เครื่องจักรสร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสาร

เชาวลิต ขันคำ[1*], สุพรัตรา แดงเจริญ[2] และ ขณิดา จรุงจิตต์[3]

[1]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง
[2,3]สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏราชนครินทร์
*chouvalit.k@rumail.ru.ac.th

บทคัดย่อ

บทความวิจัยนี้นำเสนอเครื่องจักรซอฟต์แวร์สำหรับสร้างประโยคภาษาไทยตามวัตถุประสงค์ของการสื่อสาร ได้แก่ ประโยคบอกเล่า ปฏิเสธ คำถาม และประโยคคำสั่งหรือขอร้อง เพื่อเก็บไว้ในคลังประโยคภาษาไทยสำหรับนำไปใช้สร้างเท็กซ์ในงานวิจัยสตอรีเจนเนอเรชันหรือนำเข้าสู่การเรียนรู้และสร้างตัวแบบทางภาษาต่อไป กลไกการออกแบบเครื่องจักรใหม่อาศัยฐานข้อมูลเชิงสัมพันธ์ ใช้ฟังก์ชันทางพีชคณิตเชิงสัมพันธ์แบบครอสโปรดักเป็นกฎควบคุมการสร้างประโยคตามรูปแบบไวยากรณ์ภาษาไทย การทดลองทางทฤษฎีเบื้องต้นนำเข้าประโยคบอกเล่า 14 รูปแบบ จำนวน 4.506×10^{16} ประโยค รูปแบบประโยคตามวัตถุประสงค์การสื่อสาร จำนวน 8 รูปแบบ ผลการทดลองทางทฤษฎีพบว่า เครื่องจักรสามารถสร้างประโยคแบบไม่พึ่งบริบทได้ 1.09928×10^{18} ประโยค แบ่งเป็นประโยคคำถาม 6.41245×10^{17} ประโยคปฏิเสธ 2.74819×10^{17} และประโยคคำสั่งหรือขอร้อง 1.83213×10^{17} ตามลำดับ

คำสำคัญ: การประมวลผลภาษาธรรมชาติ การสร้างประโยคภาษาไทย การสร้างเรื่องราวภาษาไทย เครื่องจักรสร้างประโยคภาษาไทย ความสัมพันธ์แบบครอสโปรดัก

THAI SENTENCES GENERATION ENGINE DEPENDED ON COMMUNICATION OBJECTS

Chouvalit Khancome[1*], Suphrattra Daengcharoen[2] and Kanida Charungchit[3]

[1]Computer Science Department Faculty of Science Ramkhamhaeng University
[2,3] Information Technology Department Faculty of Science and Technology Rajabhat Rajanagarindra University
*chouvalit.k@rumail.ru.ac.th

Abstract

This research paper presents a new software machine for generating Thai sentences according to the objectives of communication: declarative sentences, negative sentences, question sentences, and command or request sentences. Then, those sentences will be stored in a corpus of Thai sentences for use in Text Generation and Story Generation research or Machine Learning and building a language model. The inner mechanisms of machine relies on relational databases using a cross-product relational algebraic function as a rule to control sentence construction followed by Thai grammatical patterns. The preliminary theoretical experiments imported 14 Thai sentence patterns and 4.506×10^{16} sentences with 8 Thai sentence patterns of the objective communication. The results of the theoretical experiments found that the machine was able to create a context-free sentence of 1.09928×10^{18}, divided into the number of questions 6.41245×10^{17}, negative sentences 2.74819×10^{17}, and command or request sentences 1.83213×10^{17}, respectively.

Keywords: Natural Language Processing, Thai Sentence Generation, Thai Story Generation, Thai Sentence Generation Engine, Cross-product Relation

1. บทนำ

การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) เป็นวิทยาศาสตร์ในปัญญาประดิษฐ์ (Artificial Intelligent) มีจุดมุ่งหมายเพื่อช่วยให้คอมพิวเตอร์เข้าใจตลอดจนตีความและใช้ภาษามนุษย์ในการสื่อสาร การศึกษาการประมวลผลภาษาธรรมชาติต้องเกี่ยวข้องกับประโยค (sentence) ที่ประกอบด้วยคำชนิดต่างในภาษา โดยเฉพาะอย่างยิ่งในการวิจัยเกี่ยวกับการสร้างเท็กซ์ (Text Generation) หรือ สร้างเรื่องราว (Story Generation) ที่ทำให้คอมพิวเตอร์สามารถสร้างเรื่องราวต่างๆ ตามความต้องการของมนุษย์ได้โดยอัตโนมัติ นำมาประยุกต์ใช้เพื่อสร้างเรื่องเล่า นิทาน นวนิยาย วรรณกรรม สารคดี บทความวิชาการ และงานทางด้านภาษาศาสตร์อื่นๆ เป็นต้น การวิจัยตามแนวคิดนี้จำเป็นต้องกลไกการนำคำชนิดต่างๆ (parts of speech) วลี (phrases) กลุ่มคำ (clauses) มาสร้างประโยคประกอบกันเพื่อให้ได้เนื้อความเชิงความหมายที่มนุษย์สามารถเข้าใจได้ งานวิจัยเกี่ยวกับการสร้างประโยคโดยเฉพาะภาษาไทยยังค่อนข้างมีน้อย พบใน [1] ซึ่งออกแบบและทดลองเกี่ยวกับการสร้างคลังของเท็กซ์ภาษาไทยเพื่อใช้ในการสร้างเรื่องเล่า งานวิจัย [2]

นำเสนอเครื่องจักรสำหรับการสร้างประโยคภาษาไทยด้วยการกำหนดรูปแบบตายตัว [3] สร้างโครงสร้างภาษาใหม่สำหรับจัดเก็บโครงสร้างของคำเพื่อนำไปสู่การสร้างประโยคและวลีต่างๆ และ [4] สร้างเครื่องจักรซอฟต์แวร์ใหม่สำหรับสร้างประโยคโดยใช้กลไกของฐานข้อมูลเชิงสัมพันธ์เป็นต้นแบบ เป็นต้น

