วยงานของเราสามารถรับบทบาทได้เป็นอย่างดี.
2	การงดออกเสียง	abstention	N	การไม่แสดงความเห็นตามสิทธิ	
3	ตามกฎหมาย	according to law	ADV	ถูกต้องตามกฎหมาย	
4	เทคโนโลยีสมัยใหม่	advance technologies	N	เทคโนโลยีสมัยใหม่	คอมพิวเตอร์เป็นปัจจัยสำคัญในการพัฒนาเทคโนโลยี...
5	ระเบียบวาระการประชุม	agenda	N	ลำดับรายการการประชุมที่กำหนดไว้	
6	ฮอร์โมนเพศชาย	testosterone	N	สารที่สร้างขึ้นโดยต่อมไร้ท่อของร่างกายในเพศชาย	
7	เอเชีย	Asia	N	ชื่อทวีป	
8	เอเชีย	Asia	ADJ	เกี่ยวกับทวีปเอเชีย	
9	อาเซียน	Asian	N	ภูมิภาคเอเชียตะวันออกเฉียงใต้	
10	อาเซียน	Asian	ADJ	เกี่ยวกับภูมิภาคในแถบตะวันออกเฉียงใต้	
11	เป็นหลักประกัน	Assure	V	เป็นหลักประกัน	การใช้จ่ายโดยประหยัด จะเป็นหลักประกันความ...
12	แต่บางที	at other times	ADV	บางที, บางคราว, บางครั้ง	บางทีชาวไร่ควรจะแบ่งชนิดของฝ้ายตั้งแต่ตอนเก็บ
13	ในบางกรณี	at some point	DET	บางกรณี	ในบางกรณี ภาษานำเข้าของวัตถุดิบยังประกอบไม่...
14	อำนาจหน้าที่	authority and function	N	อำนาจหน้าที่	

ภาพ 4 ข้อมูล LEXITRON ที่ใช้ทดลอง

ผลการทดลอง

จากการดำเนินการพัฒนาโปรแกรมนำเข้า LEXITRON เข้าสู่ระบบจัดการฐานข้อมูลมายเอสคิวแอล และเขียนโปรแกรมตามขั้นตอนวิธี 1 ทำให้ได้ข้อมูลแต่ R_q ซึ่งเป็นตารางและจำนวนคำเพื่อใช้ทดลอง แสดงตัวอย่างผลการทดลองขั้นตอนนี้ ดังภาพ 5

เมื่อได้แต่ละ R_q ดังภาพ 7 นำเข้ารีเลชันดังกล่าวให้โปรแกรมที่พัฒนาตามขั้นตอนวิธี 2 โดยเลือกรีเลชัน R_q ที่เกี่ยวข้องกับ TP ดังนี้ $N=21,090$ คำ $V=13,028$ คำ $BE=4$ คำ $ADJ=2,879$ คำ $ADV=2,876$ คำ $NP=10,545$ คำ และ $PP=221$ คำ

การนำเสนอผลการทดลองแบ่งเป็น 2 ส่วน คือ ผลการทดลองเชิงปริมาณและผลการทดลองเชิงคุณภาพ โดยตาราง 3 และ 4 แสดงผลการทดลองเชิงปริมาณ ส่วนตาราง 5 และ 6 แสดงผลการทดลองเชิงคุณภาพพิจารณาจากประโยคที่อ่านได้ความหมายถูกต้อง

ผลการทดลองเชิงปริมาณ

ตาราง 3 แสดงรูปแบบคำประกอบประโยค ตัวอย่างประโยคที่ได้จากการทำงานของเครื่องจักร ทั้งประโยคที่อ่านได้ความหมายถูกต้องและไม่สามารถเข้าใจได้โดยแยกให้เห็นคำแต่ละชนิดที่นำมาประกอบกันเป็นประโยค ตาราง 4 แสดงผลเชิงปริมาณที่ได้จากการครอสโปรดักทั้งหมดจำแนกตามรูปแบบและจำนวนของคำและวลีที่นำมาประกอบกันตามรูปแบบ 21 รูปแบบ