แรงบันดาลใจของคณะผู้วิจัยเกิดจากงานวิจัยของสุพรัตรา แดงเจริญและคณะ [4] ซึ่งนำเสนอเครื่องจักรทางซอฟต์แวร์ใหม่สำหรับสร้างประโยคภาษาไทยอัตโนมัติเพื่อเก็บไว้ในคลังประโยค (Sentence Corpus) เพื่อนำไปใช้สำหรับสตอรีเจนเนอเรชัน เครื่องจักรนี้นำคำภาษาไทยซึ่งแบ่งตามหน้าที่ของคำ (parts of speech) มาประกอบกันเป็นประโยคตามรูปแบบไวยากรณ์ภาษาไทย โดยอาศัยกลไกของฐานข้อมูลเชิงสัมพันธ์ (Relational Database) ควบคุมการสร้างประโยคด้วยฟังก์ชันทางพีชคณิตเชิงสัมพันธ์ (Relational Algebra) เป็นกฎในการสร้างประโยค ทดลองด้วยคำจากพจนานุกรม LEXiTRON [5] จำนวน $3x10^4$ คำ กำหนดรูปแบบประโยคตาม 14 รูปแบบ ผลการทดลองได้ประโยคตามรูปแบบไวยากรณ์โดยไม่พึ่งบริบท (context free sentence) มากถึง $4.506x10^{16}$ ประโยค พิจารณาประโยคเชิงความหมาย (contextual sentence) สามารถอ่านเข้าใจมากที่สุดถึง 5-60 เปอร์เซ็นต์ ข้อด้อยของงานวิจัยคือเครื่องจักรสามารถสร้างประโยคได้เพียงแค่ประโยคเดี่ยว (Single Sentence) ซึ่งสามารถเป็นเพียงประโยคบอกเล่าเท่านั้น ยังไม่สามารถสร้างประโยคตามวัตถุประสงค์หรือเจตนาการสื่อสารอื่นๆ ได้แก่ ประโยคคำถาม ประโยคปฏิเสธ ประโยคคำสั่งหรือข้อร้องได้

ดังนั้นงานวิจัยใหม่นี้ จึงมุ่งเน้นจัดการกับข้อด้อยของงานวิจัย [4] โดยออกเครื่องจักรซอฟต์แวร์ใหม่เพื่อต่อขยายความสามารถของเครื่องจักร [4] ให้สามารถสร้างประโยคภาษาไทยตามเจตนาการสื่อสาร (บอกเล่า, คำถาม, ปฏิเสธ, คำสั่งหรือขอร้อง) ได้ สร้างกลไกภายในซึ่งประกอบด้วยอัลกอริทึมการสร้างประโยคและใช้แนวคิดของฐานข้อมูลเชิงสัมพันธ์อาศัยฟังก์ชันทางพีชคณิตเชิงสัมพันธ์แบบครอสโปรดัก (Cross Product) เป็นกฎควบคุมการสร้างประโยคตามรูปแบบไวยากรณ์ภาษาไทย ทำให้สามารถสร้างประโยคภาษาไทยตามเจตนาการสื่อสารได้ทุกประเภทของประโยคภาษาไทย

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ประโยคภาษาไทย

ประโยคคือหน่วยสมบูรณ์ทางไวยากรณ์ [6] สามารถจำแนกได้ตามสถานการณ์การใช้งานและจำแนกตามโครงสร้างได้ 4 แบบ ได้แก่ ประโยคสามัญ ประโยคซับซ้อน ประโยคประสม และประโยคเชื่อม แต่หากพิจารณาโครงสร้างประโยคตามแนวทางการใช้งานภาษาธรรมชาติที่กล่าวไว้ใน [7] ที่ระบุว่าสามารถแบ่งลักษณะของไวยากรณ์ตามลักษณะของการจัดเรียงของคำในภาษาโดยเฉพาะภาษาไทยเป็นแบบ SVO โดยที่ S คือ ประธาน (Subject) V คือ กริยา (Verb) และ O คือ กรรม (Object) เมื่อนำโครงสร้างการใช้งานของประโยคภาษาไทยเทียบกับภาษาอังกฤษแล้วพบว่าภาษาไทยมีการใช้งาน 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) จากการวิเคราะห์รูปแบบภาษายังแสดงให้เห็นว่ามีรูปแบบภาษาไทยและภาษาอังกฤษที่สอดคล้องกันอยู่อย่างน้อย 7 รูปแบบ (ที่ถูกนำมาเสนอเพื่อทดลองในงานวิจัยนี้ ดังนี้ 1) NP ADJ 2) NP BE ADV 3) NP BE NP 4) NP V 5) NP V PP 6) NP V NP และ 7) NP V NP NP โดยที่ BE คือ คำกริยาประเภท เป็น อยู่ คือ (Verb to be) ADV คือ คำวิเศษณ์ (Adverb)

การแบ่งประโยคตามเจตนาการสื่อสาร [8] แบ่งประโยคออกได้ 4 ชนิดคือ ประโยคบอกเล่า ประโยคคำถาม ประโยคปฏิเสธ ประโยคคำสั่งและขอร้อง โดยที่ประโยคคำถามแบ่งออกเป็น 2 ลักษณะคือ 1) คำถามที่ต้องการให้ตอบรับ โดยรูปแบบจะมีประโยคบอกเล่าและ มักลงท้ายด้วยคำว่า หรือไม่ ใช่ไหม ไม่ใช่หรือ ในที่นี้เขียนด้วย S+คำลงท้าย (เช่น ใช่หรือไม่) เป็นต้น และ 2) คำถามที่ต้องการให้ตอบเป็นประโยคหรืออธิบายคำตอบ จะมีคำถามขึ้นต้นประโยค เช่นคำว่า ใคร เมื่อใด อย่างไร รูปแบบที่นำมาใช้งานวิจัยนี้คือ คำถาม+S เช่น ใครทำน้ำหกเลอะพื้นห้องเรียน เป็นต้น ในขณะที่ประโยคปฏิเสธ จำแนกออก

ได้ 2 ลักษณะเช่นกัน คือ 1) แบบแยกประโยคบอกเล่าปกติออกจากกันอย่างน้อยสองส่วน แล้วคั่นด้วยคำปฏิเสธเช่น มิ ไม่ หา มิได้ รูปแบบที่ใช้เช่น Sส่วนแรก + ไม่/มิ + S ส่วนที่สอง เช่น ประโยคปกติ นกกินหนอนใบไม้ แยกได้เป็น นก+ไม่+กินหนอนใบไม้ เป็นต้น 2) แบบนำคำปฏิเสธต่อท้ายประโยค เช่น ก็หาไม่ รูปแบบ S+คำปฏิเสธ เช่น นกกินหนอนใบไม้ก็หาไม่ เป็นต้น ส่วนประโยคคำสั่งหรือขอร้องจะ มีคำสั่ง/ขอร้อง +VP ซึ่ง แสดงตัวอย่างเช่น ห้าม + เดินลัดสนาม /กรุณา+ปิดเบาๆ เป็นต้น นอกจากนั้นยังสามารถแบ่งประโยคตามชนิดของประโยคสามารถแบ่งได้เป็น ประโยคความเดียว ประโยคความรวม และประโยคความซ้อน ซึ่งในการแบ่งตามชนิดนี้สามารถแบ่งประโยคได้เพิ่มขึ้นเป็นประโยคที่มีเนื้อความคล้อยตามกัน ประโยคที่มีเนื้อความขัดแย้งกัน และประโยคที่เป็นเหตุเป็นผลกัน อีกชนิดการแบ่งคือประโยคความซ้อนที่ประกอบด้วยประโยคหลักและประโยคย่อยซึ่งเป็นวลีชนิดต่างๆ

2.2 พีชคณิตเชิงสัมพันธ์

พีชคณิตฐานข้อมูล (Database Algebra) หรือ พีชคณิตเชิงสัมพันธ์หรือฐานข้อมูลเชิงสัมพันธ์ [9] ถูกเขียนขึ้นเพื่อให้สามารถมองเห็นถึงความสัมพันธ์ที่เกิดขึ้นของข้อมูลโดยใช้พีชคณิตอธิบายความสัมพันธ์ของตารางต่างๆ ในฐานข้อมูลสามารถประยุกต์เข้ากับงานวิจัยทางด้านคอมพิวเตอร์ได้อย่างหลากหลาย ตัวดำเนินการพื้นฐาน (Basic Operations) ของพีชคณิตเชิงสัมพันธ์มี 6 ประเภท ดังนี้

(1) ซีเล็กชัน (Select Operation---**σ** (ซิกมา)) ทำหน้าที่เป็นตัวเลือกแถวข้อมูลจากเงื่อนไขที่กำหนด รูปแบบคือ $\sigma_{condition}^{(Table)}$

(2) โปรเจคชัน (Project Operation--- **π** (พาย)) ทำหน้าที่เป็นตัวเลือกคอลัมน์ข้อมูลจากเงื่อนไขที่กำหนด รูปแบบ $\pi_{attribute_1,attribute_2,...,attritube_n}^{(Table)}$

(3) ครอสโปรดัก (Cross – Product--- x) หรือ คาร์ทีเซียนโปรดัก (Cartesian Product) ทำหน้าที่แสดงความสัมพันธ์ด้วยการคูณคาทีเซียน (Cartesian) ระหว่าง 2 รีเลชัน รูปแบบคือ R x S

(4) เซตดิฟเฟอเรน (Set – Difference--- —) ทำหน้าที่หาความแตกต่างระหว่าง 2 รีเลชันเพื่อแสดงข้อมูลของแถวที่ต่างจากรีเลชันแรก รูปแบบคือ R — S

(5) ยูเนียน (Union--- U) ทำหน้าที่นำข้อมูลจาก 2 รีเลชันมารวมกัน และหากมีข้อมูลซ้ำจะแสดงเพียงแถวเดียว รูปแบบคือ R U S

(6) อินเตอร์เซกชัน (Intersection---⋂) ทำหน้าที่นำข้อมูลจาก 2 รีเลชันมารวมกัน และแสดงเฉพาะข้อมูลที่เหมือนกันจากทั้ง 2 รีเลชันเท่านั้น รูปแบบคือ R ⋂ S

งานวิจัยนี้ใช้ตัวดำเนินการโปรเจคชันร่วมกับตัวดำเนินการคาร์ทีเซียนโปรดักหรือครอสโปรดักมากำหนดเป็นกฎในการสร้างประโยค เพื่อควบคุมการเลือกตารางและคำสำหรับสร้างประโยคตามรูปแบบไวยากรณ์

2.3 งานวิจัยที่ใช้เป็นฐานสำหรับงานวิจัยนี้

เพื่อความเข้าใจงานวิจัยใหม่ที่จะนำเสนอ จึงนำเอานิยามและอัลกอริทึมภายในและโครงสร้างของเครื่องจักร [4] มานำเบื้องต้น ดังนี้

นิยามที่ 1 กำหนดให้ P แทน เซตความสัมพันธ์ (Relations) ของคำที่ใช้ในภาษาไทย (Part Of Speech) หรือวลีที่มีใช้ในภาษาไทย (Phrases)

นิยามที่ 2 กำหนดให้ความสัมพันธ์ $R_1, R_2, R_3, \dots, R_n \in P$ ที่แต่ละ R_i มีจำนวนแถว (Tuple) เท่ากับ n_i แถว โดยที่ $1 \leq i \leq n$

นิยามที่ 3 กำหนดให้ TP แทน เซตของรูปแบบประโยคภาษาไทย (Thai Sentence Pattern)

นิยามที่ 4 กำหนดให้ $T_1, T_2, T_3, \dots, T_m \in TP$ โดยที่ แต่ละ T_j ประกอบด้วยลำดับของ ความสัมพันธ์ในนิยามที่ 2 ตั้งแต่หนึ่งความสัมพันธ์ขึ้นไป โดยที่ $1 \leq j \leq m$

นิยามที่ 5 กำหนด RS คือ เซตของความสัมพันธ์ใดๆ ที่เกิดจากการสร้างประโยคภาษาไทยตามรูปแบบใดๆ ตามนิยามที่ 4

นิยามที่ 6 กำหนดให้ $rs_1, rs_2, rs_3, \dots, rs_m \in RS$ เมื่อ rs_j แทน ตารางผลของการสร้างประโยคภาษาไทย ตามรูปแบบ T_j ใดๆ

รูปแบบประโยคนำเข้าเครื่องจักร [4] จำแนกรูปแบบและพีชคณิตเชิงสัมพันธ์ ดังนี้

(1) $T_1 \rightarrow \pi_{(}N^{(noun)}\ ADJ^{(adj)}$) N x ADJ

(2) $T_2 \rightarrow \pi(N^{(noun)}BE^{(be)}ADV^{(adv)}$) N x BE x ADV

(3) $T_3 \rightarrow \pi(N^{(noun)}BE^{(be)}NP1^{(np1)}$) N x BE x NP1

(4) $T_4 \rightarrow \pi(N^{(noun)}V^{(verb)})$ N x V

(5) $T_5 \rightarrow \pi(N^{(noun)}V^{(verb)}PP^{(pp)})$ N x V x PP

(6) $T_6 \rightarrow \pi(N^{(noun)}V^{(verb)}NP1^{(np1)})$ N x V x NP1

(7) $T_7 \rightarrow \pi(N^{(noun)}V^{(verb)}NP1^{(np1)}NP2^{(np2)})$ N x V x NP1 x NP2