จากตาราง 4 อธิบายได้ว่า เครื่องจักรสร้างประโยคแบบไม่พึงบริบทได้จำนวนรวม 7.6392×10^{16} ประโยค โดยที่จำนวนของประโยคที่ได้ขึ้นตรงกับจำนวนของเรคคอร์ดที่มีในแต่ละรีเลชันและจำนวนของคำในแต่ละรูปแบบ ซึ่งผลการคำนวณที่ได้จะเป็นไปตามผลคูณคาร์ทีเซียนของรูปแบบครอสโปรดัก

id	words	words_cat	id	words	words_cat	id	words	words_cat	id	words	words_cat
▶ 2	การตอกเสียง	N	▶ 1	ดังกล่าวข้างต้น	PRON	▶ 11	เป็นหลักประกัน	V	▶ 13	ในบางกรณี	DET
4	เทคโนโลยีสมัยใหม่	N	2643	ใครก็ตาม	PRON	25	อุดหนุน	V	43	ในอีกกรณีหนึ่ง	DET
5	ระเบียบวาระการประชุม	N	3461	ไอ้	PRON	27	เพื่อประโยชน์ของ	V	5039	บุก	DET
6	ฮาร์โมนิคสกาย	N	3465	ไอ้หนู	PRON	28	รีบทรัพย์สิน	V	6992	ละ	DET
7	เอเชีย	N	4224	บางท่าน	PRON	31	คุ้มครองสิทธิ์	V	7940	ไหน	DET
9	อาเซียน	N	4303	ลื้อ	PRON	37	เตรียมรับมือ	V	8500	โน้น	DET
14	อำนาจหน้าที่	N	4313	ล้วนแต่	PRON	53	อ้างคำพูด	V	10469	ข้างต้น	DET
15	รถยนต์เกียร็อคโมติ	N	4314	ล้วนแล้วแต่	PRON	54	เรียกค่าเสียหาย	V	13084	ปวง	DET
16	เจตนาไม่บริสุทธิ์	N	4853	มหาราช	PRON	56	บาดเจ็บสาหัส	V	16860	ผอง	DET
id	words	words_cat	id	words	words_cat	id	words	words_cat	id	words	words_cat
▶ 8	เอเชีย	ADJ	▶ 3	ตามกฎหมาย	ADV	▶ 175	ด้วยเหตุฉะนี้	CONJ	▶ 3320	สำหรับ	PREP
10	อาเซียน	ADJ	12	แค่บางที	ADV	195	ครบไค	CONJ	4333	สู่	PREP
33	รายบุคคล	ADJ	39	มากยิ่งขึ้น	ADV	231	ถ้าหาก	CONJ	4527	สำหรับ	PREP
34	ระหว่างชาติ	ADJ	42	ในนามของ	ADV	247	ทำนองเดียวกัน	CONJ	4824	ตรงกลาง	PREP
35	ระหว่างประเทศ	ADJ	109	ภัยอยู่	ADV	260	หาหรือไม่	CONJ	4355	ภายใน	PREP
60	คมเข้ม	ADJ	125	ภูเขาเลากา	ADV	2161	ครบเท่าที่	CONJ	5185	แต่	PREP
143	งามสง่า	ADJ	157	จากนี้	ADV	2260	ในทางกลับกัน	CONJ	5688	ยัง	PREP
265	น่าใจหาย	ADJ	185	ตุ๊กตัก	ADV	2723	ยกตัวอย่างเช่น	CONJ	5815	ระหว่าง	PREP
268	น่าใจ	ADJ	186	ตุ๊กตัก	ADV	2863	แม้	CONJ	7862	เหนือ	PREP
id	words	words_cat	id	words	words_cat	id	words	words_cat	id	words	words_cat
▶ 175	ด้วยเหตุฉะนี้	CONJ	▶ 164	ข้างมันเถอะ	INT						
195	ครบไค	CONJ	262	นั่นสิ	INT						
231	ถ้าหาก	CONJ	419	ทับคำ	INT						
247	ทำนองเดียวกัน	CONJ	756	ท่า	INT						
760	หาหรือไม่	CONJ	1315	ว้าว	INT						
2161	ครบเท่าที่	CONJ	1348	อู้อู	INT						