(8) $T_8 \rightarrow \pi(NP1^{(np1)}ADJ^{(adj)})$ NP1 x ADJ

(9) $T_9 \rightarrow \pi(NP1^{(np1)}BE^{(be)}ADV^{(adv)})$ NP1 x BE x ADV

(10) $T_{10} \rightarrow \pi(NP1^{(np1)}BE^{(be)}NP2^{(np2)}$) NP1 x BE x NP2

(11) $T_{11} \rightarrow \pi(NP1^{(np1)}V^{(verb)})$ NP1 x V

(12) $T_{12} \rightarrow \pi(NP1^{(np1)}V^{(verb)}PP^{(pp)}$) NP1 x V x PP

(13) $T_{13} \rightarrow \pi(NP1^{(np1)}V^{(verb)}NP2^{(np2)})$ NP1 x V x NP2

(14) $T_{14} \rightarrow \pi(NP1^{(np1)}V^{(verb)}NP2^{(np2)}NP3^{(np3)})$ NP1 x V x NP2 x NP3

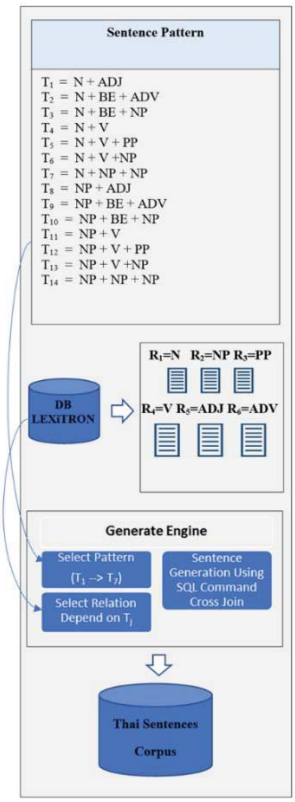
แนวคิดหลักของอัลกอริทึมและการออกแบบเครื่องจักรดังกล่าว แสดงดังรูปต่อไปนี้

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

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

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



ภาพที่ 1 เครื่องจักรที่สร้างเพื่อทดลอง [4]

เมื่อพิจารณาประโยคแบบพึ่งบริบทที่สร้างจากเครื่องจักร [4] ที่พัฒนาขึ้น ด้วยการสุ่มจากประโยคจาก 14 รูปแบบ ครั้งละ 100 ประโยค พบว่า ประโยคแบบพึ่งบริบทที่สามารถเข้าใจตามความหมายอยู่ที่ประมาณร้อยละ 5-60 ของประโยคที่ได้จากการสุ่ม โดยจำแนกประโยคที่เกิดจากรูปแบบของคำ 2, 3 และ 4 คำ สามารถเข้าใจความหมายได้มากสุดร้อยละ 60, 45 และ 5 ดังที่กล่าวไปแล้ว เครื่องจักรและอัลกอริทึมดังกล่าวนี้สามารถสร้างประโยคได้เพียงประโยคชนิดเดียวเท่านั้น และยังเป็นประโยคความเดียวที่เน้นเป็นประโยคบอกเล่าเท่านั้น ยังไม่ตอบสนองต่อประโยคที่แบ่งเจตนาการสื่อสาร ดังนั้นในงานวิจัยนี้จึงสร้างเครื่องจักรใหม่ที่สามารถนำเข้าประโยคทั้งจากในคลังคำ จากเครื่องจักรดังกล่าวมาต่อยอด ซึ่งจะได้นำเสนอในลำดับต่อไป โดยอาศัยนิยามที่กำหนดขึ้นข้างต้นนำไปสู่การสร้างอัลกอริทึมและเครื่องจักรใหม่ที่ได้นำเสนอในงานวิจัยนี้

3. วิธีดำเนินการวิจัย

เบื้องต้นเพื่อให้เข้าใจในการสร้างเครื่องจักรและอัลกอริทึมที่ทำงานในเครื่องจักรดังจะกล่าวต่อไป และรวมไปถึงการเข้าใจตัวอย่างและผลการทดลอง คณะผู้วิจัยกำหนดนิยามเพิ่มเติมดังนี้

3.1 นิยามสำหรับสร้างเครื่องจักร

นิยามที่ใช้งานดังที่กล่าวมาแล้ว นิยาม 1-6 จากงานวิจัย [4] ได้ถูกนำมาใช้ต่อเนื่องในงานวิจัยนี้ด้วย สำหรับนิยามที่เพิ่มเติม ดังนี้

นิยามที่ 7 กำหนดให้ nS[1...n] แทนรูปแบบประโยคภาษาไทยที่ซับซ้อนจากรูปแบบที่ 1 ถึง รูปแบบที่ n

นิยามที่ 8 กำหนดให้ complexPattern1...n เป็นรูปแบบประโยคภาษาไทยที่ซับซ้อนแบบใดๆ ที่สามารถบรรจุใน nS[1...n] ได้

ตัวอย่างที่ 1 แสดงตัวอย่าง complexPattern1...n ซึ่งเป็นรูปแบบประโยคซับซ้อน ยกตัวอย่างรูปแบบคำถาม เช่น
รูปแบบที่ 1 นกกินหนอน**ใช่ไหม**? (S+ใช่ไหม---จากตาราง YesNo)
รูปแบบประโยคคำถามแบบที่ 2 **ใคร**ทำน้ำหก? (Prn+VP ---โดยที่ Prn จากตาราง Ans) หรือ
ซึ่งแสดงว่าเป็น complexPattern1...2 เป็นต้น

นิยามที่ 9 กำหนดให้ Crs_i แทนตารางผลของการสร้างประโยคภาษาไทย ตามรูปแบบ ***nS[i]*** ที่เกิดจากการครอสโปรดักของตาราง rs_j ใดๆ กับ ตารางที่นำมาใช้ครอสโปรดักตามความรูปแบบ ***nS[i]***

ตัวอย่างที่ 2 ตามนิยามดังกล่าวนี้ ตารางที่นำมาครอสโปรดักได้แก่ YesNo, Ans, DenyWord และ command/ask เช่น rs_j จากรูปแบบ (6) $T_6 \rightarrow \pi(N^{(noun)}V^{(verb)}NP1^{(np1)})$ N x V x NP1 จำนวนประโยค 2,897,349,683,400 ขณะที่ ตาราง YesNo มีจำนวน 4 คำ ผลการหา $crs_i = rs_j$ x YesNo (2,897,349,683,400x4) เป็นต้น

3.2 อัลกอริทึม

แนวคิดหลักของการออกแบบเครื่องจักร แบ่งเป็นส่วนนำเข้า (input) ส่วนประมวลผล (processing) และส่วนนำผลออก (result output) สามารถเขียนแทนในแนวทางของอัลกอริทึมได้ดังนี้

Algorithm : CrossProductGeneratorEngine

Input: Relation of Senetences rs_j , typeofSentece(0,...4), complexPattern1...n,

Processing:

If typeofSentence = 0 Then

Return rs_j and exit

Else

Do case :

1: convert pattern of rs_j to S+YesNo format form of complexPattern1..n to nS[1...n]

2: convert pattern of rs_j to Ans+VP format form of complexPattern1..n to nS[1...n]

3: convert pattern of rs_j to S+DenyWord+... format form of complexPattern1..n to nS[1...n]

4: convert pattern of rs_j to command/ask+VP format form of complexPattern1..n to nS[1...n]

End of Do case

For i=1 to n Do

Crs_i = rs_j x nS[i]----where each nS[i] depended on YesNo, Ans, DenWord, command/ask Relation

End of For

End of If

Return: $crs_1, crs_2, crs_3, \dots, crs_n$

จากแนวคิดของเครื่องจักรดังกล่าว แสดงการสร้างประโยคภาษาไทยได้ดังตัวอย่างต่อไปนี้ เมื่อ

ตัวอย่างที่ 3 แสดงการนำเข้าครอสโปรดักเพื่อสร้างประโยคซับซ้อน (6) $T_6 \rightarrow \pi(N^{(noun)}V^{(verb)}NP1^{(np1)})$ N x V x NP1 (ตัวอย่างประโยค เช่น นกกินหนอนใบไม้) มีจำนวนประโยคที่ได้จากการทดลองใน [4] จำนวน 2,897,349,683,400 ประโยค และผ่านค่าประโยคซับซ้อนที่มีรูปแบบเป็นคำถามที่มีค่าเบื้องต้น 2 แบบ ได้แก่ วางไว้หน้าประโยค และต้องการตอบเพียงใช่และไม่ใช่ อาศัยตารางจาก YesNo และ Ans เป็นคำหรือวลีสั้นๆ ที่นำเข้ามาครอสโปรดักตามแนวคิดของเครื่องจักรนี้

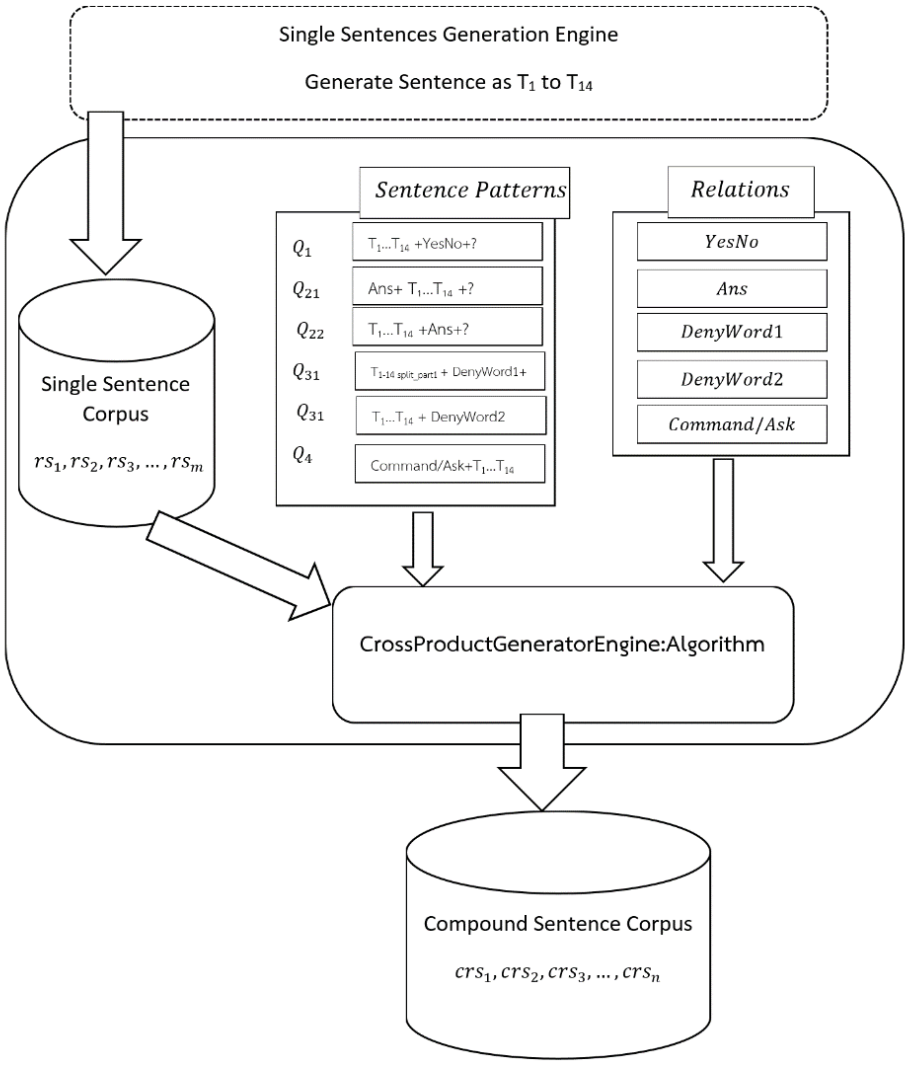
ส่วนนำเข้า: ได้แก่ rs_j จากรูปแบบ (6) $T_6 \rightarrow \pi(N^{(noun)}V^{(verb)}NP1^{(np1)})$ N x V x NP1 จำนวน 2,897,349,683,400 ประโยค typeofSentece(1, 2) ---(รูปแบบ S+คำลงท้าย และ คำถาม+S) complexPattern1...2 ขณะที่ ตาราง YesNo มีจำนวน 4 คำ และตาราง Ans มี 5 คำ

ส่วนการประมวลผล: ทำการ Cross join ด้วย $crs_1 = rs_j$ x YesNo (2,897,349,683,400x4) และ crs_2= Ansrs_j (5x2,897,349,683,400)