ภาพ 5 ผลการทดลองสร้างรีเลชันแต่ละ Rq (เลือกเอาเฉพาะที่เกี่ยวข้อง—N, V, ADJ, BE, PP, ADV,)

ตาราง 3

ตัวอย่างผลการทดลองของการสร้างประโยคจากการครอสโปรดักตามรูปแบบ T1-T21

รูปแบบ	ตัวอย่างประโยคที่อ่านได้ความหมายถูกต้อง	ตัวอย่างประโยคที่อ่านไม่ได้ความหมาย
T ₁ (N+ADJ)	หวัตรจักร มหัศจรรย์	ขนมเบื้อง สมรรถ
T ₂ (N+BE+ADV)	นักอนุรักษ์นิยม เสมือนว่า ลงมา	พระนม เป็น เท่าทุน
T ₃ (N+BE+NP1)	การชะลอ คือ การงด	คลังน้ำมัน เป็น การยึดมั่น
T ₄ (N+V)	การเล่นพรรคเล่นพวก โหดเสียง	ถึงน้ำ เล่น แอลบ
T ₅ (N+V+PP)	ในกรณี ที่ พูด แทรกผู้อื่น	เชียงใหม่ หาบ พิตร
T ₆ (N+V+NP1)	เจ้าเบี้ย เปิดโปง ผู้พิทักษ์สันติราษฎร์	ภาชีรถยนต์ รดน้ำ ต้นไม้ ความอยากรู้
T ₇ (N+V+NP1+NP2)	หนังสือสำคัญ ยอมให้ การโปรย นักโปรแกรม	กอล์ฟ บาดเจ็บ ความโอบอ้อมอารี ผู้มีส่วน
T ₈ (NP1+ADJ)	การพ่วง วิจิตรบรรจง	ความขึ้น ปั่น แบน
T ₉ (NP1+BE+ADV)	การทำลายล้าง คือ เข้าไปซ้ำมา	นักการศึกษา คือ เกินควร
T ₁₀ (NP1+BE+NP2)	ความลึกลับ คือ การสมมติ	ห้องรับแขก อยู่ การสน

ตาราง 3 (ต่อ)

รูปแบบ	ตัวอย่างประโยคที่อ่านได้ความหมายถูกต้อง	ตัวอย่างประโยคที่อ่านไม่ได้ความหมาย
T ₁₁ (NP1+V)	พนักงานเก็บเงิน เบี่ยน้อยหยอยน้อย	การบนบานศาลกล่าว จ้อคอหยอย
T ₁₂ (NP1+V+PP)	การฝึกฝน จับ ไก่อ	สถานที่ประกอบการ หมด เขต
T ₁₃ (NP1+V+NP2)	งานเฉลิมฉลอง จูโจม เจ้าพนักงานราชทัณฑ์	ใบกำกับภาษี ออก จั่ว ผู้พิมพ์จำหน่าย
T ₁₄ (NP1+V+NP2+NP3)	ผู้ส่งสาส์น แทง ข้างหลัง การซื้อขายปัญญาประดิษฐ์	คำพิพากษา ถึงที่สุด อ่า สิ่งละเล็กละน้อย ลูกดอก
T ₁₅ (V+NP1)	มุ่งเน้น การศึกษานอกระบบ	กระโดดกระเดก ทวารประดับ
T ₁₆ (V+NP+ N)	มหัศจรรย์ใจ การอบรม การอบรมทหารประจำการ	อาสา การแผ่รังสี การแผ่รังสี
T ₁₇ (NP1+NP2+V)	ตัวประกอบ การค้า เจาะตลาด	การอาบน้ำ ระดับปานกลาง คุ่มกะลาหัว
T ₁₈ (V+NP1+NP2)	เปรี้ยวจัด ผู้กุมอำนาจ หน่วยงาน	เข้ามา กึ่ง ก้าน สาขา สาระประโยชน์
T ₁₉ (V+NP1+NP2+NP3)	เข้า สังคม การเข้าสู่ การเซ็นสัญญา ข้าราชการประจำ	ตกฟาก อำนาจ คำขอบคุณ ความเค็ม
T ₂₀ (NP1+NP2+V+NP3)	ข้อใหญ่ใจความ การเกษตร เติบโต ใหญ่ ปัญญาประดิษฐ์	คำชวน การเพื่อ ชม นามหลัก
T ₂₁ (NP1+NP2)	เงินวางประกัน การไฟฟ้าฝ่ายผลิตแห่งประเทศไทย	รพยบาล ผู้กำกับ การแสดง