ส่วนนำผลออก: crs_1 ด้วยปริมาณเรคอร์ดประโยคภาษาไทยจำนวน (2,897,349,683,400x4) เรคอร์ด (ประโยค) และ crs_2 ด้วยปริมาณเรคอร์ดประโยคภาษาไทยจำนวน (5x2,897,349,683,400) เรคอร์ด (ประโยค) ซึ่งการนำเข้าข้อมูลสู่เครื่องจักรครั้งนี้สามารถผลิตประโยคภาษาไทยที่เป็นคำถามได้จำนวน (2,897,349,683,400x4) + (5x2,897,349,683,400) เป็นต้น

3.3 การสร้างเครื่องจักรเพื่อทดลอง

การสร้างเครื่องจักรสร้างประโยคภาษาไทยให้กับคลังประโยคภาษาไทย เริ่มจากศึกษารูปแบบและโครงสร้างประโยคภาษาไทยตามวัตถุประสงค์การสื่อสาร 8 รูปแบบ จากนั้นแยกคำ/กลุ่มคำในฐานข้อมูลออกเป็นตารางตามประเภทและรูปแบบของประโยค หลังจากนั้นออกแบบเครื่องจักรตามแนวคิดได้ดังภาพที่ 1



ภาพที่ 2 เครื่องจักรที่สร้างเพื่อทดลอง

3.4 รูปแบบสำหรับนำไปสู่การทดลอง/กลุ่มเป้าหมาย

เนื่องจากรูปแบบประโยคภาษาไทยที่ได้จาก [10] แจงได้ 14 รูปแบบดังแสดงในภาพที่ 1 ใน [4] คณะผู้วิจัยจึงแบ่งรูปแบบการทดลองออกเป็นสองกลุ่ม กลุ่มแรกคือรูปแบบที่ 1-7 และกลุ่มที่สอง คือรูปแบบที่ 8-14

รูปแบบประโยคที่ซับซ้อนที่ใช้ออกแบบในงานวิจัยใหม่นี้มีจำนวน 4 รูปแบบดังนี้

1) ประโยคบอกเล่ามีทั้งหมด 14 รูปแบบ (complexPattern1...14) ดังที่กล่าวข้างต้น
2) ประโยคคำถามมี 2 รูปแบบคือ 1) ถามให้ตอบใช่หรือไม่ใช่ มีข้อมูลเบื้องต้นจำนวน 4 คำ (ใช่หรือไม่, ใช่ไหม, หรือไม่, หรือยัง) และ 2) ถามให้ตอบเป็นการอธิบาย มีข้อมูลเบื้องต้น (ใคร, อะไร, ที่ไหน, เมื่อไร, อย่างไร) จำนวน 5 คำ ซึ่งแบ่งย่อยออกเป็นการวางไว้ด้านหน้า และด้านหลังประโยค

ตาราง YesNo(word) -----(Primary Data: ใช่หรือไม่, ใช่ไหม, หรือไม่, หรือยัง)

ตาราง Ans(word) -----(Primary Data: ใคร, อะไร, ที่ไหน, เมื่อไร, อย่างไร)

รูปแบบที่นำไปสู่การทดลอง : 1) $T_1...T_{14}$ +YesNo+?

2) Ans+ $T_1...T_{14}$ +?

3) $T_1...T_{14}$ +Ans+?

3) ประโยคปฏิเสธมี 2 รูปแบบ คือ 1) ใช้คำแทรกระหว่างคำนามกับคำกริยาในประโยค จำนวน 2 คำ (ไม่, มิ) และ 2) ใช้ปฏิเสธต่อท้ายประโยค จำนวน 4 คำ (หามิได้, ก็หาไม่, หรือก็เปล่า, ซะเมื่อไหร่)

ตาราง DenyWord1(word) -----(Primary Data: ไม่, มิ)

ตาราง DenyWord2(word) -----(Primary Data: หามิได้, ก็หาไม่, หรือก็เปล่า, ซะเมื่อไหร่)

รูปแบบที่นำไปสู่การทดลอง : 1) $T_{1\text{-}14\ split\_part1}$ + DenyWord1+ $T_{1\text{-}14\ split\_part2}$

2) $T_1 \ldots T_{14}$ + DenyWord2

4) ประโยคคำสั่งและขอร้อง มี 1 รูปแบบ ด้วยการนำคำสั่งและขอร้องมาวางข้างหน้าประโยค จำนวนคำ 4 คำ (ห้าม, กรุณา, โปรด, อย่า)

ตาราง Command/Ask(word) -----(Primary Data: ห้าม, กรุณา, โปรด, อย่า)

รูปแบบที่นำไปสู่การทดลอง : 1) Command/Ask+$T_1 \ldots T_{14}$

4. ผลการวิจัย

ผลการทดลองเชิงทฤษฎีการสร้างประโยคแบบไม่พึ่งบริบท ได้นำจำนวนประโยคจาก $rs_1, rs_2, rs_3, \ldots, rs_{14} \in RS$ เมื่อ rs_j คือตารางรูปแบบของประโยคที่สร้างขึ้นมาในงานวิจัย [ภ] ซึ่งหากนำประโยคจากการทดลองที่ผ่านมา นำเข้าเครื่องจักรสร้างประโยคนี้จะได้ค่าเท่ากันเพราะถือว่าเป็นประโยคบอกเล่า ซึ่งได้จำนวนดังนี้

ตารางที่ 1 ประโยคจากโครงสร้างประโยคแบบกลุ่มที่ 1 (T_1-T_7)

ลำดับที่	โครงสร้างประโยค	จำนวนการรวมคำ				จำนวนประโยค
		1	2	3	4	
1.	N + ADJ	21,090 (N)	2,879 (ADJ)	-	-	60,718,110
2.	N + BE + ADV	21,090 (N)	4 (BE)	2,876 (ADV)	-	242,619,360
3.	N + BE + NP	21,090 (N)	4 (BE)	10,545 (NP)	-	889,576,200
4.	N + V	21,090 (N)	13,028 (V)	-	-	274,760,520
5.	N + V + PP	21,090 (N)	13,028 (V)	221 (PP)	-	60,722,074,920
6.	N + V + NP	21,090 (N)	13,028 (V)	10,545 (NP)	-	2,897,349,683,400
7.	N + V + NP + NP	21,090 (N)	13,028 (V)	10,545 (NP)	10,545 (NP)	30,552,552,411,453,000

ตารางที่ 2 ประโยคจากโครงสร้างประโยคแบบกลุ่มที่ 2 (T_8-T_{14})

ลำดับที่	โครงสร้างประโยค	จำนวนการรวมคำ				จำนวนประโยค
		1	2	3	4	
1.	NP + ADJ	10,545 (NP)	2,879 (ADJ)	-	-	30,359,055
2.	NP + BE + ADV	10,545 (NP)	4 (BE)	2,876 (ADV)	-	121,309,680
3.	NP + BE + NP	10,545 (NP)	4 (BE)	10,545 (NP)	-	444,788,100
4.	NP + V	10,545 (NP)	13,028 (V)	-	-	137,380,260
5.	NP + V + PP	10,545 (NP)	13,028 (V)	221 (PP)	-	30,361,037,460
6.	NP + V + NP	10,545 (NP)	13,028 (V)	10,545 (NP)	-	1,448,674,841,700
7.	NP + V + NP + NP	10,545 (NP)	13,028 (V)	10,545 (NP)	10,545 (NP)	15,276,276,205,726,500

รูปแบบประโยคที่ซับซ้อนประโยคคำถามจะได้จำนวนประโยคดังนี้

ประโยคคำถามมี 2 รูปแบบคือ 1) ถามให้ตอบใช่หรือไม่ใช่ มีข้อมูลเบื้องต้นจำนวน 4 คำ(ใช่หรือไม่, ใช่ไหม, หรือไม่, หรือยัง Q1 ---รูปแบบ $T_1 \ldots T_{14}$ +YesNo+? และ 2) ถามให้ตอบเป็นการอธิบาย มีข้อมูลเบื้องต้น (ใคร, อะไร, ที่ไหน, เมื่อไร, อย่างไร) จำนวน 5 คำ ซึ่งแบ่งย่อยออกเป็นการวางไว้ด้านหน้า และด้านหลังประโยค (2 แบบ กำหนด Q21, Q22) ตามรูปแบบ---- Ans+ $T_1 \ldots T_{14}$ +? และ $T_1 \ldots T_{14}$ +Ans+?

ตารางที่ 3 จำนวนประโยคคำถามที่ได้

รูปแบบประโยคเดี่ยว (T1-T14)	จำนวนประโยคเดี่ยว (rsj)	ประโยคที่ได้จากการสร้างด้วย Q1(rsjx4)	ประโยคที่ได้จากการสร้างด้วย Q21(rsjx5)	ประโยคที่ได้จากการสร้างด้วย Q22(rsjx5)
T_1	60,718,110	242,872,440	303,590,550	303,590,550
T_2	242,619,360	970,477,440	1,213,096,800	1,213,096,800
T_3	889,576,200	3,558,304,800	4,447,881,000	4,447,881,000
T_4	274,760,520	1,099,042,080	1,373,802,600	1,373,802,600
T_5	60,722,074,920	242,888,299,680	303,610,374,600	30,3610,374,600
T_6	2,897,349,683,400	11,589,398,733,600	14,486,748,417,000	14,486,748,417,000
T_7	30,522,522,411,453,000	122,090,089,645,812,000	152,612,612,057,265,000	152,612,612,057,265,000
T_8	30,359,055	121,436,220	151,795,275	151,795,275
T_9	121,309,680	485,238,720	606,548,400	606,548,400
T_{10}	444,788,100	1,779,152,400	2,223,940,500	2,223,940,500
T_{11}	137,380,260	549,521,040	686,901,300	686,901,300
T_{12}	30,361,037,460	121,444,149,840	151,805,187,300	151,805,187,300
T_{13}	1,448,674,841,700	5,794,699,366,800	7,243,374,208,500	7,243,374,208,500
T_{14}	15,276,276,205,726,500	61,105,104,822,906,000	76,381,381,028,632,500	76,381,381,028,632,500

ประโยคปฏิเสธมี 2 รูปแบบ คือ 1) D1 ใช้คำแทรกระหว่างคำนามกับคำกริยาในประโยค จำนวน 2 คำ (ไม่, มิ) และ 2) D2 ใช้ปฏิเสธต่อท้ายประโยค จำนวน 4 คำ (หามิได้, ก็หาไม่, หรือก็เปล่า, ซะเมื่อไหร่)--- 1) $T_{1\text{-}14\ split\_part1}$ + DenyWord1+ $T_{1\text{-}14\ split\_part2}$ 2) $T_1 \ldots T_{14}$ + DenyWord2

ตารางที่ 4 จำนวนประโยคปฏิเสธที่ได้

รูปแบบประโยคเดี่ยว (T1-T14)	จำนวนประโยคเดี่ยว (rsj)	ประโยคที่ได้จากการสร้างด้วย D1(rsjx2)	ประโยคที่ได้จากการสร้างด้วย D2(rsjx4)
T_1	60,718,110	121,436,220	242,872,440
T_2	242,619,360	485,238,720	970,477,440
T_3	889,576,200	1,779,152,400	3,558,304,800
T_4	274,760,520	549,521,040	1,099,042,080
T_5	60,722,074,920	121,444,149,840	242,888,299,680
T_6	2,897,349,683,400	5,794,699,366,800	11,589,398,733,600
T_7	30,522,522,411,453,000	61,045,044,822,906,000	122,090,089,645,812,000
T_8	30,359,055	60,718,110	121,436,220
T_9	121,309,680	242,619,360	485,238,720
T_{10}	444,788,100	889,576,200	1,779,152,400
T_{11}	137,380,260	274,760,520	549,521,040
T_{12}	30,361,037,460	60,722,074,920	121,444,149,840
T_{13}	1,448,674,841,700	2,897,349,683,400	5,794,699,366,800
T_{14}	15,276,276,205,726,500	30,552,552,411,453,000	61,105,104,822,906,000

ประโยคคำสั่งและขอร้อง มี 1 รูปแบบ (C1) ด้วยการนำคำสั่งและขอร้องมาวางข้างหน้าประโยค จำนวนคำ 4 คำ (ห้าม, กรุณา, โปรด, อย่า)--- รูปแบบ Command/Ask+$T_1 \ldots T_{14}$

ตารางที่ 5 จำนวนประโยคคำสั่งและขอร้องที่ได้

รูปแบบ (T1-T14)	จำนวนประโยคเดี่ยว (rsj)	ประโยคที่ได้จากการสร้างด้วย C1(rsjx4)
T_1	60,718,110	242,872,440
T_2	242,619,360	970,477,440
T_3	889,576,200	3,558,304,800
T_4	274,760,520	1,099,042,080
T_5	60,722,074,920	242,888,299,680
T_6	2,897,349,683,400	11,589,398,733,600
T_7	30,522,522,411,453,000	122,090,089,645,812,000
T_8	30,359,055	121,436,220
T_9	121,309,680	485,238,720
T_{10}	444,788,100	1,779,152,400
T_{11}	137,380,260	549,521,040
T_{12}	30,361,037,460	121,444,149,840
T_{13}	1,448,674,841,700	5,794,699,366,800
T_{14}	15,276,276,205,726,500	61,105,104,822,906,000