ตาราง 4

ปริมาณประโยคที่ได้จากการครอสโปรดักจำแนกตามรูปแบบ T1-T21 และจำนวนคำในแต่ละรหัสชั้น

รูปแบบ	จำนวนคำ				ประโยคที่ได้
T ₁	21,090	2,879	-	-	6.0718x107
T ₂	21,090	4	2,876	-	2.4262x108
T ₃	21,090	4	10,545	-	8.8958x108
T ₄	21,090	13,028	-	-	2.7476x108
T ₅	21,090	13,028	221	-	6.0722x1010
T ₆	21,090	13,028	10,545	-	2.8973x1012
T ₇	21,090	13,028	10,545	10,545	3.0553x1016
T ₈	10,545	2,879	-	-	3.0359x107
T ₉	10,545	4	2,876	-	1.2131x108
T ₁₀	10,545	4	10,545	-	4.4479x108
T ₁₁	10,545	13,028	-	-	1.3738x108
T ₁₂	10,545	13,028	221	-	3.0361x1010
T ₁₃	10,545	13,028	10,545	-	1.4487x1012
T ₁₄	10,545	13,028	10,545	10,545	1.5276x1016
T ₁₅	13,028	10,545	-	-	1.3738x108
T ₁₆	13,028	10,545	21,090	-	2.8973x1012

ตาราง 4 (ต่อ)

รูปแบบ	จำนวนคำ				ประโยคที่ได้
T ₁₇	10,545	10,545	13,028	-	1.4487x1012
T ₁₈	13,028	10,545	10,545	-	1.4487x1012
T ₁₉	13,028	10,545	10,545	10,545	1.5276x1016
T ₂₀	10,545	10,545	13,028	10545	1.5276x1016
T ₂₁	10,545	10,545	-	-	1.1120x108
รวม	7.6392x1016				

ผลการทดลองเชิงคุณภาพ

ตาราง 5 เป็นผลเชิงคุณภาพที่นำข้อมูลที่ได้จากการครอสโปรดักแมววิเคราะห์เพื่อหาจำนวนของประโยคที่สามารถอ่านได้ความหมายถูกต้อง จำแนกเป็นร้อยละ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ส่วนตาราง 6 แสดงการเปรียบเทียบจำนวนร้อยละเฉลี่ยของการพิจารณาประโยคที่อ่านได้ความหมายถูกต้องเมื่อพิจารณาตามประโยคที่ประกอบด้วยคำ 2 3 และ 4 คำ

ทั้งสองตารางนำเสนอ ดังนี้ ผลการทดลองเป็นผลการทดลองสุ่มประโยคอย่างอิสระโดยใช้โปรแกรมคอมพิวเตอร์ ด้วยการเขียนโปรแกรมเพื่อสุ่มประโยคในตาราง 1 แต่ละรูปแบบ (21 รูปแบบ) จำนวน 12 ชุด ชุดละ 25 ประโยค (300 ประโยคต่อชุด) รวมทั้งสิ้น 6,300 ประโยค จากนั้นวิเคราะห์ผลลัพธ์ของแต่ละรูปแบบด้วยคณะผู้วิจัย ที่นำออกมาวิเคราะห์ทีละประโยค ด้วยการให้ความเห็นของคณะผู้วิจัยตัดสินว่าประโยคที่สร้างขึ้นสามารถอ่านได้ความหมายถูกต้องหรือไม่ ทั้งนี้ได้กำหนดบริบทตายตัวสำหรับจัดประเภทเพียงแต่ตัดสินว่าสามารถอ่านได้ใจความหรือไม่เท่านั้น

ผลการวิเคราะห์โดยรวม (macro analysis) พบว่า สามารถอ่านได้ความหมายถูกต้องจำนวน 2,312 ประโยค คิดเป็นร้อยละ 36.67 ผลการวิเคราะห์ดังกล่าวแสดงดังตาราง 5

พิจารณาจากตาราง 5 อธิบายได้ว่า ผลเชิงคุณภาพที่สามารถอ่านได้ความหมายถูกต้อง โดยเฉลี่ยทั้ง 21 รูปแบบ รูปแบบละ 300 ประโยค จำนวนที่อ่านความหมายได้ถูกต้องได้เฉลี่ย 110.10 ประโยค คิดเป็นร้อยละ 36.70 โดยมีค่าน้อยสุดอยู่ที่ร้อยละ 13.33 คือ รูปแบบประโยคที่ 14 ค่ามากที่สุดได้ถึงร้อยละ 64 คือ รูปแบบประโยคที่ 1 และมีค่าจำนวนประโยคเฉลี่ยจาก 25 ประโยคในแต่ละชุดข้อมูลคือ 9.17 ประโยค

วิเคราะห์ในส่วนรายละเอียด (micro analysis) นำผลการทดลองในตาราง 5 มาวิเคราะห์เพิ่มเติม โดยเปรียบเทียบความยาวของประโยคที่ประกอบด้วย 2 คำ 3 คำ และ 4 คำ โดยประโยค 2 คำ คือ รูปแบบที่ 1 4 8 11 15 และ 21 ประโยคแบบ 3 คำ คือ รูปแบบที่ 2 3 5 6 9 10 12 13 16 17 และ 18 ประโยคแบบ 4 คำ คือ รูปแบบที่ 7 14 19=20 ผลการทดลองที่ได้พบว่า เมื่อประโยคที่ประกอบด้วยคำ 2 คำ สามารถอ่านได้ความหมายถูกต้องมีร้อยละเฉลี่ยระหว่าง 44.00-64.00 คิดเฉลี่ยเป็นร้อยละ 53.05 ขณะที่ประโยคที่ประกอบด้วยคำ 3 คำ ได้ร้อยละเฉลี่ยระหว่าง 22.33-57.67 คิดเป็นค่าเฉลี่ยร้อยละ 34.57 และประโยคที่ประกอบด้วยคำ 4 คำ พบว่า ได้ประโยคที่มีความหมาย ร้อยละเฉลี่ยระหว่าง 13.33-21.00 คิดเป็นค่าเฉลี่ยร้อยละ 18.00 แสดงร้อยละเฉลี่ยดังตาราง 6

ตาราง 5

ผลรวม ร้อยละ และส่วนเบี่ยงเบนมาตรฐานของประโยคแบบ T_1-T_{21} ที่สามารถอ่านความหมายได้ถูกต้อง

รูปแบบประโยค	ผลรวม	ร้อยละ	$\bar{X}$	SD
T_1	192.00	64.00	16.00	3.07
T_2	116.00	38.67	9.67	3.26
T_3	104.00	34.67	8.67	2.90
T_4	168.00	56.00	14.00	3.64
T_5	173.00	57.67	14.42	5.47
T_6	109.00	36.33	9.08	2.27
T_7	63.00	21.00	5.25	2.26
T_8	166.00	55.33	13.83	4.02
T_9	101.00	33.67	8.42	4.64
T_{10}	98.00	32.67	8.17	4.99
T_{11}	132.00	44.00	11.00	4.00
T_{12}	146.00	48.67	12.17	4.15
T_{13}	76.00	25.33	6.33	2.77
T_{14}	40.00	13.33	3.33	2.39
T_{15}	138.00	46.00	11.50	3.78
T_{16}	67.00	22.33	5.58	2.61
T_{17}	76.00	25.33	6.33	1.92
T_{18}	75.00	25.00	6.25	2.30
T_{19}	51.00	17.00	4.25	3.28
T_{20}	62.00	20.67	5.17	3.97
T_{21}	159.00	53.00	13.25	2.70
เฉลี่ย	110.10	36.70	9.17	3.35

ตาราง 6

ผลร้อยละเฉลี่ยเมื่อพิจารณาความยาวของคำ

ความยาวคำ	ผลรวมเฉลี่ย	$\bar{X}$	ร้อยละ	SD
2	159.17	13.26	53.05	3.54
3	103.73	8.64	34.57	3.39
4	54.00	4.50	18.00	2.98

อภิปรายผล

จากผลการทดลองเชิงปริมาณที่ได้จากการนำคำเข้าเครื่องจักรโดยประมาณ 3×10^4 คำ ทำให้ได้ประโยคที่สามารถอ่านได้ความหมายถูกต้องอยู่ที่ร้อยละ 13.33-64 หรือ 1.0183×10^{16} - 4.8891×10^{16} ประโยค ค่าเฉลี่ยโดยรวมร้อยละ 36.70 คิดเป็นประโยคที่ได้ 2.8036×10^{16} ประโยค ซึ่งเป็นปริมาณประโยคจำนวนมาก ดังนั้นหากมีการคำเข้าสู่เครื่องจักรที่สร้างขึ้นมากกว่าการทดลองในงานวิจัยนี้จะทำให้ได้ปริมาณของประโยคที่มากยิ่งขึ้น อย่างไรก็ตาม ยังมีจำนวนคำมากขึ้นจำเป็นต้องอาศัยขีดความสามารถของเครื่องคอมพิวเตอร์มากยิ่งขึ้น รวมถึงจำเป็นต้องพิจารณาขีดความสามารถของระบบจัดการฐานข้อมูลที่สูงขึ้นอีกด้วย ฉะนั้น จำเป็นสำหรับผู้ที่จะใช้งานขั้นตอนวิธีและเครื่องจักรที่สร้างขึ้นจากงานวิจัยนี้ จำเป็นต้องปรับให้เหมาะกับข้อมูลนำเข้าของตนให้เหมาะสม

จากผลการทดลองเชิงคุณภาพ หากพิจารณาโดยจำแนกประโยคที่เกิดจากรูปแบบของประโยคที่ประกอบด้วยคำ 2 คำ ได้ 7.5180×10^8 ประโยค สามารถเข้าใจความหมายถูกต้องได้ร้อยละ 53.05 คิดเป็น 3.9883×10^8 ประโยค ประโยคที่ประกอบด้วยคำ 3 คำ ได้ 1.0234×10^{13} ประโยค สามารถเข้าใจความหมายได้เฉลี่ยร้อยละ 34.57 คิดเป็น 3.5377×10^{12} ประโยค พิจารณาประโยคที่ประกอบด้วยคำ 4 คำ ได้ประโยคจำนวน 7.6381×10^{16} ประโยค สามารถเข้าใจความหมายได้เฉลี่ยร้อยละ 18.00 คิดเป็น 1.3749×10^{16} ประโยค ในกรณีนี้ เมื่อพิจารณาเปรียบเทียบค่าเฉลี่ยร้อยละของประโยคที่มีความยาวและสั้น จะพบว่าประโยคสั้น (ประโยคที่ประกอบด้วยคำน้อย) จะมีค่าเฉลี่ยมากกว่าประโยคที่มีความยาว (ประโยคที่ประกอบด้วยคำมาก) แต่หากพิจารณาให้ละเอียดลงไปในปริมาณที่ได้ของประโยค พบว่า แม้ว่าค่าเฉลี่ยร้อยละจะน้อยกว่าแต่ความจริงแล้วจำนวนประโยคมีมากขึ้นซึ่งสวนทางกันเนื่องจากการใช้วิธีการครอสโปรดักที่ได้ปริมาณประโยคเป็นแบบผลคูณคาร์ทีเซียนนั่นเอง ยกตัวอย่างจากผลการทดลองนี้ เมื่อเปรียบเทียบประโยคที่ประกอบด้วยคำ 2, 3 และ 4 คำ ค่าเฉลี่ยร้อยละ 53.05 34.57 และ 18.00 แต่ปริมาณประโยคอยู่ที่ 3.9883×10^8 1.0234×10^{13} และ 1.3749×10^{16} เป็นต้น

เครื่องจักรที่สร้างขึ้นสามารถทำได้ดีหากยึดปริมาณของประโยคเป็นหลัก แต่หากพิจารณาคุณภาพเป็นหลัก โดยเฉพาะความยาวขนาด 2 คำ และ 3 คำ จะมีค่าร้อยละการอ่านได้ความหมายถูกต้องสูง หากจำนวนคำมากในประโยคทำให้ค่าเฉลี่ยของการอ่านได้ของประโยคจากการคิดจากประโยคทั้งหมดจะได้น้อยลงตามลำดับ ข้อด้อยของเครื่องจักรนี้ คือ เป็นการสร้างประโยคเป็นพื้นฐาน ยังไม่ได้มีการกำหนดบริบทการสร้าง หรือคิดกระบวนการที่ซับซ้อนที่ทำให้ได้ประโยคตามวัตถุประสงค์ของการใช้งาน วิธีการนี้จำเป็นจะต้องหากระบวนการที่สามารถกรองหรือสร้างเงื่อนไขเพิ่มเติมในการสร้างประโยคซึ่งอาจทำให้ค่าเฉลี่ยของการอ่านได้สูงขึ้นก็ได้ หรือหากต้องการให้ได้ตามวัตถุประสงค์ ควรนำงานวิจัย สร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสาร (Khancome, Daengcharoen & Charungchit, 2022) มาผนวกหรือประยุกต์คาดว่าจะทำให้ได้ประโยคที่มีประสิทธิภาพมากยิ่งขึ้น

สรุปผลการวิจัย

งานวิจัยนี้พัฒนาเครื่องจักรซอฟต์แวร์สำหรับสร้างประโยคภาษาไทยเก็บไว้ในคลังประโยคภาษาไทย เพื่อสนับสนุนการสร้างเรื่องราวในสตอรี่เจนเนอเรชัน ออกแบบโดยอาศัยตัวดำเนินการครอสโปรดักในฐานข้อมูลเชิงสัมพันธ์และใช้พีชคณิตเชิงสัมพันธ์เป็นกฎการสร้างประโยค ตามรูปแบบไวยากรณ์ ทดลองจากคำในพจนานุกรมไทย-อังกฤษ LEXITRON ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ สามารถสร้างประโยคที่สามารถอ่านได้ความหมายถูกต้องตามรูปแบบประโยคไวยากรณ์ภาษาไทย 21 รูปแบบ ได้จำนวนเรคอร์ดหรือประโยคแบบผลคูณคาร์ทีเซียน ซึ่งขึ้นกับจำนวนรีเลชันจำนวนเรคอร์ด จำนวนคำในแต่ละรีเลชัน และจำนวนรูปแบบที่นำเข้าเครื่องจักร โดยการศึกษาวิจัยนี้ได้ปริมาณประโยครวม 7.6392×10^{16} ประโยค ได้ประโยคแบบฟังก์ชันตั้งแต่ร้อยละ 13.33-64 เครื่องจักรสร้างประโยคที่สามารถอ่านได้ความหมายถูกต้องที่ค่าเฉลี่ยร้อยละ 53.05 34.57 และ 18.00 เมื่อพิจารณาประโยคที่ประกอบด้วยคำ 2 คำ 3 คำ และ 4 คำ ตามลำดับ ทั้งนี้หากจัดเตรียมคำชนิดต่าง ๆ ด้วยเงื่อนไขและรูปแบบไวยากรณ์ที่หลากหลายมากขึ้น และ

คำแนะนำที่เกี่ยวข้องกับบริบทที่ต้องการมากขึ้น คาดว่าน่าจะได้ประโยคที่สามารถอ่านความหมายได้ถูกต้องมากขึ้นตามไปด้วย

ข้อเสนอแนะ

แนวคิดจากงานวิจัยนี้ทำให้เกิดการสร้างประโยคภาษาไทยได้เป็นจำนวนมาก เหมาะสำหรับนำไปต่อยอดเพื่อสกัดเอาประโยคที่ดีสามารถนำเข้าสู่การเรียนรู้เครื่องจักร (machine learning) หรือการเรียนรู้เชิงลึก (deep learning) เพื่อใช้เรียนรู้ รวมถึงการนำเข้าเพื่อสอนและทดสอบ (train

and test) ในตัวแบบภาษาธรรมชาติ (natural language model) ได้ และสามารถนำไปต่อยอดสู่การพัฒนาเป็นตัวแบบสร้างประโยคภาษาไทยหรือปรับใช้สู่ภาษาอังกฤษและภาษาอื่น ๆ ได้

การวิจัยในลำดับต่อไปควรมีรูปแบบและตัวแบบในการสกัดเอาประโยคแบบพึงบริบทได้ หรือกำหนดรูปแบบการสร้างประโยคตามแนวทางมาร์คอฟเชน (markov chain) โดยการกำหนดค่าความน่าจะเป็นร่วมกับการโครสโปรดักเพื่อสร้างประโยค ซึ่งน่าจะได้ความแม่นยำในเชิงความหมายได้ถูกต้องมากยิ่งขึ้น



References

- Bui, V., Abbass, H. S., & Bender, A. (2010). Evolving stories: Grammar evolution for automatic plot generation. *IEEE Congress on Evolutionary Computation 2010* (pp. 1-8). Barcelona, Spain: IEEE doi: 10.1109/CEC.2010.5585934
- Daengcharoen, D., Charungchit, K., & Khancome, C. (2020). Thai sentence generation engine for Thai sentences corpus. *Proceedings of the 3rd Conference on Innovation Engineering and Technology for Economy and Society 2020* (pp. 199-204). Bangkok: Kasem Bundit University (in Thai)
- Dekpituksirikul, K. (2008). *A comparative study of phrases, sentences and discourses of normal and tistic children* (Master's thesis). Silpakorn University. Nakhon Pathom (in Thai)
- Elmasri, R., & Navathe, S. (2010). *Fundamentals of database system* (4th ed.). USA.: Pearson Education Inc.
- Grandi, F., Mandreoli, F., martoglia, R., & Penzo, W. (2017). A relational algebra for streaming tables living in a temporal database world. *24th International Symposium on Temporal Representation and Reasoning (TIME 2007)* (pp. 1-15). Wadern: Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH, Wadern
- Hamburg University. (2021). *Story generator algorithms*. Retrieved from http://wikis.sub.uni-hamburg.de/lhn/index.php/Story_Generator_Algorithms.
- Jampaibool, T., & Arunmanakul, V. (2016). The complexity of the index noun phrase indicates the complexity of the text: A case study of textbooks for Thai subjects Grade 1-3. *Humanities Journal*, 23(2), 148-177. (in Thai)
- Khancome, C., Daengcharoen, D., & Charungchit, K. (2022). The machine creates Thai sentences according to the purpose of communication. *The 14th NPRU National Academic Conference* (pp. 612-623). Nakhon Pathom Nakhon: Pathom Rajabhat University (in Thai)

- Krukaset, W., Krukaset, N., & Khancome, C. (2017). Thai sentence generation machine employing fixed patterns". *2017 IEEE International Conference on High Performance Computing and Communications Workshops* (pp. 70-73). Bangkok: IEEE. (in Thai)
- Kybartas, B., & Bidarra, R. (2010). A survey on story generation techniques for authoring computational narratives. *IEEE Trans. on Computational Intelligence and AI in Games*, 3(5), 776-786.
doi:10.1109/TCIAIG.2016.2546063
- LEXiTRON Data. (2018). *Lecitron data is an ideal database for people to develop or study*. Retrieved from https://lexitron.nectec.or.th/2009_1/index.php?q=common_manager/download#latest_version.
- Li, C., Chen-Chuan Chang, K., Ilyas, I. F., & Song, S. (2005). RankSQL: Query algebra and optimization for relational top-k queries. USA: SIGMOD 2005 Bultimore Maryland
- Limpanadudadee, W., Punyabukkanna, P., & Poobrasert, O. (2014). Text corpus for natural language story-telling sentence generation: A design and evaluation. *11th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 80-85). Bangkok: IEEE. (in Thai)
- Meehan, J. R. (1976). *The metanovel: Write stories by computer* (Doctoral dissertation). Yale University. USA.
- Meehan, J. R. (1977). Tale-Spin, an interactive program that writes stories. *Proceedings of the 5th International Joint Conference on Artificial Intelligence* (pp. 91-98). USA: Morgan Kaufmann Publishers Inc.
- Ontanon, S., & Zhu, J. (2011). The SAM algorithm for analogy-based story generation. *Proceedings of the 9th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 67-72). USA: AIIDE
- Thongkaol, S. (2007). *Artificial Intelligence*. Songkhla: Faculty of Science and Technology Songkhla Rajabhat University. (in Thai)