อย่างไรก็ตาม ผลการวิจัยดังกล่าวเป็นเชิงทฤษฎีที่สามารถสร้างการครอสโปรดักเพื่อสร้างประโยคได้จำนวนมหาศาล อีกทั้งตามอัลกอริทึมภายในเครื่องจักรที่สร้างขึ้นมิได้จำกัดสำหรับเพียงรูปแบบประโยคเริ่มต้นเพียง 14 รูปแบบ และรูปแบบตามวัตถุประสงค์เพียงแค่ 8 รูปแบบ และประโยคซับซ้อนที่ยกขึ้นมาแสดงในผลเชิงทฤษฎีนี้เท่านั้น หากวิเคราะห์รูปแบบประโยคภาษาไทยยังมีอีกจำนวนมาก ซึ่งสามารถนำเข้าและสร้างประโยคได้จากเครื่องจักรดังกล่าวนี้ได้เป็น จำนวนมากด้วยเช่นกัน

5. สรุปผลและข้อเสนอแนะ

งานวิจัยนี้พัฒนาเครื่องจักรทางซอฟต์แวร์ใหม่สำหรับสร้างประโยคภาษาไทยตามวัตถุประสงค์การสื่อสารอัตโนมัติเพื่อเก็บไว้ในคลังประโยคตามแนวทางการประมวลผลภาษาธรรมชาติ มุ่งเน้นศึกษาการออกแบบอัลกอริทึมและกำหนดรูปแบบประโยคตามไวยากรณ์ภาษาไทย อาศัยกลไกของฐานข้อมูลเชิงสัมพันธ์ควบคุมการสร้างประโยคด้วยฟังก์ชันทางพีชคณิตเชิงสัมพันธ์เป็นกฎในการสร้างประโยค เครื่องจักรซอฟต์แวร์ที่สร้างขึ้นสามารถสร้างประโยคภาษาไทยแบบไม่พึ่งบริบทด้วยการครอสโปรดักรีเลชัน ทำให้ได้จำนวนประโยคเป็นค่าผลคูณคาร์ทีเชียนของจำนวนเรคคอร์ดในตารางที่ถูกเลือกมาสร้างประโยค ทั้งสามารถใช้ต่อยอดในงานประมวลผลภาษาธรรมชาติโดยเฉพาะสตอรีเจนเนอเรชันหรือสกัดเอาประโยคที่ดี นำเข้าสู่การเรียนรู้เครื่องจักรและตัวแบบภาษาธรรมชาติได้ นอกจากนั้นยังสามารถนำไปปรับใช้กับภาษาอังกฤษและภาษาอื่นๆ ได้อีกด้วย

ข้อเสนอแนะสำหรับการวิจัยในครั้งต่อไป ควรมีรูปแบบและโมเดลในการสกัดเอาประโยคที่ใช้งานได้ดี หรือกำหนดรูปแบบการสร้างประโยคเพิ่มเติมตามแนวทางของมาร์คอฟเชน (Markov Chain) ที่ระบุค่าความน่าจะเป็นร่วมในการสร้างประโยค ซึ่งน่าจะช่วยให้การสร้างประโยคมีความแม่นยำและได้ประโยคแบบพึ่งบริบทได้อย่างถูกต้องมากยิ่งขึ้น

เอกสารอ้างอิง

[1] Limpanadudadee W., Punyabukkanna P. and 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)** (80-85).

[2] Krukaset W., Krukaset N. and Khancome C. (2017). Thai Sentence Generation Machine Employing Fixed Patterns. **2017 IEEE International Conference on High Performance Computing and Communications Workshops** (70-73).

[3] เชาวลิต ขันคำและ สุภัทรา สหพงศ์. (2563). ภาษาโครงสร้างสำหรับจัดเก็บคำไทยเชิงความหมาย (Data Structure Language for Storing Thai Semantic Word). **การประชุมวิชาการวิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยีและสถาปัตยกรรรมศาสตร์ ครั งที่ 11** (1417-1423).

[4] สุพัตรา แดงเจริญ, ขณิดา จรุงจิตต์, และ เชาวลิต ขันคำ. (2563). เครื่องจักรสร้างประโยคภาษาไทยสำหรับคลังประโยคภาษาไทย. **การประชุมวิชาการนวัตกรรมด้านวิศวกรรมและเทคโนโลยีเพื่อเศรษฐกิจและสังคม ครั้งที่ 3 (Proceedings of the 3rd Conference on Innovation Engineering and Technology for Economy and Society 2020)** (199-204).

[5] **LEXiTRON Data : ข้อมูลเล็กซิตรอนเป็นฐานข้อมูลที่เหมาะสำหรับผู้ที่จะนำไปพัฒนาหรือศึกษาต่อ**. ค้นเมื่อ 11 พฤศจิกายน 2561 จากhttps://lexitron.nectec.or.th/2009_1/index.php?q=common_manager/download#latest_version

[6] เกียรติชัย เดชพิทักษ์ศิริกุล. (2551). **การศึกษาเปรียบเทียบวลี ประโยค และสัมพันธสารของเด็กปกติและเด็กทิสติก**. วิทยานิพนธ์ปริญญามหาบัณฑิตอักษรศาสตร์ สาขาภาษาไทย บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร.

[7] ณรงค์กรรณ รอดทรัพย์.(2555). โครงสร้างและวากยสัมพันธ์ของนามวลีภาษาไทย ในป้ายรณรงค์หาเสียงเลือกตั้งสมาชิกสภาผู้แทนราษฎร พ.ศ. 2555 กรณีศึกษาเขตพื้นที่กรุงเทพมหานครและพิษณุโลก. **วารสารปาริชาต ฉบับพิเศษผลงานวิจัยจากการประชุมวิชาการ ครั้งที่ 12** (64-74).

[8] **ชนิดและโครงสร้างของวลีและประโยค. ค้นเมื่อ** 25 พฤศจิกายน 2561. จาก http://kb.psu.ac.th/psukb/bitstream/2010/6891/8/Chapter3.pdf

[9] Elmasri R. and Navathe S. (2010). **Fundamentals of Database System Fourth Edition**. USA: Pearson Education Inc.

[10] ศศลักษณ์ ทองขาว. (2550). **ปัญญาประดิษฐ์**. สงขลา: คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา.