ระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทย

โดย

นางสาวณิชชา บำรุง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2558

ระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทย

โดย

นางสาวณิชชา บำรุง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2558

An Automatic Compound Noun Extraction System for Thai Sentences

BY

Ms. Nidcha Bumrung

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCE AND TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2015

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

นางสาวณิชชา บำรุง

เรื่อง

ระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทย

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

เมื่อวันที่ วันที่ 7 มกราคม พ.ศ. 2559

ประธานกรรมการสอบวิทยานิพนธ์ _____
(ดร.ปกป้อง ส่องเมือง)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.รัชฎา คงคะจันทร์)

กรรมการสอบวิทยานิพนธ์ _____
(ดร.วสิศ ลิ้มประเสริฐ)

กรรมการสอบวิทยานิพนธ์ _____
(ดร.มารุต บูรณรัช)

คณบดี _____
(รองศาสตราจารย์ปกรณ์ เสริมสุข)

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

นางสาว ณิชชา บำรุง

เรื่อง

ระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทย

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต

เมื่อวันที่ วันที่ 4 มกราคม พ.ศ. 2559

ประธานกรรมการสอบวิทยานิพนธ์ _____
(ดร.ปกป้อง ส่องเมือง)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก _____
(ผู้ช่วยศาสตราจารย์ ดร.รัชฎา คงคะจันทร์)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม _____
(ดร.วสิศ ลิ้มประเสริฐ)

กรรมการสอบวิทยานิพนธ์ _____
(ดร.มารุต บูรณรัช)

คณบดี _____
(รองศาสตราจารย์ปกรณ์ เสริมสุข)

หัวข้อวิทยานิพนธ์	ระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทย
ชื่อผู้เขียน	นางสาว ณิชชา บำรุง
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา/คณะ/มหาวิทยาลัย	วิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร.รัชฎา คงคะจันทร์
ปีการศึกษา	2558

บทคัดย่อ

วิทยานิพนธ์ฉบับนี้นำเสนอระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทยโดยระบบประกอบด้วยสี่ส่วน คือ การประมวลผลก่อน การวิเคราะห์คำนามประสมโดยใช้รูปแบบ การวิเคราะห์คำนามประสมถาวร และ การวิเคราะห์คำนามประสมโดยใช้ความหมาย โดยการหาหน้าที่ของคำของประโยคแต่ละประโยคจะใช้โปรแกรม swath เป็นตัวติดป้ายแบ่งแยกโครงสร้างของคำแต่ละคำซึ่งหน้าที่ของคำแต่ละคำและจะมีการพิจารณาคำแต่คำด้วยกฎขอไวยากรณ์ซึ่งเป็นกฎสำหรับคำนามประสมที่ได้มาจากผู้เชี่ยวชาญโดยคำที่ตรงกันตามกฎจะถือว่าเป็นคำนามประสมจากนั้นจึงจะเข้าสู่การขั้นตอนแรกซึ่งเป็นการตรวจสอบคำนามประสมโดยนำคำที่ได้จากกฎจากข้างต้นมาตรวจสอบกับพจนานุกรม กรณีที่พบคำนามประสมในพจนานุกรมจะนับว่าเป็น “permanent compound noun” ส่วนกรณีที่ไม่พบคำนามประสมในพจนานุกรมต้องเข้าสู่ขั้นตอนที่สองการพิจารณาคำนามประสมซึ่งประกอบด้วยสี่ขั้นตอนด้วยกันคือความสัมพันธ์แบบจ่ากลุ่ม-ลูกกลุ่ม คำนามประสมไม่สามารถแยกได้ คำนามประสมสลับที่ไม่ได้ และ ไม่มีความเป็นเจ้าของ กรณีที่พบทั้งสี่ขั้นตอนจะถือว่าเป็นคำนามประสม จากการทดลองคำที่นำมาตรวจสอบจะต้องผ่านส่วนที่ 1, 2, และ 3 หรือ 4 ถึงจะสามารถสรุปได้ว่าคำที่นำมาตรวจสอบเป็นคำนามประสม การทดลองเป็นการใช้บทความจากสารานุกรมของคลังข้อความภาษาไทย BEST Corpus Training Set 1 Release 3 ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ เป็นตัวทดสอบประสิทธิภาพซึ่งจากผลการทดสอบได้ค่าความครบถ้วนเป็น 55.38% และ ค่าความแม่นยำเป็น 41.38%

Thesis Title	An Automatic Compound Noun Extraction System for Thai Sentences
Author	Ms. Nidcha Bumrung
Degree	Master of Science (Computer Science)
Major Field/Faculty/University	Department of Computer Science Faculty of Science and Technology Thammasat University
Thesis Advisor	Dr. Rachada Kongkachandra
Academic Years	2015

ABSTRACT

This Thesis is proposed an automatic system for extracting some compound nouns from Thai sentences. The system includes four parts i.e. preprocessing, pattern-based compound noun analysis, permanent compound noun analysis and semantic-based compound noun analysis. The input sentences are firstly segmented and tagged their part-of-speeches by using SWATH. Then, the POS tags of each word are considered with syntactical rules for compound noun provided by the experts. The matched words according to the rules are counted as compound noun candidates. The first feature is to consider their appearances in the standard dictionary. If they are found in the dictionary, they then counted as the "permanent compound noun". All candidates that are not found in the dictionary are then re-considered by their semantics. The four features as the superordinate-subordinate criterion, inseparable compound noun, unchangeable position compound noun and no ownership are employed. The candidates are considered as the compound nouns when the feature 1,2, and (3 or 4) are true. The five million word from BEST2010 encyclopedia are exploited to evaluate the system performance. The precision and recall of the extraction system are 41.37% and 55.38%.

Keywords: Thai Compound nouns analysis, Automatic compound noun extraction-system, permanent compound noun, Semantic-based criteria, Superordinate-subordinate semantic

กิตติกรรมประกาศ

วิทยานิพนธ์เหล่านี้สำเร็จลุล่วงได้ด้วยความเมตตาจากผู้ช่วยศาสตราจารย์ ดร.รัชฎา คงคะจันทร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ตลอดจนคณาจารย์ และคณะกรรมการสอบวิทยานิพนธ์ทุกท่าน ที่สละเวลาอันมีค่าในการให้คำแนะนำ ความรู้ แง่คิด และทักษะต่างๆ ที่เกี่ยวกับการทำงานวิจัย ตลอดจนการตรวจสอบความถูกต้องในการจัดทำวิทยานิพนธ์ฉบับนี้ จนงานวิจัยและวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี รวมถึงขอขอบคุณเจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์ทุกท่าน ที่ให้ความอนุเคราะห์และอำนวยความสะดวกในการทำวิทยานิพนธ์มาโดยตลอด

ขอกราบขอบพระคุณ มารดาที่อยู่ช่วยตลอดตั้งแต่เริ่มจนจบการทดสอบระบบรวมถึงรุ่นพี่ที่ธรรมศาสตร์ยอมสละเวลาคอยให้คำปรึกษาและสอบถามข้อสงสัย ตลอดจนเพื่อนๆพี่ๆและน้องๆที่ช่วยทำแบบสอบถามในการทดสอบระบบ ขอขอบคุณ วาสนา ที่ช่วยในการให้คำปรึกษาและพัฒนาระบบ ขอขอบคุณคุณพรทิวาและคุณนิลวนาที่ช่วยเป็นที่ปรึกษาในเรื่องภาษาอังกฤษ ขอขอบคุณจุฑามาศและคุณนิศาชลที่สละเวลามาเป็นที่ปรึกษาและเป็นผู้เชี่ยวชาญ สุดท้ายนี้ขอขอบคุณเพื่อนๆที่คอยให้กำลังใจและช่วยเหลือจนวิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ขึ้นได้ ขออันเชิญคุณพระศรีรัตนตรัยและสิ่งศักดิ์สิทธิ์ทั้งหลาย คุ้มครองปกป้องรักษาให้ทุกท่าน มีความสุข มีความเจริญ มีโภคทรัพย์และมีสุขภาพพลานามัยสมบูรณ์แข็งแรงตลอดไปด้วยเทอญ

นางสาว ณิชชา บำรุง

สารบัญ

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การตรวจสอบคำนามประสมจากประโยคภาษาไทยมีวัตถุประสงค์เพื่อทำให้คอมพิว เตอร์ สามารถรับรู้และสามารถตรวจสอบหาคำนามประสมที่อยู่ภายในประโยคก่อนที่คอมพิวเตอร์จะนำข้อมูลไปประมวลผลในส่วนงานอื่นๆ โดยข้อมูลที่ใช้ในการศึกษาเป็นบทความจากสารานุกรมของคลังข้อความภาษาไทย BEST Corpus Training Set 1 Release 3 ของ NECTEC

การศึกษาเกี่ยวกับคำในภาษาไทยเรื่องของการนำคำต่างๆ มาประกอบกันเป็นกลุ่มคำเพื่อใช้ในการสื่อสาร นักไวยากรณ์ไทยแต่เดิมมา เรียกกลุ่มคำที่ประกอบกันบางแบบว่าประโยค (พระยาอุปกิตศิลปสาร ๒๕๔๘) และ เรียกกลุ่มคำที่ประกอบกันแบบอื่นๆ ว่า วลี การตั้งชื่อวลีตั้งตามชนิดของคำที่อยู่ต้นวลี อาทิ วลีที่มีคำนามอยู่ต้นวลี จะเรียกว่า นามวลี ส่วนวลีที่มีคำกริยาอยู่ต้นวลี จะเรียก กริยาวลี ต่อมามีการศึกษาเกี่ยวกับการนำหน่วยคำอิสระตั้งแต่ 2 หน่วยขึ้นไปมาประสมกันทำให้เกิดเป็นคำใหม่ที่มีความหมายใหม่หนึ่งหน่วยความหมาย โดยจะเรียกคำนั้นว่า คำประสม ซึ่งคำประสมจะมีความแตกต่างกับวลีในเรื่องของความหมาย กล่าวคือ วลีจะเกิดจากคำหรือหน่วยคำอิสระตั้งแต่ 2 คำขึ้นไปมาเรียงต่อกันและไม่เกิดความหมายใหม่ ส่วนคำประสมเมื่อนำหน่วยคำอิสระตั้งแต่ 2 คำมาเรียงต่อกันจะต้องเกิดเป็นความหมายใหม่หนึ่งหน่วยความหมาย ซึ่งในงานวิจัยฉบับนี้จะเน้นมุ่งไปที่การศึกษาในส่วนของคำนามประสม

รูปแบบของคำนามประสมทำให้สามารถสะท้อนให้เห็นถึงลักษณะทางสังคมและวัฒนธรรมได้ ดังที่ คุณอัญชลี สิงห์น้อย ได้นำเสนอคือ คำนามประสมเป็นรูปแบบหนึ่งของภาษาที่สามารถแสดงหรือสะท้อนให้เห็นถึงพื้นฐานทางสังคม และวัฒนธรรมของผู้พูดภาษานั้นๆ ได้อย่างชัดเจน ผู้ที่พูดภาษาต่างกันและมีวัฒนธรรมที่ต่างกัน ย่อมมีพื้นฐานทางความคิดที่แตกต่างกันในการสร้างคำประสมเพื่อใช้เรียกสิ่งของ คุณอัญชลี สิงห์น้อยได้ให้ตัวอย่างเช่น คำว่า แม่บ้าน ในภาษาไทยเกิดจากการนำเอาคำว่า แม่ และ บ้าน มาประสมกัน เพื่อใช้เรียกภรรยาของพ่อบ้าน หญิงผู้จัดการงานในบ้าน ซึ่งขณะเดียวกัน ในภาษาอังกฤษก็มีการประสมคำที่ใช้ในความหมายเดียวกันนี้ แต่เลือกคำที่แตกต่างออกไปมาประสมกัน คือ คำว่า wife แทนที่จะเป็นคำว่า mother เหมือนดังเช่นในภาษาไทย มาประสมกับคำว่า house เป็น housewife มาใช้เรียกบุคคลดังกล่าว

สำหรับคำนามประสมในภาษาไทยอาจประกอบด้วยคำ 2 คำหรือมากกว่านั้น โดยขึ้นอยู่กับวิวัฒนาการในการใช้และการยอมรับกันเป็นสากลไม่ได้มีกฎเกณฑ์ใดตายตัว ซึ่งสามารถตรวจสอบการใช้ที่ถูกต้องได้จากพจนานุกรมที่มีการปรับปรุงล่าสุด จากปัญหาข้างต้นทำให้คอมพิวเตอร์ไม่สามารถตรวจสอบได้ว่า ภายในประโยค 1 ประโยค มีคำนามประสมอยู่ร่วมด้วย ซึ่งจะส่งผลกระทบต่อการนำข้อมูลไปประมวลผลในลำดับถัดไป

ดังนั้น ผู้ศึกษาจึงเกิดแนวความคิดในการแก้ปัญหา โดยนำเทคนิควิทยาการคอมพิวเตอร์ซึ่งเป็นการนำโครงสร้างของคำนามประสมมาใช้ในการตรวจสอบ เพื่อคัดแยกคำนามประสมออกจากประโยคด้วยวิธีทางอัลกอริทึม จากนั้นจึงนำไปตรวจสอบ เพื่อคัดแยกคำนามประสมออกจากประโยคด้วยวิธีการ 2 ขั้นตอนคือ การตรวจสอบคำนามประสม และ การพิจารณาคำนามประสม

1.2 วัตถุประสงค์งานวิจัย

1. เพื่อศึกษาหลักการในการคัดแยกคำนามประสมออกจากประโยค
2. เพื่อศึกษาหลักการในการตรวจสอบคำนามประสมภายในประโยคของภาษาไทย
3. เพื่อนำเสนอวิธีการแก้ไขปัญหาการคัดแยกคำนามประสมออกจากประโยคในภาษาไทย ได้แก่ การจับกลุ่มจากโครงสร้าง Part of Speech ของคำนามประสม, การตรวจสอบคำนามประสม และ การพิจารณาคำนามประสม

1.3 ขอบเขตงานวิจัย

1. ประโยคที่นำมาใช้ในการทดสอบการคัดแยกคำนามประสม เป็นบทความที่มีประโยคไม่ยาวมากนัก และมีการเว้นวรรคของแต่ละประโยคอย่างชัดเจน
2. ข้อมูลที่ใช้ในการทดสอบ จะใช้ swath ในการหาส่วนประกอบของประโยค แต่ไม่สามารถตรวจสอบคำที่ swath ค้นหาส่วนประกอบออกมาผิดได้
3. คำแต่ละคำภายประโยคจะต้องมีความถูกต้องตามหลักไวยากรณ์
4. ข้อมูลที่ใช้ในการทดสอบ เป็นบทความจากสารานุกรมของคลังข้อความภาษาไทย BEST Corpus Training Set 1 Release 3 ของ NECTEC

1.4 คำจำกัดความ

1.4.1 การจับคู่คำจากความสัมพันธ์ของคำ

เป็นการแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น โดยอ้างอิงจากความสัมพันธ์ของคำ ด้วยวิธีการจับคู่ความสัมพันธ์ของคำ

1.4.2 Compounds noun

คำนามประสม คือ คำนามที่เกิดจากการนำเอาคำตั้งแต่ 2 คำ ขึ้นไป มารวมกันแล้วเกิดเป็นคำนามคำใหม่ขึ้นมา โดยจะมีคำคำหนึ่งที่จะอยู่ข้างท้ายจะทำหน้าที่เป็นคำหลัก (head) และคำที่อยู่ข้างหน้า ทำหน้าที่เป็นส่วนขยาย (modifier) เมื่อรวมกันแล้วอาจเกิดเป็นคำใหม่คำเดียวกัน หรือแยกเป็น 2 คำที่ยังคงคำเดิม หรือใส่ – (hyphen) เชื่อมระหว่างคำ ทั้งนี้ขึ้นอยู่กับวิวัฒนาการในการใช้และการยอมรับกันเป็นสากลไม่ได้มีกฎเกณฑ์ใดตายตัว ซึ่งสามารถตรวจสอบการใช้ที่ถูกต้องได้จากพจนานุกรม (dictionary) ที่มีการปรับปรุงล่าสุด

1.4.3 การแยกคำนามประสม

การแยกคำนามประสมเป็นเทคนิคในการแบ่งแยกคำนามประสมออกจากประโยค โดยใช้โครงสร้างของคำนามประสมในการคำนวณหาวิธีการแบ่งแยกคำนามประสมออกจากประโยค

1.4.4 N-gram

แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดอักขระ (Character Sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงกัน (Word Sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของชุดอักขระหรือคำ ประมาณได้จากคลังข้อมูลที่สร้างไว้ ซึ่ง N-Gram ได้ใช้หลักการของสถิติในหลาย ๆ ด้านมาประยุกต์ใช้

1. การประมาณค่าด้วย 2-Gram (Probability Bigram) คือ การประมาณค่าความน่าจะเป็นของชุดอักขระที่เกิดขึ้นร่วมกันที่จะพบ อักขระ (คำ) ทีละ 2 ตัว (คำ) ติดกันในชุดอักขระนั้น

2. การประมาณค่าด้วย 3- Gram (Probability Trigram) คือ การประมาณค่าความน่าจะเป็นของชุดอักขระที่เกิดขึ้นร่วมกันที่จะพบ อักขระ (คำ) ทีละ 3 ตัว (คำ) ติดกันในชุดอักขระนั้น

3. การประมาณค่าด้วย 4- Gram (Probability Quadgram) คือ การประมาณค่าความน่าจะเป็นของชุดอักขระที่เกิดขึ้นร่วมกันที่จะพบ อักขระ (คำ) ทีละ 4 ตัว (คำ) ติดกันในชุดอักขระนั้น หรืออาจประมาณค่าความน่าจะเป็นจากความยาวของเอ็นแกรมมากกว่า 4-แกรม ก็ได้ขึ้นอยู่กับความจำเป็นในการทดลอง แต่ระบบของเอ็นแกรมจะยิ่งซับซ้อนมากขึ้นตามลำดับ

1.4.5 Lexitron

พจนานุกรมสื่ออิเล็กทรอนิกส์ไทย อังกฤษ สามารถใช้งานได้สองรูปแบบ คือ ระบบออนไลน์ให้บริการผ่านทางเว็บไซต์และระบบออฟไลน์ ผู้ใช้สามารถดาวน์โหลดไปใช้งานบนเครื่องคอมพิวเตอร์ส่วนบุคคลได้ โดยคำศัพท์จะแสดงความหมาย ประเภทของคำ นิยามของคำศัพท์ คำพ้อง คำตรงข้าม คำลักษณะนาม ประโยคตัวอย่างที่มีการใช้งานจริง

1.4.6 Superordinate-Subordinate

X-x = Xx/x ซึ่ง X เป็นคำว่า ปลา ส่วน x เล็ก คือ ดุก เมื่อนำมาผสมกันจะได้ความสัมพันธ์เป็นลักษณะของ superordinate-subordinate คือ ปลา-ดุก

1.4.7 ความถาวร

การจำแนกอรรถสัมพันธ์พื้นฐาน โดยพื้นฐานความเชื่อที่เกี่ยวกับความถาวร และความเป็นปกติวิสัยของสิ่งต่างๆ เช่น ประเทศไทย ประเทศจีน เป็นการใช้เรียกชื่อของประเทศแต่ละประเทศ ซึ่งมีความถาวรหรือเป็นสิ่งทำเป็นปกติวิสัย เป็นคำที่ใช้กันทั่วไปและคนทั่วไปสามารถเข้าใจความหมายได้

1.4.8 Part of Speech

ส่วนที่ประกอบออกมาเป็นคำพูด หรือออกมาเป็นประโยค สามารถแบ่งออกเป็น 8 ชนิด ได้แก่

1. คำนาม (Noun) คือคำที่ใช้เรียกชื่อคน สัตว์ สิ่งของ สถานที่
2. คำสรรพนาม (Pronoun) คือคำที่ใช้แทนคำนาม
3. คำคุณศัพท์ (Adjective) ใช้ขยายนามหรือสรรพนาม
4. คำวิเศษณ์ (Adverb) คือคำที่ใช้แสดงความเคลื่อนไหวของนามหรือสรรพนาม
5. คำกริยา (Verb) คือคำที่ใช้ขยายกริยา คุณศัพท์ และกริยาวิเศษด้วยกัน
6. คำบุพบท (Preposition) คือคำที่ใช้บอกความสัมพันธ์กับนามหรือกริยา
7. คำสันธาน (Conjunction) ใช้เชื่อมประโยค หรือคำ
8. คำอุทาน (Interjection) ใช้เพื่อแสดงความรู้สึก เสียใจ หรือประหลาดใจ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สร้างโปรแกรมที่สามารถตรวจสอบคำนามประสมจากประโยคได้อย่างถูกต้อง
2. ทำให้คอมพิวเตอร์สามารถแยกคำนามประสมออกจากประโยคได้อย่างถูกต้อง
3. สามารถแก้ปัญหาการแยกคำนามประสมออกจากประโยค ด้วยใช้วิธีการทางอัลกอริทึมอย่างถูกต้องได้

1.6 รายละเอียดวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้ประกอบด้วยบทนำในบทที่ 1 จะกล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของงานวิจัย ขอบเขตของงานวิจัย คำจำกัดความ และผลประโยชน์ที่คาดว่าจะได้รับ บทที่ 2 กล่าวถึงทฤษฎีต่างๆ ที่เกี่ยวข้องกับงานวิจัย รวมทั้งงานวิจัยอื่นๆ ที่เกี่ยวข้อง บทที่ 3 เป็นการอธิบายวิธีการดำเนินงานวิจัย แนวคิดพื้นฐานและขั้นตอนการดำเนินงาน บทที่ 4 เป็นส่วนของผลการทดลองและการประเมินผลการทดลอง และบทที่ 5 บทสุดท้าย เป็นส่วนของการสรุปผลการทดลองของงานวิจัย รวมทั้งรายละเอียดและข้อเสนอแนะ

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

ในวิทยานิพนธ์นี้นำเสนอเกี่ยวกับคำนามประสม ซึ่งจัดอยู่ในประเภทหนึ่งของ Multiword Expression งานวิจัยฉบับนี้เป็นการตรวจสอบคำนามประสมจากประโยคภาษาไทย โดยมีวัตถุประสงค์ เพื่อให้คอมพิวเตอร์สามารถรับรู้และสามารถตรวจสอบหาคำนามประสมภายในประโยค ก่อนที่คอมพิวเตอร์จะนำข้อมูลไปประมวลผลในลำดับถัดไป

2.1.1 Multiword Expressions (MWEs)

บางวลีในพจนานุกรม จะประกอบด้วย หนึ่งคำหรือสองคำหรือมากกว่านั้น โดยเฉพาะ phrase (วลี) ยกตัวอย่างเช่น strong tea ไม่ได้หมายถึง ชาที่แข็งแรงหรืออ้วน แม้ว่าคำคุณศัพท์เหล่านั้นอาจจะใช้อธิบายชาที่มีมากกว่าปริมาณปกติของรสชาติ แต่บางครั้งวลีอาจมีความหมายที่แตกต่างจากคำที่นำมาประกอบกัน เช่น วลี kick the Bucket ซึ่งมีความหมายว่า ตาย โดยเรียกวลีข้างต้นว่า MWEs

Multiword Expressions สามารถแบ่งออกเป็นหลายประเภท คือ idioms, compounds noun, lexical collocation, complex lexical items, particle verbs, named entities, figurative expressions, Institutionalized phrases, terminology และ light verb

2.1.1.1 compounds noun

compounds noun หมายถึง การนำคำตั้งแต่ 2 คำ มารวมกัน แล้วเกิดเป็นคำนามใหม่ขึ้นมา โด ยเมื่อรวมกันแล้ว อาจเกิดเป็นคำใหม่ที่มีความหมายต่างจากคำที่นำมารวมกัน หรือยังคงมีเค้าความหมายเดิม ทั้งนี้ขึ้นอยู่กับวิวัฒนาการในการใช้และการยอมรับกันเป็นสากลไม่ได้มีกฎเกณฑ์ใดตายตัว ซึ่งสามารถตรวจสอบการใช้ที่ถูกต้องได้จากพจนานุกรมที่มีการปรับปรุงล่าสุด

2.1.1.2 Idioms

Idioms หมายถึง สำนวนหรือถ้อยคำที่ไม่ได้มีความหมายตรงตัว ซึ่งสามารถแสดงตัวอย่างได้ดังนี้

so far หมายถึง จนถึงทุกวันนี้, จนกระทั่งปัจจุบัน, จนเดี๋ยวนี้

to a limited extent หมายถึง เพียงแค่นี้, เพียงเท่านี้

2.1.1.3 lexical collocation

lexical collocation หมายถึง กลุ่มคำที่ปรากฏร่วมกัน มักประกอบด้วย คำนาม คำกริยา คำคุณศัพท์ หรือ คำกริยาวิเศษณ์

สำหรับผู้เรียนภาษาอังกฤษเป็นภาษาที่สอง จะพบว่ามีพจนานุกรมแบบ collocation ที่ช่วยบอกและอธิบายว่าคำไหนที่นิยมใช้คู่กับคำไหนบ่อยๆ เช่นคำว่า knowledge ก็จะมีคำที่นิยมใช้คู่กันคือ considerable, great, vast, comprehensive, sound, thorough, detailed, broad, extensive, inside, common, factual, practical, up-to-date เป็นต้น

2.1.1.4 complex lexical items

complex lexical items หมายถึง คำประสาน ซึ่งก็คือกลุ่มคำที่ปรากฏร่วมกัน และเมื่อแบ่งคำออกเป็นหน่วยย่อย จะมีส่วนหนึ่งที่สามารถอยู่ตามลำพังได้ และอีกส่วนหนึ่งที่ไม่สามารถอยู่ตามลำพังได้

ตัวอย่างคำในภาษาอังกฤษ

pretest precook immaterial immature devalue decode quickly happily noted lifted bigger longer

จากตัวอย่าง คำที่สามารถอยู่ตามลำพังได้ คือคำว่า test, cook, material, mature, value, code, quick, happy, note, lift, big และ long ส่วนคำที่ไม่สามารถอยู่ตามลำพังได้ คือคำว่า pre, im, de, ly, er และ ed

2.1.1.5. Particle verbs

Particle verbs หมายถึง กริยาวลี โดยสามารถเรียกชื่อได้หลายรูปแบบ คือ two-part verbs, three-part verbs, two-word verbs และ multi-word verbs เป็นต้น แต่ตามไวยากรณ์ในภาษาอังกฤษ กริยาวลี คือ phrasal verb โดยลักษณะของกริยาวลี จะมีคำกริยาหนึ่งคำและคำอื่นๆ อีกหนึ่งหรือสองคำ โดยเมื่อประกอบกันขึ้นเป็นกลุ่มคำ จะทำให้เกิดความหมายใหม่ขึ้นมาหรืออาจจะไม่มีเค้าความหมายของคำกริยาเดิมเลย

คำอื่นๆ ที่กล่าวถึงจากบริบทแรก จะเรียกว่าส่วนประกอบหรือ particles โดยมากจะเป็น down, in, off, out, up เป็นต้น เมื่อส่วนประกอบมาอยู่รวมกับคำกริยาจะเกิดเป็นความหมายใหม่ โดยส่วนประกอบจะเปลี่ยนความหมายเดิมของคำกริยา ซึ่งแตกต่างจากการใช้คำกริยาร่วมกับคำบุพบท หรือการใช้คำกริยาร่วมกับคำกริยาวิเศษณ์ เพราะคำบุพบทและ

คำกริยาวิเศษณ์ ไม่ได้เปลี่ยนความหมายเดิมของคำกริยาแต่ particle จะเปลี่ยนความหมายของคำกริยา

2.1.1.6 Named Entity

Named Entity หมายถึง คำนามเฉพาะ ซึ่งคือคำที่ทำหน้าที่ระบุชี้เฉพาะถึงสิ่งต่างๆ เช่น บุคคล ชื่อองค์กร ชื่อสถานที่ รวมไปถึงนิพจน์แสดงวันเวลา ปริมาณเงิน และเปอร์เซ็นต์

2.1.1.7 Figurative expressions

Figurative expressions จะแบ่งรูปแบบการเปรียบเทียบออกเป็น 2 รูปแบบดังนี้

1. Simile หมายถึง เป็นการนำของสองสิ่งมาเปรียบเทียบกัน โดยมีคำเชื่อม like /as เพื่อแสดงความคล้ายคลึงกันดังนี้

ตัวอย่าง

The grass is like a green carpet.
This hall is as quiet as a graveyard.
Love is like a rocket.

2. Metaphor หมายถึง คำอุปมาอุปไมย เป็นการเปรียบเทียบของสิ่งหนึ่งเป็นอีกสิ่งหนึ่ง โดยอาศัยความหมายแฝง ซึ่งหากดูตามโครงสร้างประโยคแล้วเหมือนไม่ใช่การเปรียบเทียบ

ตัวอย่าง

Prasit ’s home is an antique shop.
Non is a tiger when he is angry.
Pim was a goddess of beauty.

2.1.1.8 Institutionalized phrases

Institutionalized phrases หมายถึง การรวมกันของกลุ่มคำ ซึ่งคำที่เกิดขึ้นมาจากการผสมคำที่มีความหมายในตัวเองเข้าด้วยกัน แล้วกลายเป็นคำใหม่ที่มีเค้าความหมายเดิม หรือไม่คงเหลือเค้าความหมายเดิมดังตัวอย่างด้านล่าง

ตัวอย่าง เค้าความหมายเดิม

Traffic หมายถึง การจราจร

Light หมายถึง แสง

Traffic Light หมายถึง สัญญาณไฟจราจร

จากตัวอย่างข้างต้นการนำคำว่า traffic กับ light มาผสมกันจะเกิดเป็นคำใหม่ที่มีความหมายว่า สัญญาณไฟจราจร ซึ่งยังคงมีเค้าความหมายเดิม แต่ทั้งนี้ทั้งนั้นความหมายนั้นอาจจะเปลี่ยนไปโดยสิ้นเชิง และไม่คงเหลือเค้าความหมายเดิมของคำที่นำมาประสมกันเลยดังตัวอย่างด้านล่าง

ตัวอย่าง ไม่คงเหลือเค้าความหมายเดิม

Salt หมายถึง เกลือ

Pepper หมายถึง พริกไทย

Salt and Pepper หมายถึง ผมหงอก

2.1.1.9 terminology

terminology หมายถึง คำศัพท์เฉพาะทาง เป็นคำที่ใช้เฉพาะกลุ่มหรือเฉพาะอาชีพ ซึ่งสามารถยกตัวอย่างได้ดังนี้

ตัวอย่าง

white dwarf คือ ดาวแคระขาวเป็นดาวที่อุณหภูมิผิวสูงมากแต่ไม่ค่อยสว่าง

Law of Mass Action คือ อัตราการเกิดปฏิกิริยาเคมีซึ่งจะเป็นสัดส่วน โดยตรงกับความเข้มข้นของสารตั้งต้นที่เกิดปฏิกิริยา

2.1.1.10 light verb

light verb คือ กริยาเฉพาะที่มีความหมายในตัวเองตัวอย่าง เช่น do หรือ take แต่ที่แสดงความหมายซับซ้อนมากขึ้น เมื่อรวมกับคำอื่นโดยส่วนมากเป็นคำนาม เช่น do a trick or take a bat

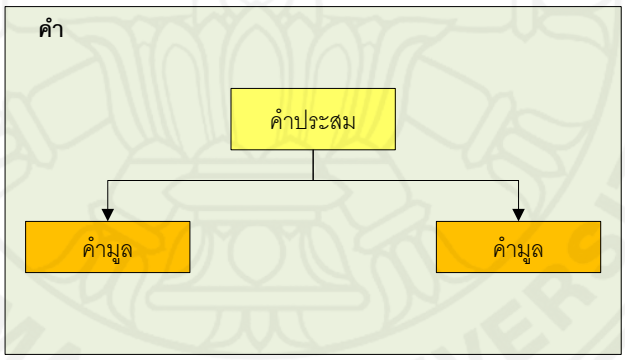
สำหรับในงานวิจัยฉบับนี้จะมุ่งเน้นศึกษาเกี่ยวกับคำนามประสม compounds noun ในภาษาไทยเพื่อทำให้คอมพิวเตอร์สามารถรับรู้และสามารถตรวจสอบเพื่อค้นหาคำนามประสมภายในประโยคก่อนที่คอมพิวเตอร์จะนำข้อมูลไปประมวลผลในลำดับถัดไป

2.1.2 คำและการสร้างคำในภาษาไทย

1. คำ คือ เสียงที่เปล่งออก แต่ต้องมีความหมายหนึ่งความหมาย

2. คำมูล คือ “คำพื้นฐานที่มีความหมายสมบูรณ์ในตัวเอง กล่าวคือ เป็นคำที่สร้างขึ้นโดยเฉพาะ เพื่อใช้เรียกสิ่งใดสิ่งหนึ่ง อาการใดอาการหนึ่ง โดยอาจเป็นคำไทยดั้งเดิมหรือเป็นคำที่มาจากภาษาอื่น ” (สืบค้นจาก โรงเรียนมหิดลวิทยานุสรณ์ , เอกสารประกอบการเรียน รายวิชา ท๔๐๑๐๕ หลักภาษาไทยในชีวิตประจำวัน)

3. คำประสม คือ คำที่เกิดจากการนำคำมูลตั้งแต่สองคำขึ้นไปมาผสมกัน แล้วเกิดความหมายใหม่หนึ่งหน่วยความหมาย หรือยังคงมีเค้าความหมายเดิม จากคำอธิบายข้างต้นสามารถสรุปออกมาได้ดังภาพที่ 2.1



ภาพที่ 2.1 การสร้างคำ

2.1.2.1 ประเภทของคำประสม

(1) คำที่นำมาประสมกันนำมาจากภาษาใดก็ได้

1. คำประสมที่เป็นคำไทยกับคำไทย ตัวอย่างเช่น แม่+บ้าน = แม่บ้าน , พ่อ+ตา = พ่อตา, โรง+เรียน = โรงเรียน

2. คำประสมที่เป็นคำไทยกับบาลีตัวอย่างเช่น ราช+วัง = ราชวัง, พล+เรือน = พลเรือน, รถ+ไฟ = รถไฟ

3. คำประสมระหว่างภาษาบาลี สันสกฤตหรือคำที่มาจากภาษาอื่น เช่น กิจ+ธุระ = กิจธุระ, วงศ์+ญาติ = วงศ์ญาติ, น้ำ+ซุป = น้ำซุป

(2) คำมูลที่นำมาผสมกัน อาจเป็นนาม สรรพนาม กริยา วิเศษณ์ บุพบท

1. คำประสมที่มีคำนามเป็นตัวตั้ง เช่น พ่อตา ปากจัด คนใช้ น้ำหวาน คนนอก ภาคใต้

2. คำประสมที่มีคำกริยาเป็นตัวตั้ง เช่น ยิงปืน เชื่อใจ ห่อหมก จับจอง อวดดี

3. คำประสมที่มีคำวิเศษณ์เป็นตัวตั้ง เช่น หลายใจ น่ารัก หวานเย็น

4. คำประสมที่มีคำบุพบทเป็นตัวตั้ง เช่น ใต้เท้า ที่นอน ใต้ดิน

2.1.3 ลักษณะคำประสม

1. ความหมายแต่งต่างจากเดิม เป็นการนำคำมูลตั้งแต่สองคำขึ้นไปมาผสมกัน โดยคำใหม่ที่เกิดขึ้นจะไม่มีเค้าความหมายเดิมของคำมูลที่นำมาผสมกัน ดังแสดงในตารางที่ 2.1

ตารางที่ 2.1

ความหมายแตกต่างจากเดิม

คำประสม	ความหมาย
พ่อครัว	ชายที่มีหน้าที่ประกอบอาหาร
แม่ทัพ	ชายหรือหญิงที่เป็นหัวหน้าของกองทัพ
ลูกน้ำ	ลูกของยุงซึ่งอยู่ในน้ำ
ตู้เย็น	ภาชนะคล้ายตู้ ให้ความเย็นแก่ของที่บรรจุอยู่

2. คงเค้าความหมายเดิม เป็นการนำคำมูลตั้งแต่สองคำขึ้นไปมาผสมกัน โดยคำใหม่ที่เกิดขึ้นจะมีเค้าความหมายเดิมของคำมูลที่นำมาผสมกัน ดังแสดงในตารางที่ 2.2

ตารางที่ 2.2

คงเค้าความหมายเดิม

คำประสม	ความหมาย
น้ำแข็ง	น้ำที่ถูกความเย็นจัดจนแข็งตัวเป็นก้อน
หมาป่า	ชื่อหมาหลายชนิดในวงศ์ Canidae มีถิ่นกำเนิดเกือบทั่วทุกภูมิภาคของโลก ขนลำตัวมีสีต่าง ๆ เช่น น้ำตาลเทา เทาปนแดง ฟันและเขี้ยวคมมาก นิสัยดุร้าย
แกงส้ม	แกงพวกหนึ่ง มีลักษณะคล้ายแกงเผ็ดแต่ไม่ใส่เครื่องเทศ กะทิ หรือ น้ำมัน

คำประสมสามารถแบ่งออกเป็นหลายประเภท ซึ่งในงานวิจัยฉบับนี้จะขอกล่าวถึงเฉพาะคำนามประสม

2.1.4 คำนามประสม

คำนามประสม คือ การนำคำมูลตั้งแต่สองคำขึ้นไปมาผสมกันและคำใหม่ที่เกิดขึ้นจะทำหน้าที่เป็นคำนาม ซึ่งจะมีลักษณะการประสมคำ ดังตารางที่ 2.3

ตารางที่ 2.3

Parts of speech Compound Nouns

Compound-elements	**Examples**	**ภาษาไทย**	**Examples**
noun + noun	bedroom water tank	นาม + นาม	แม่น้ำ พ่อบ้าน แปรงสีฟัน ฯลฯ
noun + verb	rainfall haircut train-spotting	นาม + กริยา	แบบเรียน เข็มกลัด ยาดม
noun + adverb	hanger-on passer-by	นาม + วิเศษณ์	น้ำแข็ง ถั่วเขียว หัวหอม

Compound-elements	Examples	ภาษาไทย	Examples
verb + noun	washing machine driving licence	กริยา + นาม	กินใจ เล่นตัว เข้าใจ ได้หน้า ฯลฯ
adverb + noun	onlooker bystander	วิเศษณ์ + คำนาม	อ่อนข้อ สองหัว
adjective + verb	dry-cleaning		
adverb + verb	output overthrow		
Participle (-ing) + คำนาม (Noun)	Dancing-teacher		
verb + verb		กริยา + กริยา	กันชน ไขควง

2.1.5 หนังสือคำนามประสม ศาสตร์และศิลป์ในการสร้างคำไทย

แนวทางการพิจารณาความแตกต่างระหว่างคำนามประสมกับนามวลี และ ประโยค

1. อรรถศาสตร์ เกณฑ์ทางอรรถศาสตร์เป็นเกณฑ์สำคัญที่มักนำไปเป็นหลักในการพิจารณาคำประสม กล่าวคือ หากข้อความใดมีความหมายที่ได้จากการรวมเข้าด้วยกันของหน่วยสมาชิกอย่างตรงไปตรงมา ข้อความนั้นจะถือว่าเป็นวลีหรือประโยค หากข้อความใดที่มีความหมายผิดไปจากการรวมของหน่วยย่อย เกิดเป็นความหมายใหม่ที่ไม่อาจคาดเดาได้จะถือว่าเป็นคำประสม ดังนั้นคำที่เกิดติดต่อกัน เช่น คอหอย ที่มีความหมายว่าเป็น คอของหอย ซึ่งเป็นความหมายที่ได้จากการประกอบกันโดยตรงของคำว่า คอ และ หอย จะนับว่าเป็นวลี แต่หากมีความหมายว่า ลูกกระเดือก ซึ่งเป็นสิ่งอื่นที่ไม่สามารถคาดคะเนได้ จากความหมายของคำว่า คอ และ หอย โดยตรง จะนับว่าเป็นคำประสม

ลักษณะทางอรรถศาสตร์อีกตัวอย่างหนึ่งที่มีประโยชน์ในการบอกความแตกต่างระหว่างคำประสม วลี และ ประโยค ยกตัวอย่างเช่น วลีมีโครงสร้างแบบคำนาม-คำนาม สามารถมีรูปแบบอรรถสัมพันธ์ได้เพียงสองแบบคือ สิ่งที่ถูกครอบครอง-เจ้าของ เช่น บ้านฉัน ลูกคุณ และ สิ่งประดิษฐ์-วัสดุ แต่คำนามประสม มีอรรถสัมพันธ์ระหว่างหน่วยสมาชิกที่หลากหลายมากกว่าวลี เป็นต้น (อัญชลี สิงห์น้อย, 2548, น. 133-134)

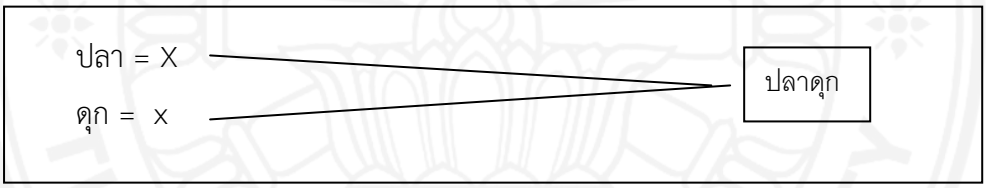
2. วากยสัมพันธ์ ลักษณะทางวากยสัมพันธ์ที่สำคัญเป็นที่รู้จักกันดีประการหนึ่งในการทดสอบคำประสมคือ สมาชิกคำนามประสมไม่อาจแยกกันได้ไม่ว่าจะเป็นการแยกกันหรือการแทรกคำอื่นระหว่างสมาชิก โดยยกตัวอย่าง คำที่เกิดต่อเนื่องกันเช่น หูช้าง ดังนั้นประโยค ฉันดัดหูช้าง

หากตีความหมายว่าเป็นวลี จะหมายถึง หูของช้าง แต่ยังสามารถแยกจากกันได้เป็น ช้างฉันดัดหู ซึ่งหากตีความว่าเป็นคำประสมที่หมายถึงกระจกข้างของรถแล้วจะไม่สามารถแยก หู และ ช้าง ออกจากกันได้ (อัญชลี สิงห์น้อย, 2548, น. 135-136)

3. สมาชิกคำนามประสมไม่อาจเชื่อมกับคำอื่นได้ คำประสมโดยทั่วไปไม่มีคำเชื่อม เช่น และ หรือ แต่ ฯลฯ ปรากฏระหว่างหน่วยสมาชิก ดังนั้นข้อความที่ขีดเส้นใต้ เช่น บ้านเล็กแต่เก่าคอห่าน และเป็ด หมูหัน หรือย่าง ฯลฯ จึงไม่น่าพิจารณาว่าเป็นคำนามประสม (อัญชลี สิงห์น้อย, 2548, น. 138)

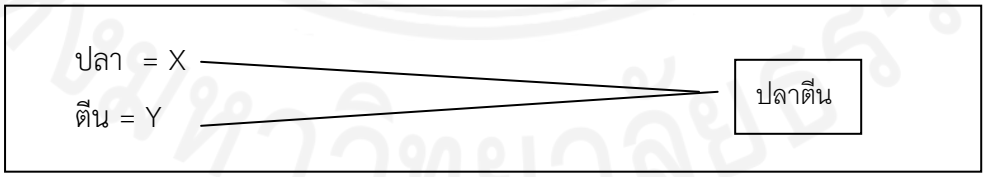
4. อรรถความสัมพันธ์พื้นฐานของคำนามประสมมีอรรถสัมพันธ์พื้นฐานได้ 2 รูปแบบ (อัญชลี สิงห์น้อย, 2548, น. 81-83)

1. X-x = Xx/x ซึ่ง X เป็น superordinate ส่วน x เล็กคือ subordinate เมื่อนำมาผสมกันจะได้ความสัมพันธ์ดังภาพที่ 2.2



ภาพที่ 2.2 superordinate-subordinate

2. X-Y = Xy ซึ่ง X และ Y เป็นคำสองคำที่นำมาผสมกันและทำให้เกิด unrelated meaning เมื่อนำมาผสมกันจะได้ดังภาพ 2.3

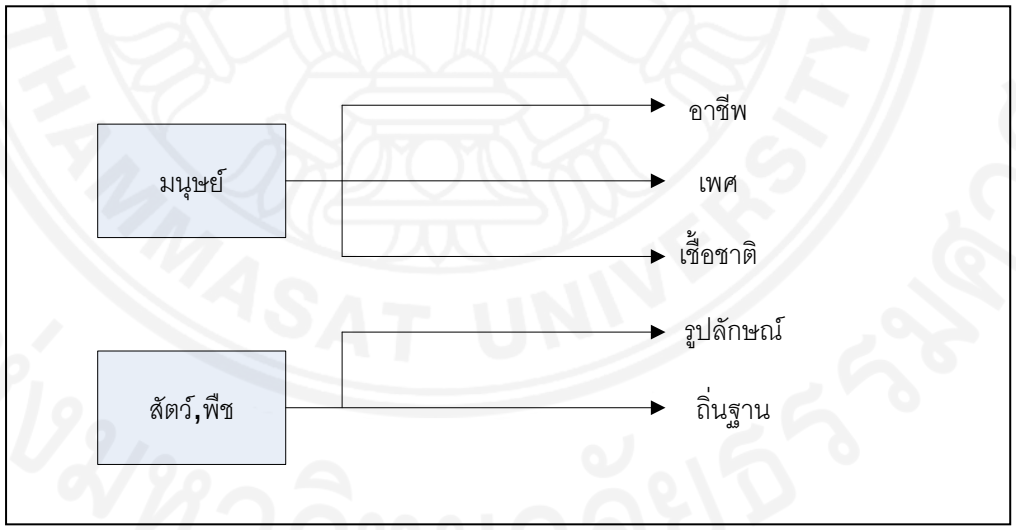


ภาพที่ 2.3 unrelated meaning

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 Downing and Levi (2520) On the creation and use of English compound nouns. Language

เป็นการจำแนกอรรถสัมพันธ์พื้นฐาน โดยพื้นฐานความเชื่อที่เกี่ยวกับถาวร และความเป็นปกติวิสัยของสิ่งต่างๆ จากการศึกษาคำประสมของ Downing แสดงให้เห็นว่าคำนามประสม ชนิดคำนามล้วน สามารถจำแนกออกเป็นอรรถสัมพันธ์รูปแบบต่างๆ บนพื้นฐานของความคิดหรือความเชื่อเกี่ยวกับ ความถาวร และ ความเป็นปกติวิสัยนี้เองที่กล่าวกันว่าสรรพสิ่งธรรมชาติอันได้แก่ มนุษย์ สัตว์ พืช มักถูกจำแนกตามลักษณะพื้นฐานของสิ่งนั้นๆ เช่น มนุษย์ มักมีการจำแนกโดยแบ่งตามรูปลักษณ์ และ ถิ่นฐาน ส่วนวัตถุธรรมชาติมักจำแนกตามส่วนประกอบที่มาและแหล่ง ตัวอย่างเช่น กรงนก จะถูกตีความ โดยอาศัยหลักความเป็นปกติวิสัยบนพื้นฐานของวัฒนธรรมได้ว่าเป็นที่อยู่อาศัยที่เตรียมไว้สำหรับนกอยู่มากกว่าที่จะมีใครตีความว่า สถานที่ที่เตรียมไว้สำหรับนกบินผ่านเป็นต้นดังภาพที่ 2.4



ภาพที่ 2.4 ความถาวร

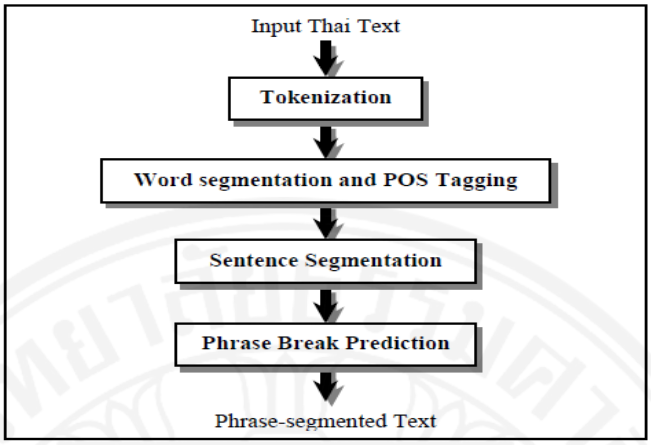
2.2.2 ณัฐกานต์ เพ่งผล (2545) การวิเคราะห์นามวลีภาษาไทยโดยอ้างอิงข้อมูลสถิติและข้อสนเทศทางภาษา

งานวิจัยฉบับนี้เป็นการค้นหาแนวทางและพัฒนาโปรแกรมวิเคราะห์นามวลีที่สามารถแยกแยะและกำหนดขอบเขตของแก่นนามวลี รวมทั้งสามารถระบุตำแหน่งของคำหลักแท้ของนามวลี ซึ่งแบ่งกระบวนการทำงานแบ่งออกเป็น 4 ขั้นตอน ได้แก่ การรวมคำให้เป็นหน่วยความหมายที่ใหญ่ขึ้น การวิเคราะห์ขอบเขตนามวลี การแบ่งนามวลีย่อย และ การหาตำแหน่งคำหลักแท้ของนามวลี ซึ่งในแต่ละขั้นตอนจะประมวลโดยอิงข้อมูลสถิติหรือข้อสนเทศทางภาษา โดยนำข้อมูลจากนิตยสารกินนรีจำนวน 100 เอกสาร และ นิตยสาร อ.ส.ท. จำนวน 100 เอกสาร โดยให้นักวิจัยทางภาษาสร้างกฎนามวลีและกฎการรวมคำจากเอกสาร ซึ่งมีทั้งหมด 83 กฎ และกฎการรวมคำ 90 กฎ

จากผลการทดลองพบว่า การรวมคำให้เป็นหน่วยความหมายที่ใหญ่ขึ้นสามารถเพิ่มความถูกต้องของการกำหนดขอบเขตของนามวลีจาก 78% เป็น 90% และ โปรแกรมสามารถหาคำหลักแท้ให้นามวลีได้ถูกต้อง 55% แต่ยังพบปัญหาในเรื่องของผลการวิเคราะห์คำประสมและนามวลีสำหรับเอกสารที่ระบบยังไม่ได้เรียนรู้ ซึ่งมีความถูกต้องไม่มากนักเนื่องจากฐานความรู้ภาษาที่รวบรวมได้จากเอกสาร 200 เอกสารไม่ครอบคลุมไวยากรณ์ในภาษาไทยทั้งหมดและการวิเคราะห์การทดลองจำกัดอยู่เฉพาะเอกสารที่ทำการทดลอง 2 กลุ่มเอกสาร ทำให้ไม่สามารถจัดการกับเอกสารประเภทอื่นได้ดีเท่าที่ควร

2.2.3 C. Hansakunbutheung, A. Thangthai, C. Wutiwiwatchai and R.Siricharoenchai (2548) Learning Methods and Features for Corpus-Based Phrase Break Prediction on Thai

บทความนำเสนอโปรแกรม 5 ชนิด ที่มีชื่อเสียงในการหาขอบเขตของวลี ได้แก่ POS Sequence Model, CART, RIPPER, SLIPPER และ Neural Network ซึ่งการทดลองเป็นการเปรียบเทียบวิธีการ 5 ชนิด เพื่อที่ใช้ในการการหาขอบเขตของวลี โดยข้อความที่นำมาใช้ในการทดสอบจะไม่มีคำที่ไม่มีความหมาย ซึ่งการทดลองและการเรียนรู้แต่ละวิธีการมาจากการ Train เพื่อหาจำนวนสูงสุดของข้อมูลจากใน Leaf node และ จำนวนของพาทิชั่นสำหรับค่าที่ดีที่สุด โดยลำดับแรกเป็นการนำข้อความภาษาไทยมาแบ่งเป็นหลายกลุ่มตามประเภทของ Thai Character, Digit หรือ สัญลักษณ์จากนั้น Thai Character จะถูกแบ่งเป็นคำ และ ติดแท็กด้วย POS จากนั้นจึงจะเข้าสู่การวิเคราะห์การทำนายหาขอบเขตของวลีดังแสดงในภาพที่ 2.5



ภาพที่ 2.5 กระบวนการหาขอบเขตของวลี, โดย C. Hansakunbutheung, A. Thangthai, C. Wutiwiwatchai and R.Siricharoenchai, 2548, NECTEC, Thailand

จากผลการทดลองวิธีการที่ดีที่สุดคือ CART โดยสามารถแบ่งได้ถูกต้อง 80.14% และมีชุดเชื่อมต่อที่ถูกต้อง 94.40% และไม่สามารถแบ่งได้ถูกต้อง 2.37% เพื่อเพิ่มประสิทธิภาพในการเรียนรู้ต้องเลือกลักษณะข้อมูลให้เหมาะสมกับคุณลักษณะที่จะนำมาใช้ในการทดสอบและจากการทดลองพบว่า โครงสร้างของ Part of Speech ได้รับการพิสูจน์แล้วว่าให้ประสิทธิภาพสูงสุดในการแยกคำ ซึ่งเป็นปัจจัยหนึ่งที่ส่งผลต่อการทำนายขอบเขตของวลีในภาษาอังกฤษ ดังนั้น ถ้าโครงสร้างของ Part of Speech มีความถูกต้องมากขึ้นจะเป็นปัจจัยหนึ่งที่ทำให้การทำนายหาขอบเขตของวลีมีประสิทธิภาพมากขึ้น

2.2.4 Kanyanut Kriengket, Krit Kosawat, Sunant Anchaleenukul (2550) A Computation Linguistics Study of Compound Nouns in Thai

เป็นงานวิจัยที่ทำการแบ่งคำประสมตามลักษณะทางความหมายซึ่งสามารถแบ่งเป็น 2 ประเภทดังนี้

1. คำประสมแบบเข้าศูนย์ (endocentric compound) คือคำประสมที่แสดงชนิดของส่วนประกอบหลัก ส่วนประกอบของคำประสมประเภทนี้มีหน่วยหลักและหน่วยขยายทำหน้าที่ร่วมกันดังภาพที่ 2.6

Structures				Elements	Words
Logic	Function	Semantic	POS		
$P\text{-}A_1$ ($\text{-}A_2$)	$H\text{-}M_1$ ($\text{-}M_2$)	[SHP-PUR]	N-V	ก้าม-วัด /kâ:m/–/wát/ (pincers)-(measure)	ก้ามวัด /kâ:m wát/ (callipers)
		[INS-PUR]	N-VP (VT-N)	เครื่อง-(ถ่วง-ล้อ) /khrɯ̂aŋ/–(/thùaŋ lɔ́:/) (instrument)-(balance-a wheel)	เครื่องถ่วงล้อ /khrɯ̂aŋ thùaŋ lɔ́:/ (wheel balance)
		[PROC-PUR]	N-VP (VT-N)-VP (VT-N)	ตู้-(อบ-ความร้อน)-(ฆ่า-เชื้อโรค) /tû:/–(/?òp khwamrɔ́:n/)–(/khâ: chɯ́arô:k/) (cabinet)-(fumigate with heat)-(destroy infection)	ตู้อบความร้อนฆ่าเชื้อโรค /tû: ?òp khwamrɔ́:n khâ: chɯ́arô:k/ (hot-air sterilizer)
		[PUR-OBJ]	VT-N	พัด-ลม /phát/-/lom/ (fan)-(wind)	พัดลม /phát lom/ (fan)
		[PUR-GOAL]	VT-VT	กัน-ชน /kan/-/chon/ (protect)- (crash)	กันชน /kan chon/ (bumper)

ภาพที่ 2.6 คำประสมแบบเข้าศูนย์ (endocentric compound)

โดย Kanyanut Kriengket, KritKo Sawat, Sunant Anchaleenukul, 2550, NECTEC, Thailand

2. คำประสมแบบออกนอกศูนย์ (exocentric compound) คำประสมที่มีส่วนความหมายแตกต่างไปจากส่วนประกอบ ส่วนประกอบแต่ละส่วนที่ประกอบกันเป็นคำประสมไม่มีส่วนหนึ่งส่วนใดเป็นส่วนหลัก แต่ทั้งสองส่วนจะทำหน้าที่ร่วมกันเป็นคำๆ เดียว แสดงความหมายใหม่ ดังภาพที่ 2.7

Structures				Elements	Words
Logic	Function	Semantic	POS		
$P_1\text{-}P_2$	$H_1\text{-}H_2$	[PART-PUR]	N-VT	หู-ฟัง /hŭ:/-/faŋ/ (ear)- (listen)	หูฟัง /hŭ: faŋ/ (earphone)
		[MAIN-POW]	N-N	แม่-แรง /mê:/–/rɛ:ŋ/ (mother)- (power)	แม่แรง /mê: rɛ:ŋ/ (jack)
		[FUNC-PUR]	N-N	สะพาน-ไฟ /sàpha:n/-/faj/ (bridge)-(electric)	สะพานไฟ /sàpha:n faj/ (cut-out)
		[QLT-THN]	CL-N	พวง-มาลัย /phuaŋ/-/ma:laj/ (bunch)-(wreath)	พวงมาลัย /phuaŋ ma:laj/ (steering wheel)
		[PUR-MET]	VT-VT	ไข-ควง /khăj/–/khwuaŋ/ (drive)-(screw)	ไขควง /khăj khwuaŋ/ (screwdriver)

ภาพที่ 2.7 คำประสมแบบออกนอกศูนย์ (exocentric compound)

โดย Kanyanut Kriengket, KritKo Sawat, Sunant Anchaleenukul, 2550, NECTEC, Thailand

2.2.5 กัญญาณัฐ เกรียงเกตุ (2550) การศึกษาคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์แนวภาษาศาสตร์คอมพิวเตอร์

งานวิจัยฉบับนี้เป็นการนำคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์มาวิเคราะห์โครงสร้างความสัมพันธ์ระหว่างคำที่มีต่อกันทั้ง วลี คำประสม และ คำมูล เป็นการศึกษาวิเคราะห์โครงสร้างหลายระดับของคำนามแสดงอุปกรณ์ด้ายวิทยาศาสตร์ ได้แก่ โครงสร้างผิวซึ่งประกอบด้วยโครงสร้างระดับชนิดของคำและระดับหน้าที่ โครงสร้างระดับตรรกะหรือโครงสร้างระดับกลางและโครงสร้างระดับลึก ได้แก่ โครงสร้างระดับความหมายเป็นการทำการทดลอง โดยใช้โปรแกรม Unitex มาสร้างโครงสร้างความสัมพันธ์ระหว่างส่วนประกอบของคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์ เพื่อทดสอบการตรวจจับคำจากโครงสร้างที่ประยุกต์มาจากโปรแกรม Unitex การทดลองเป็นการตรวจจับคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์โดยแบ่งออกเป็น 2 ส่วน คือ กราฟหลักสำหรับทดสอบการตรวจจับโครงสร้างกับคลังข้อมูลที่เป็นรายการคำศัพท์จากพจนานุกรม และ กราฟหลักสำหรับการตรวจจับโครงสร้างกับคลังข้อมูลที่ได้จากการรวบรวมบทความในนิตยสารที่เผยแพร่ทางอินเตอร์เน็ต

ผลการทดสอบรายการคำศัพท์จากพจนานุกรมพบว่า เครื่องคอมพิวเตอร์สามารถตรวจจับคำได้ 974 คำ เป็นคำที่ถูกต้อง 973 คำ และ ตรวจจับคำผิดอีก 1 คำ และ มีค่าความครบถ้วนเป็นร้อยละ 100 และ ค่าความแม่นยำเป็นร้อยละ 99.90

ผลการทดสอบจากคลังข้อมูลที่เป็นบทความแบ่งออกเป็น 2 ประเภท คือ คำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์ จำนวน 815 คำ และ คำนามแสดงอุปกรณ์ที่อยู่นอกสาขาวิทยาศาสตร์ จำนวน 1209 คำ

จากการทดสอบของคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์พบว่า มีค่าความครบถ้วนเป็น 71.41% และ ค่าความแม่นยำเป็น 30.62% ส่วนคำนามที่อยู่นอกสาขามีค่าความครบถ้วนเป็น 81.47% และ ค่าความแม่นยำเป็น 51.81% ส่วนปัญหาที่พบจากการทดสอบเกิดจากคำที่มาจากคลังข้อมูล ไม่ได้ปรากฏเป็นรายการคำศัพท์โดดๆ เช่นเดียวกับในพจนานุกรมแต่ปรากฏในประโยคที่มีบริบทแวดล้อม ทำให้เครื่องคอมพิวเตอร์ตรวจจับคำเกินความยาวที่ต้องการ อีกทั้งยังเกิดความผิดพลาดอื่นๆ คือ เครื่องคอมพิวเตอร์ตรวจจับคำ วลี หรือ หน่วยสร้างอื่น ที่มีการเว้นวรรคระหว่างส่วนประกอบเข้ามาด้วย จึงจำเป็นต้องอาศัยเงื่อนไขอื่นๆ เพิ่มเติมเพื่อช่วยในการคัดกรอง

2.2.6 ณรงค์กรรณ รอดทรัพย์ (2554) โครงสร้าง และ วากยสัมพันธ์ของนามวลีภาษาไทยในป้ายรณรงค์หาเสียงเลือกตั้งสมาชิกสภาผู้แทนราษฎร

เพื่อวิเคราะห์โครงสร้าง วากยสัมพันธ์ และอรรถสัมพันธ์ของหน่วยสมาชิกนามวลีภาษาไทย ซึ่งปรากฏในป้ายรณรงค์หาเสียงเลือกตั้งสมาชิกสภาผู้แทนราษฎร 2554 ผลการศึกษาพบว่า นามวลีภาษาไทยต้องประกอบด้วยอย่างน้อยคำนามและอาจมีหน่วยขยายนามอยู่ทางซ้ายของคำนามโดยหน่วยขยายนามปรากฏได้ตั้งแต่ 1 หน่วยขึ้นไปถึงมากที่สุดคือ 6 หน่วย ได้แก่ ประโยค คุณศัพท์ วลีแสดงปริมาณ วลีแสดงเจ้าของ บุพบทวลี คำนามขยาย หรือ คำชี้เฉพาะ ส่วนขยายเหล่านี้อาจปรากฏอย่างใดอย่างหนึ่งและ/หรืออาจประกอบกันเป็นส่วนขยายของคำนามหลัก แนวคิดซึ่งใช้วิเคราะห์และอธิบายภาษาในบทความวิจัยนี้ คือ ทฤษฎีไวยากรณ์หน้าที่นิยมแบบลักษณ์ภาษา (Functional-Typological Approach) ของ Talmy Givón โดยการวิจัยแย่งออกเป็นสองประเด็น คือ โครงสร้าง และ วากยสัมพันธ์นามวลีภาษาไทย กับ อรรถสัมพันธ์ของนามวลีภาษาไทยโดยใช้กรณีศึกษาเป็นนามวลีที่ปรากฏในป้ายรณรงค์หาเสียงเลือกตั้งสมาชิกสภาผู้แทนราษฎร

จากผลการทดลองโครงสร้าง และ วากยสัมพันธ์ของนามวลีภาษาไทยปรากฏในป้ายรณรงค์หาเสียงเลือกตั้งสมาชิกสภาผู้แทนราษฎร 2554 ปรากฏทั้งรูปแบบโครงสร้างคำนามคำเดียว และ รูปแบบโครงสร้างคำนามกับส่วนขยาย ส่วนอรรถสัมพันธ์ของหน่วยสมาชิกนามวลีภาษาไทยซึ่งปรากฏในป้ายรณรงค์หาเสียงเลือกตั้งปรากฏออกมา 6 ลักษณะ คือ สิ่งที่ถูกครอบครอง-เจ้าของ ประเด็น-สรรพสิ่ง สรรพสิ่ง-จุดประสงค์ สรรพสิ่ง-แหล่ง สรรพสิ่ง-ลำดับ ผู้ชำนาญ-สิ่งที่ชำนาญ

ตารางที่ 2.4

การเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

ผู้เขียน	เทคนิคที่ใช้	จุดเด่น	จุดด้อย
Downing and Levi (2520)	เป็นการจำแนกอรรถสัมพันธ์พื้นฐาน โดยพื้นฐานความเชื่อที่เกี่ยวกับถาวร และ ความเป็นปกติวิสัยของสิ่งต่างๆ	จำแนกคำนามประสมด้วยอรรถสัมพันธ์พื้นฐานจากพื้นฐานความเชื่อที่เกี่ยวกับความถาวร และ ความเป็นปกติวิสัย	การศึกษาไม่ใช่เป็นการวิจัยแบบทดลองเป็นเพียงแนวคิดพื้นฐานที่ใช้ในการคัดแยกคำนามประสม
ณัฐกานต์ เพ่งผล (2545)	การวิเคราะห์ขอบเขตนามวลี การแบ่งนามวลีย่อย และ การหาตำแหน่งคำหลักแท้ของนามวลี	การรวมคำให้เป็นหน่วยความหมายที่ใหญ่ขึ้นสามารถเพิ่มความถูกต้องของการกำหนดขอบเขตของนามวลี	ฐานความรู้ภาษาที่รวบรวมได้จากเอกสาร 200 เอกสารไม่ครอบคลุมไวยากรณ์ในภาษาไทยทั้งหมด
C.Hansakunbutheung, A. Thangthai, C.Wutiwiwatchai and R.Siricharoenchai (2548)	การทดลองเป็นการเปรียบเทียบวิธีการ 5 ชนิดเพื่อที่ใช้ในการการหาขอบเขตของวลีได้แก่ POS Sequence Model, CART, RIPPER, SLIPPER และ Neural Network	- วิธีการที่ดีที่สุดคือ CART - โครงสร้างของ Part of Speech ได้รับการพิสูจน์แล้วว่าให้ประสิทธิภาพสูงสุดในการแยกคำ	การเรียนรู้ต้องเลือกลักษณะข้อมูลให้เหมาะสมกับคุณลักษณะที่จะนำมาใช้ในการทดสอบ

ตารางที่ 2.4 (ต่อ)

การเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

ผู้เขียน	เทคนิคที่ใช้	จุดเด่น	จุดด้อย
Kanyanut Kriengket, KritKo Sawat, Sunant Anchaleenukul (2550)	งานวิจัยที่ทำการแบ่งคำประสมตามลักษณะทางความหมาย คือ คำประสมแบบเข้าศูนย์ และ คำประสมแบบออกนอกศูนย์	สามารถแบ่งประเภทของคำนามประสมออกเป็นสองประเภท คือ คำประสมแบบเข้าศูนย์ และ คำประสมแบบออกนอกศูนย์	การศึกษาไม่ใช่เป็นการวิจัยแบบทดลองเป็นหลักการที่ใช้ในการแบ่งประเภทคำนามประสม
กัญญาณัฐ เกรียงเกตุ (2550)	งานวิจัยฉบับนี้เป็นการนำคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์มาวิเคราะห์โครงสร้างความสัมพันธ์ระหว่างคำที่มีต่อกันทั้ง วลี คำประสม และ คำมูล	การทดสอบรายการคำศัพท์จากพจนานุกรมเครื่องสามารถตรวจจับได้ 974 คำ เป็นคำที่ถูกต้อง 973 คำ และตรวจจับคำผิดอีก 1 คำ และ มีค่าความครบถ้วนเป็นร้อยละ 100	การทดสอบคำนามที่มาจากคลังข้อมูลจะไม่ได้ปรากฏเป็นรายการคำศัพท์โดดในพจนานุกรมแต่ปรากฏในประโยคที่มีบริบทแวดล้อมทำให้เครื่องคอมพิวเตอร์ตรวจจับคำเกินความยาวที่ต้องการ

ตารางที่ 2.4 (ต่อ)

การเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

ผู้เขียน	เทคนิคที่ใช้	จุดเด่น	จุดด้อย
ณรงค์กรรณ รอดทรัพย์ (2554)	วิเคราะห์โครงสร้าง วากยสัมพันธ์ และ อรรถสัมพันธ์ของหน่วยสมาชิกนามวลีภาษาไทย	ส่วนอรรถสัมพันธ์ของหน่วยสมาชิกนามวลีภาษาไทยซึ่งปรากฏในป้ายรณรงค์หาเสียงเลือกตั้งปรากฏออกมา 6 ลักษณะ คือ สิ่งที่ถูกครอบครอง-เจ้าของ ประเด็น-สรรพสิ่ง สรรพสิ่ง-จุดประสงค์ สรรพสิ่ง-แหล่ง สรรพสิ่ง-ลำดับ ผู้ชำนาญ-สิ่งที่ชำนาญ	ข้อมูลที่ใช้ในการทดลองอาจยังไม่ครอบคลุมอรรถสัมพันธ์ของหน่วยสมาชิกนามวลีภาษาไทยลักษณะอื่นๆ

บทที่ 3
วิธีการดำเนินงานวิจัย

งานวิจัยฉบับนี้นำเทคนิควิทยาการคอมพิวเตอร์ ซึ่งเป็นการนำโครงสร้างของคำนามประสมมาใช้ในการตรวจสอบ เพื่อคัดแยกคำนามประสมออกจากประโยค ด้วยวิธีทางอัลกอริทึม จากนั้นจึงแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น โดยอ้างอิงจากลำดับของคำจากซ้ายไปขวาเพื่อจัดกลุ่มของคำก่อนนำไปวิเคราะห์ เพื่อใช้ในการตัดแยกคำนามประสมออกจากประโยค ด้วยวิธีการสองขั้นตอนคือ การตรวจสอบคำนามประสมและการพิจารณาคำนามประสม ทำให้คอมพิวเตอร์สามารถนำข้อมูลที่ผ่านการตรวจสอบไปใช้งานได้อย่างมีประสิทธิภาพมากขึ้น โดยมีขั้นตอนการดำเนินการ ดังต่อไปนี้

ขอบเขตการทดลอง แนวคิดพื้นฐาน ขั้นตอนการทำงานของระบบ องค์ประกอบภาพรวมของระบบ ขั้นตอนการทดลอง และผลการทดลอง พร้อมทั้งอธิบายถึงปัญหาต่างๆ ที่เกิดขึ้นในการทดลอง เครื่องมือที่ใช้สำหรับพัฒนาระบบ การฝึกฝนทดลอง และการวัดผลการทดลอง

3.1 ความเป็นมาและความสำคัญของปัญหา

การทดลองสำหรับงานวิจัยนี้ จะทำการทดลองกับคลังข้อมูล 5 Million Words For BEST2010 ประเภท encyclopedia ที่มีการหา Part of speech จากโปรแกรม Swath 2.0.1 เรียบร้อยแล้ว ซึ่งบทความที่ใช้ในการคัดแยกคำนามประสมออกจากประโยคเป็นบทความที่มีประโยคไม่ยาวมากนัก มีการเว้นวรรคของแต่ละประโยคอย่างชัดเจนและไม่สามารถตรวจสอบคำที่ Swath ไม่สามารถแบ่งคำ หรือ หา Part of Speech ออกมาได้ ดังแสดงในภาพที่ 3.1 โดยโปรแกรม Swath มองคำว่าสัตว์เลี้ยงเป็นคำเดียวกัน ทำให้ระบบไม่สามารถนำคำว่า สัตว์เลี้ยง ไปวิเคราะห์เพื่อคัดแยกคำนามประสมออกจากประโยคในขั้นตอนต่อไปได้

id	name	POSW	ConvertSW	POSLex	ConvertLW
1	หมู	NCMN	N	N	N
2	เป็น	VSTA	V	ADJ	S
3	สัตว์เลี้ยง	NCMN	N	N	N
4	ที่	PREL	O	N	N

ภาพที่ 3.1 คำที่ตรวจสอบไม่ได้

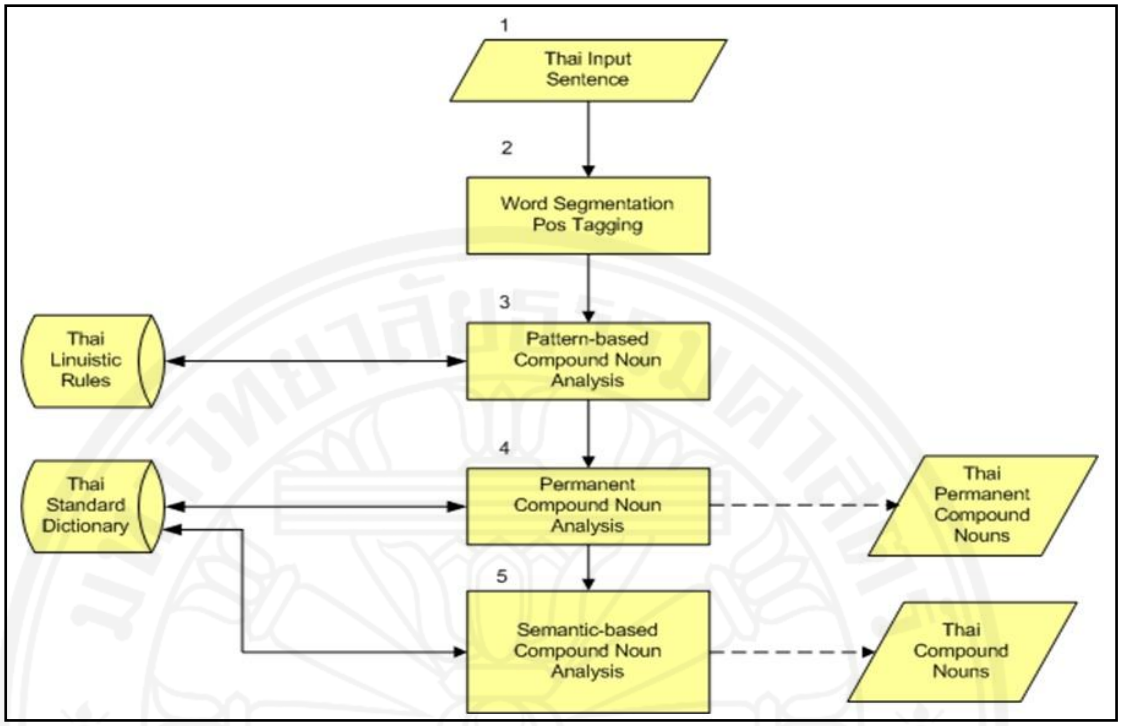
3.2 แนวคิดพื้นฐาน

แนวคิดที่เสนอในงานวิจัยนี้ได้ทำการศึกษาเกี่ยวกับการวิเคราะห์และคัดแยกคำนามประสมออกจากประโยค โดยใช้โครงสร้างของคำนามประสม เพื่อเป็นคัดแยกคำนามประสมออกจากประโยค จากนั้นจึงแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น ด้วยวิธีการจับคู่ความสัมพันธ์ของคำ ก่อนนำไปวิเคราะห์ เพื่อใช้ในการคัดแยกคำนามประสมออกจากประโยค ซึ่งคำที่เกิดอาจมีตั้งแต่สองพยางค์ขึ้นไปจนถึงสี่พยางค์ จากนั้นจึงนำผลลัพธ์ที่ได้มาตรวจสอบด้วยวิธีการทางอัลกอริทึมเพื่อคัดแยกคำนามประสมออกจากประโยค

ภายในประโยคของภาษาไทย ประกอบด้วย หน่วยย่อยภายในประโยค ซึ่งประกอบด้วยคำหลายประเภททำให้การหาโครงสร้างของคำนามประสมค่อนข้างมีความซับซ้อนและเกิดความคลุมเครือในการกำหนดขอบเขตของคำนามประสม การนำโครงสร้างของคำนามประสมมาช่วยในการคัดแยกคำนามประสมออกจากประโยคเป็นส่วนหนึ่งในการช่วยลดความซับซ้อนนี้ลงได้ แต่ก็ยังไม่ใช่ทั้งหมดจึงต้องมีการแยกส่วนประกอบย่อยๆ ออกมา เพื่อนำไปวิเคราะห์ต่อในอีกสองขั้นตอนคือ การตรวจสอบคำนามประสม และ การพิจารณาคำนามประสม

3.3 ขั้นตอนการทำงานของระบบ

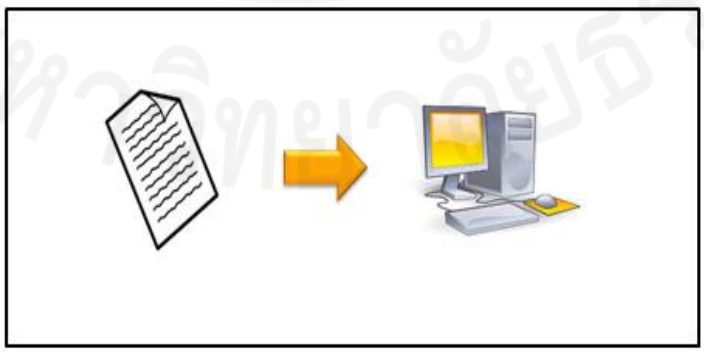
การทำงานของระบบเป็นการนำบทความภาษาไทยจากคลังข้อมูล 5 Million Words For BEST2010 ประเภท encyclopedia มาผ่านโปรแกรม swath เพื่อคัดแยกและติดป้าย Part of Speech ของคำแต่ละคำ จากนั้นจึงนำโครงสร้างของคำนามประสมมาคัดแยกคำนามประสมออกจากประโยค เพื่อคัดแยกคำแต่ละคำที่อาจจะเป็นส่วนประกอบของคำนามประสมออกจากประโยค จากนั้นจึงแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น โดยอ้างอิงการจับคู่ความสัมพันธ์ของคำ เพื่อจับกลุ่มของคำก่อนนำไปวิเคราะห์ เพื่อใช้ในการคัดแยกคำนามประสมออกจากประโยค ด้วยวิธีการสองขั้นตอนคือ การตรวจสอบคำนามประสมและการพิจารณาคำนามประสม



ภาพที่ 3.2 องค์ประกอบภาพรวมของระบบ

3.3.1 Thai Input Sentence

การนำบทความเข้าสู่ระบบดังภาพที่ 3.3 เป็นการเตรียมข้อมูลก่อนทำการ extract compound noun ซึ่งเป็นส่วนหนึ่งที่อยู่ในขั้นตอนการทำงานของระบบดังภาพที่ 3.2 โดยการตัดประโยคออกจากบทความทีละหนึ่งบรรทัด เพื่อนำเข้าสู่ระบบและตรวจสอบหาคำเชื่อมภายในประโยค ซึ่งจะเป็นส่วนช่วยในการแบ่งกลุ่มคำออกจากกันก่อนที่จะนำประโยคไปประมวลผลในลำดับถัดไป



ภาพที่ 3.3 การนำบทความเข้าสู่ระบบ

3.3.2 Word Segmentation Pos Tagging

เป็นการหาโครงสร้างของคำแต่ละคำภายในประโยค โดยใช้หลักการดังนี้

1. เป็นการหาโครงสร้างของคำแต่ละคำซึ่งเป็นส่วนหนึ่งที่อยู่ในขั้นตอนการทำงานของระบบดังภาพที่ 3.2 โดยใช้ Part of Speech ของโปรแกรม swath ดังภาพที่ 3.4 (POS = Part of Speech ของ swath, Description = คำอธิบายกลุ่มของ POS) เป็นตัวติดป้ายแบ่งแยกโครงสร้างของคำแต่ละคำดังแสดงในภาพที่ 3.6 โดยเป็นการแบ่งแยกโครงสร้างของคำแต่ละคำ (POSW = Part of Speech ของ swath, ConvertSw = Convert Part of Speech ของ Swath เพื่อให้เข้ากับกฎคำนามประสมของงานวิจัยฉบับนี้ , PosLex = Pos of Speech ของ Lexitron, ConvertLW = Convert Part of Speech ของ Lexitron เพื่อให้เข้ากับกฎคำนามประสมของงานวิจัยฉบับนี้)

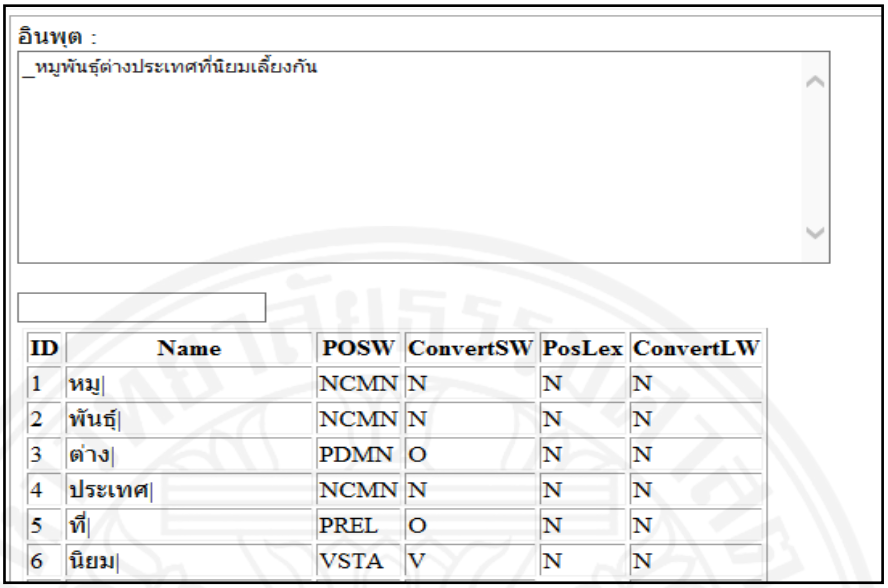
No.	POS	Description
1	NPRP	Proper noun
2	NCNM	Cardinal number
3	NONM	Ordinal number
4	NLBL	Label noun
5	NCMN	Common noun
6	NTTL	Title noun
7	PPRS	Personal pronoun

ภาพที่ 3.4 Part of Speech ของ swath

2. เป็นการหาโครงสร้างของคำแต่ละคำซึ่งเป็นส่วนหนึ่งที่อยู่ในขั้นตอนการทำงานของระบบดังภาพที่ 3.2 โดยใช้พจนานุกรม Lexitron ดังภาพที่ 3.5 (POS Lexitron = POS ของพจนานุกรม Lexitron) เป็นตัวติดป้ายแบ่งแยกโครงสร้างของคำแต่ละคำ ดังภาพที่ 3.6 เป็นการแบ่งแยกโครงสร้างของคำแต่ละคำโดย (POSW = Part of Speech ของ swath, ConvertSw = Convert Part of Speech ของ Swath เพื่อให้เข้ากับกฎคำนามประสมของงานวิจัยฉบับนี้ , PosLex = Pos of Speech ของ Lexitron, ConvertLW = Convert Part of Speech ของ Lexitron เพื่อให้เข้ากับกฎคำนามประสมของงานวิจัยฉบับนี้)

POS Lexitron
ADV,ADJ
V
CLAS,CLTV
CONJ
N,PRON

ภาพที่ 3.5 Part of Speech ของ Lexitron


อินพุต :

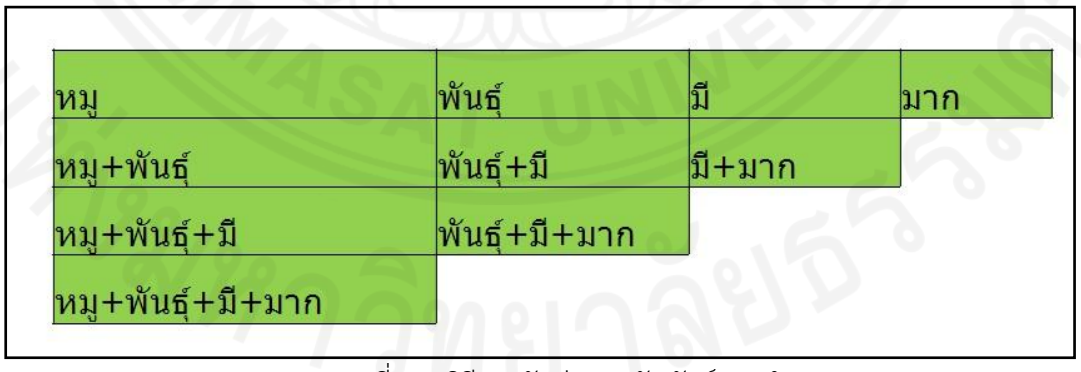
_หมูพันธุ์ต่างประเทศที่นิยมเลี้ยงกัน

ID	Name	POSW	ConvertSW	PosLex	ConvertLW
1	หมู|	NCMN	N	N	N
2	พันธุ์|	NCMN	N	N	N
3	ต่าง|	PDMN	O	N	N
4	ประเทศ|	NCMN	N	N	N
5	ที่|	PREL	O	N	N
6	นิยม|	VSTA	V	N	N

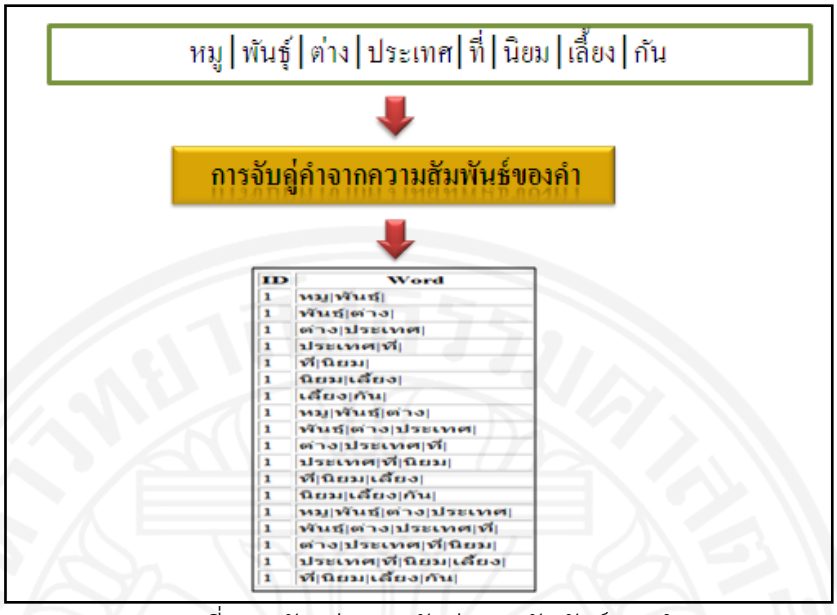
ภาพที่ 3.6 แบ่งแยกโครงสร้างของคำแต่ละคำ

3.3.3 Pattern-based Compound Noun Analysis

เป็นการแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้นซึ่งเป็นส่วนหนึ่งที่อยู่ในขั้นตอนการทำงานของระบบดังภาพที่ 3.2 โดยอ้างอิงจากความสัมพันธ์ของคำด้วยวิธีการจับคู่ความสัมพันธ์ของคำตั้งแต่ 2-4 Gram ดังแสดงในภาพที่ 3.7 และ ภาพที่ 3.8 (ID = จากภาพหมายเลข 1 หมายถึง คำ 1 คำ, Word = คำแต่ละคำที่ถูกจับคู่ความสัมพันธ์ให้กลายเป็นคำ 1 คำ)



ภาพที่ 3.7 วิธีการจับคู่ความสัมพันธ์ของคำ



ภาพที่ 3.8 ตัวอย่างการจับคู่ความสัมพันธ์ของคำ

3.3.4 Thai Linguistic Rules

เป็นการจับกลุ่มจากโครงสร้าง Part of Speech ของคำนามประสมซึ่งเป็นส่วนหนึ่งที่อยู่ในขั้นตอนการทำงานของระบบดังภาพที่ 3.2 โดยเป็นการใช้โครงสร้างคำนามประสม เพื่อรวมคำที่จะเป็นนามประสมออกจากประโยคซึ่งกฎที่นำมาใช้นำมาจากเอกสารต่างดังนี้

1. หนังสือคำนามประสม ศาสตร์และศิลป์ในการสร้างคำไทย (อัญชลี สิงห์น้อย, 2548)

2. หนังสือระบบคำภาษาไทย (สุนันท์ อัญชลีนุกูล, 2546)

3. การศึกษาคำนามแสดงอุปกรณ์ ด้านวิทยาศาสตร์แนวภาษาศาสตร์คอมพิวเตอร์ (กัญญาณัฐ เกรียงเกตุ, 2550)

4. คำนามประสม (Compound Nouns) สืบค้นจาก (Guessing the Meanings of Compound Words, 2555)

5. COMPOUND NOUNS IN ENG สืบค้นจาก (Edufind, Parts of speech of Compound nouns)

ตารางที่ 3.1

กฎคำนามประสม

จำนวนคำที่ผสมกันแล้วเป็นคำนามประสม					เอกสารลำดับที่
2-Gram		กฎ	คำที่พบ		
	1	น+น	ปากกา	ปาก+กา	2,3,4,5
			พวงมาลัย	พวง+มาลัย	
	2	น+ก	ตราชู	ตรา+ชู	1,2,3,5
			ปากคีบ	ปาก+คีบ	
			หัวเผา	หัว+เผา	
	3	ก+น	พัดลม	พัด+ลม	1,2,3,4,5
	4	ก+ก	ไขควง	ไข+ควง	3,2
			กันชน	กัน+ชน	
ว=วิเศษณ์	5	น+ว	น้ำแข็ง	น้ำ+แข็ง	5
	6	ว+น	หวานใจ	หวาน+ใจ	4,5
จน=ลำดับที่,ลน ลักษณะนาม	7	จน+ลน	สามขา	สาม+ขา	3
	8	น+บ	ไฟข้าง	ไฟ+ข้าง	1,2,4
3,4-Gram	9	น+น+น	เกจหัวเทียน	เกจ+หัว+เทียน	3
			แผงแป้นอักษร	แผง+แป้น+อักษร	
			กระปุกเกียร์ธรรมดา	กระปุก+เกียร์+ธรรมดา	
			ปั้มเชื้อเพลิงไฟฟ้า	ปั้ม+เชื้อเพลิง+ไฟฟ้า	
	10	น+ก+น	เครื่องพิมพ์เลเซอร์	เครื่อง+พิมพ์+เลเซอร์	1,2,3
			เครื่องฟอกไอเสีย	เครื่อง+ฟอก+ไอเสีย	
	11	น+น+ก	สายไฟพ่วง	สาย+ไฟ+พ่วง	3
	12	น+น+น+น	ล็อคแกนพวงมาลัย	ล็อค+แกน+พวง+มาลัย	3
	13	น+จน+ลน	ขวดสามคอ	ขวด+สาม+คอ	3

3.3.5 Permanent Compound Noun Analysis

การตรวจสอบคำนามประสม เมื่อผ่านกฎคำนามประสมแล้ว ขั้นตอนต่อไปเป็นการนำคำที่ผ่านกฎมาเข้าขั้นตอนการตรวจสอบคำนามประสม ดังแสดงในภาพที่ 3.9 โดยจะทำการตรวจสอบดังนี้

1. พบคำที่ค้นหาในพจนานุกรม Lexitron (Check_Lexitron) และ ทำหน้าที่เป็นคำนามสำหรับกรณีที่พบ และคำที่ไม่ใช่คำนามจะถูกนำเข้าสู่ขั้นตอนการพิจารณาคำนามประสม

2. พบคำที่ค้นหาในพจนานุกรมฉบับราชบัณฑิตยสถาน (Check_RadChaBudit) และทำหน้าที่เป็นคำนามกรณีที่พบ และ คำที่ไม่ใช่คำนามจะถูกนำเข้าสู่ขั้นตอนการพิจารณาคำนามประสม

A	B	C	D	E	F	G	H	I	J
id	Word	Swatch	Lexitron	PredicSw	PredicLw	Check_Lexitron	Check_RadChaBudit	IR	Score
1	การ|ให้|	N	T	N|	NV	0	0	17	50
2	ให้|อาหาร|	N	T	|N	VN	0	0	18	50
3	การ|ให้|อาหาร|	N	T	N|N	NVN	0	0	9	50

ภาพที่ 3.9 การตรวจสอบคำนามประสม

3.3.6 Semantic-based Compound Noun Analysis

การพิจารณาคำนามประสม กรณีที่ไม่พบคำนามประสมในพจนานุกรมต้องพิจารณาความแตกต่างระหว่างคำนามประสม นามวลี และประโยค ดังภาพที่ 3.10

1. กรณีที่พบว่าเป็นคำ superordinate-subordinate จะถือว่าเป็นคำนามประสม โดยใช้ความหมายของ subordinate เป็นตัวตั้งจากนั้นจึงนำ superordinate เข้าไปค้นหาในความหมายของ subordinate ว่ามี superordinate อยู่ในความหมายของ subordinate นั้นหรือไม่ ถ้าพบว่าเป็นส่วนหนึ่งของ subordinate และ ทั้ง superordinate และ subordinate มี Part of Speech เป็นคำนาม (NN) จะถือว่าเป็นคำนามประสมตัวอย่างเช่น superordinate = ปลา(N), subordinate = ดุก(N) โดยการตรวจสอบ ดุก หมายถึง ชื่อ ปลาน้ำจืดทุกชนิดในสกุล clarias จากการตรวจสอบข้างต้นจะเห็นว่ามีคำว่า ปลา คือส่วนที่ขีดเส้นใต้อยู่ในส่วนความหมายของคำว่า ดุก และ คำว่า ปลา และ ดุกต่างทำหน้าที่เป็นคำนาม จึงสามารถสรุปได้ว่า ปลาดุก คือคำนามประสม

2. อรรถศาสตร์ อ้างอิงเรื่องโครงสร้างการแสดงความเป็นเจ้าของ (Owner) เป็นการนำคำที่ต้องการค้นหาเติมคำว่า “ของ” ลงไปจากนั้นจึงนำไปค้นในคลังข้อมูล ถ้ามีความเป็นเจ้าของจะถูกคัดออก

3. ลำดับการเรียงคำ เนื่องจากคำนามประสมจะไม่สามารถสลับที่ (Swap) และแยกคำได้ (Sperate) เป็นการนำคำที่ต้องการค้นหามาทดลองสลับที่หรือแยกคำ จากนั้นจึงนำไปค้นหาในคลังข้อมูล เพื่อใช้ในการตรวจสอบ กรณีที่พบว่า สามารถสลับที่หรือแยกคำได้จะถูกคัดออก

id	Word	superordinate-subordinate	Swap	Score	Sperate	Score	Owner	Score
1	การ|ลวก|	0	0	1	0	1	0	1
2	ชาวบ้าน|ทั่วไป|	0	0	1	0	1	0	1
3	ก้าวหน้า|ไป|	0	0	1	0	1	0	1
4	ปรับปรุง|ขึ้น|	0	0	1	0	1	0	1
5	ยกพื้น|สูง|	0	0	1	0	1	0	1
6	ผู้|เลี้ยง|	0	0	1	0	1	0	1
7	ผู้|เลี้ยง|	0	0	1	0	1	0	1
8	ผู้|เลี้ยง|หมู|	0	0	1	0	1	0	1
9	การ|ให้|อาหาร|	0	0	1	0	1	0	1

ภาพที่ 3.10 พิจารณาคำนามประสม

3.3.7 คำนวณเปอร์เซ็นโอกาสที่จะเป็นคำนามประสม

แบ่งเป็นสองส่วนคือ ส่วนตรวจสอบคำนามประสม และ ส่วนพิจารณาคำนามประสมโดยคิดเป็นเปอร์เซ็นดังนี้

3.3.7.1. ตรวจสอบคำนามประสม

1. กรณีพบคำที่ค้นหาในพจนานุกรม Lexitron จะให้ค่าคะแนนเป็น 1 กรณีที่ไม่พบ จะให้ค่าคะแนนเป็น 0

2. กรณีพบคำที่ค้นหาในพจนานุกรมฉบับราชบัณฑิตยสถาน จะให้ค่าคะแนนเป็น 1 กรณีที่ไม่พบ จะให้ค่าคะแนนเป็น 0

กรณีที่พบข้อใดข้อหนึ่งจะถือว่ามีโอกาสที่จะเป็นคำนามประสม 100% และจะไม่มีการนำมาตรวจสอบในส่วนพิจารณาคำนามประสม

3.3.7.2. พิจารณาคำนามประสม

1. กรณีที่พบว่าเป็น superordinate-subordinate จะถือว่าเป็นคำนามประสม 100%

2. กรณีที่ไม่มีโครงสร้าง การแสดงความเป็นเจ้าของ จะให้ค่าคะแนนเป็น 1

3. ลำดับการเรียงคำ

3.1 ไม่สามารถสลับที่ได้ จะให้ค่าคะแนนเป็น 1

3.2 ไม่สามารถแยกคำได้ จะให้ค่าคะแนนเป็น 1

จากข้อ 2, 3.1, 3.2 สามารถนำมาคิดเป็นเป็นเปอร์เซ็นได้ดังนี้

1. กรณีที่พบทั้ง 3 ข้อ สามารถสรุปได้ว่ามีโอกาสที่จะเป็นคำนามประสม 100%

2. กรณีที่พบทั้ง 2 ข้อ สามารถสรุปได้ว่ามีโอกาสที่จะเป็นคำนามประสม 0%

3. กรณีที่พบ 1 ข้อ สามารถสรุปได้ว่ามีโอกาสที่จะเป็นคำนามประสม 0%

3.4 เครื่องมือที่ใช้สำหรับพัฒนาระบบ

เครื่องมือและซอฟท์แวร์คอมพิวเตอร์ที่ใช้ในงานวิจัยฉบับนี้แสดงในตารางที่ 3.2

ตารางที่ 3.2

เครื่องมือที่ใช้สำหรับงานวิจัย

Hardware	- Intel Core i7
Operating System	Microsoft Windows XP Professional
Software	- Swath 2.0.1 - Paint - Notepad -Microsoft visual studio 2010 -MySql

3.5 การฝึกฝนทดลอง

ผลการทดลองในการคัดแยกคำนามประสมออกจากประโยค จากการทดลองได้นำประโยค 100 ประโยค มาทำการคัดแยกคำนามประสมออกจากประโยค พบว่าสามารถคัดแยกคำนามประสมออกจากประโยคได้ดี เมื่อโปรแกรม swath สามารถแบ่งคำออกจากประโยคได้ถูกต้อง เพราะถ้าแบ่งคำผิดตั้งแต่เริ่มต้น จะทำให้กระบวนการอื่นๆ มีความผิดพลาดตามไปด้วย

3.6 การวัดผลการทดลอง

ข้อมูลดิบได้มาจากคลังจากคลังข้อมูล 5 Million Words For BEST 2010 ประเภท encyclopedia จำนวน 100 บรรทัด เป็นบทความทางด้านสารานุกรม เพื่อใช้ทดสอบคัดแยกคำนามประสมออกจากประโยคจำนวน 130 คำ ประกอบด้วย ตัวอักษรจำนวน 1013 ตัว นำมาคัดแยกคำนามประสมออกจากประโยค

การวัดผลความถูกต้องของการคัดแยกคำนามประสมออกจากประโยคของงานวิจัยนี้เป็นการคำนวณค่าความครบถ้วน (recall) และ ค่าความแม่นยำ (precision) ของเครื่องคอมพิวเตอร์

3.6.1 ค่าความครบถ้วน (Recall) คือ ค่าที่ได้จากการนำคำที่เครื่องคอมพิวเตอร์มาคัดแยกคำนามประสมออกจากประโยคได้และตรวจสอบแล้วว่าเป็นคำที่ถูกต้อง ซึ่งในที่นี่คือจำนวนคำนามประสมที่คอมพิวเตอร์สามารถคัดแยกออกมาได้ถูกต้อง หารด้วยจำนวนคำนามประสมที่ถูกต้องทั้งหมด แล้วคูณด้วย 100 เพื่อให้ค่าความครบถ้วนเป็นเปอร์เซ็นต์ ดังภาพที่ 3.11

สูตรในการหาค่าความครบถ้วน มีดังนี้

$$\frac{\text{คำที่ถูกต้อง}}{\text{คำที่ใช้ในการทดสอบทั้งหมด}} \times 100 = \text{Recall (\%)}$$

ภาพที่ 3.11 ค่าความครบถ้วน (Recall)

3.6.2 ค่าความแม่นยำ (Precision) คือ ค่าที่ได้จากการนำคำที่เครื่องคอมพิวเตอร์คัดแยกคำนามประสมออกจากประโยคได้ถูกต้อง ซึ่งในที่นี่คือจำนวนคำนามประสมที่คอมพิวเตอร์สามารถคัดแยกออกมาได้ถูกต้อง หารด้วยจำนวนคำที่เครื่องคอมพิวเตอร์สามารถบอกขอบเขตคำนามประสมออกจากประโยคได้ทั้งหมด แล้วคูณด้วย 100 เพื่อให้ได้ค่าความแม่นยำเป็นเปอร์เซ็นต์ ดังแสดงในภาพที่ 3.12

สูตรในการหาค่าความแม่นยำ มีดังนี้

$$\frac{\text{คำที่ถูกต้อง}}{\text{คำที่พบ}} \times 100 = \text{Precision (\%)}$$

ภาพที่ 3.12 ค่าความแม่นยำ (Precision)

บทที่ 4
ผลการวิจัย

ในบทนี้จะกล่าวถึงการทดลองและผลการทดลองคัดแยกคำนามประสมออกจากประโยคโดยวิเคราะห์การคัดแยกคำนามประสมออกจากประโยคด้วยเครื่องคอมพิวเตอร์เปรียบเทียบความถูกต้องกับการคัดแยกคำนามประสมจากผู้เชี่ยวชาญ พร้อมทั้งอธิบายถึงผลความผิดพลาดที่เป็นปัญหาก่อนการประมวลผลขั้นต้น และลักษณะปัญหาต่างๆที่เกิดขึ้นในการทดลอง

การคัดแยกคำนามประสมออกจากประโยคเป็นการนำโครงสร้างของคำนามประสมมาใช้ในการตรวจสอบ เพื่อคัดแยกคำนามประสมออกจากประโยค ด้วยวิธีทางอัลกอริทึม จากนั้นจึงแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น โดยอ้างอิงจากลำดับของคำจากซ้ายไปขวา เพื่อจัดกลุ่มของคำก่อนนำไปวิเคราะห์ เพื่อใช้ในการคัดแยกคำนามประสมออกจากประโยค โดยด้วยวิธีการสองขั้นตอนคือ การตรวจสอบคำนามประสม และการพิจารณาคำนามประสม

4.1 การเตรียมข้อมูล (Preprocessing)

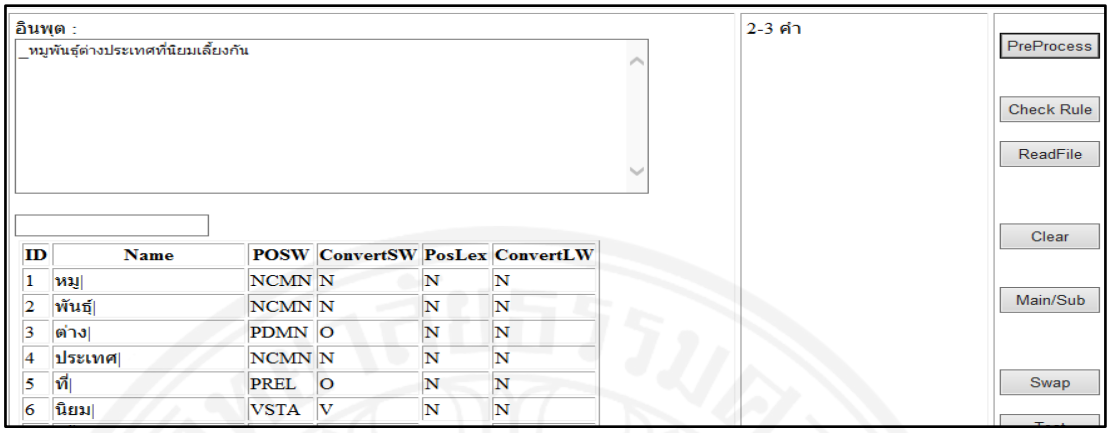
เป็นการเตรียมข้อมูลก่อนทำการสกัดคำนามประสมออกมาจากประโยค โดยการตัดประโยคออกจากบทความทีละหนึ่งบรรทัด จากนั้นจึงตรวจสอบ Part of Speech (POSW = POS ที่ได้มาจากโปรแกรม swath, PosLex = POS ที่ได้มาจากฐานข้อมูล Lexitron) ของคำแต่ละคำที่อยู่ภายในประโยคดังภาพที่ 4.1 เพื่อเป็นการเตรียมข้อมูลก่อนที่จะนำเข้าสู่ระบบ แล้วจึงตรวจสอบคำเชื่อมภายในประโยค เพื่อจะเป็นส่วนช่วยในการแบ่งกลุ่มคำออกจากกัน จากนั้นจึงแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น โดยอ้างอิงจากความสัมพันธ์ของคำด้วยวิธีการจับคู่ความสัมพันธ์ของคำดังภาพที่ 4.2

4.1.1 เครื่องมือที่ใช้สำหรับการทดสอบระบบ

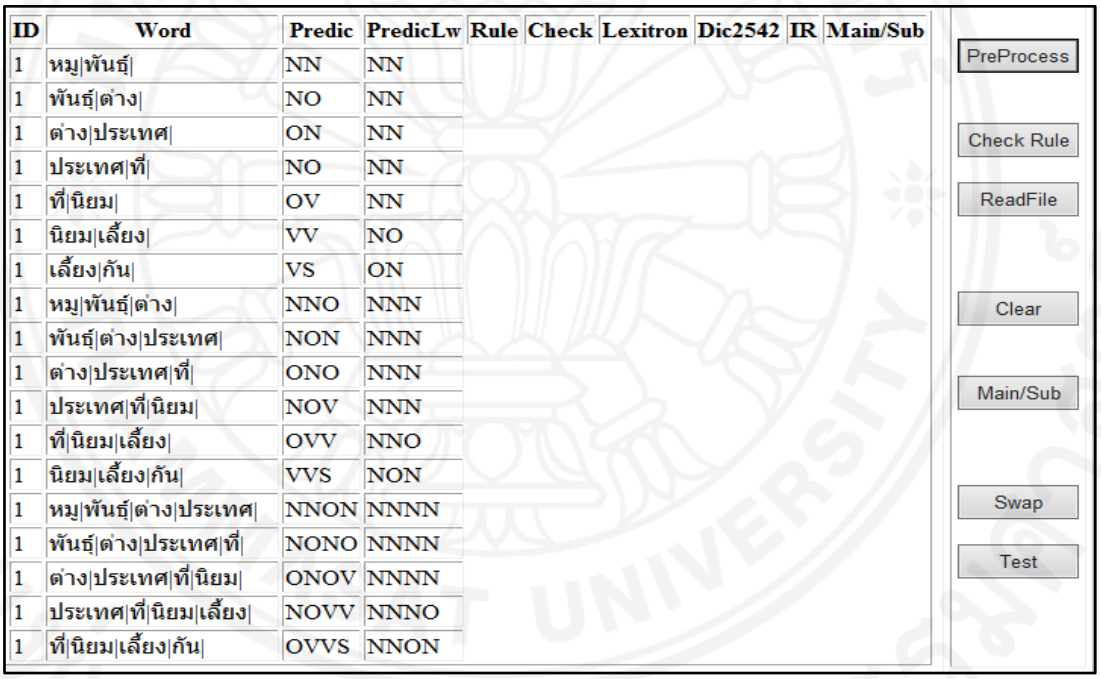
1. คัดประโยค 100 ประโยคออกมากจากคลังข้อความภาษาไทย BEST Corpus Training Set 1 Release 3 ของ NECTEC ซึ่งประกอบด้วยคำ 1013 คำจากประโยคทั้งหมด 100 บรรทัด

2. เตรียมพจนานุกรม Lexitron ซึ่งประกอบด้วย คำศัพท์ทั้งหมด 81,708 คำ เพื่อใช้เป็นพจนานุกรมที่ใช้สำหรับตรวจสอบคำนามประสม

3. เตรียมพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2542 ประกอบด้วยคำศัพท์ 37,000 คำ เพื่อใช้เป็นพจนานุกรมที่ใช้สำหรับตรวจสอบคำนามประสม



ภาพที่ 4.1 การแบ่งกลุ่มคำออกจากกัน



ภาพที่ 4.2 การจับคู่ความสัมพันธ์ของคำ

4.2 การทดลอง

4.2.1 Permanent Compound Noun Analysis (การตรวจสอบคำนามประสม)

เป็นการนำคำที่ได้จากการจับคู่ความสัมพันธ์ของคำมาผ่านกฎของคำนามประสม ทั้ง 13 กฎ จากนั้นจึงนำมาตรวจสอบกับพจนานุกรม เพื่อเป็นขั้นตอนแรกในการสกัดคำนามประสมออกจากประโยคซึ่งประกอบด้วยขั้นตอนต่างๆ ดังนี้

1. Pattern-based Compound Noun Analysis (การจับคู่ความสัมพันธ์ของคำ) จากการทดลองจะใช้ประโยค หมูพันธุ์ต่างประเทศที่นิยมเลี้ยงกัน มาใช้ในการทดลอง โดยจะแบ่งการแตกคำแต่ละคำที่อยู่ภายในประโยคออกมาทีละลำดับชั้น โดยอ้างอิงจากความสัมพันธ์ของคำด้วยวิธีการจับคู่ความสัมพันธ์ของคำตั้งแต่ 2-4 Gram ดังภาพที่ 4.3

ID	Word
1	หมู|พันธุ์|
1	พันธุ์|ต่าง|
1	ต่าง|ประเทศ|
1	ประเทศ|ที่|
1	ที่|นิยม|
1	นิยม|เลี้ยง|
1	เลี้ยง|กัน|
1	หมู|พันธุ์|ต่าง|
1	พันธุ์|ต่าง|ประเทศ|
1	ต่าง|ประเทศ|ที่|
1	ประเทศ|ที่|นิยม|
1	ที่|นิยม|เลี้ยง|
1	นิยม|เลี้ยง|กัน|
1	หมู|พันธุ์|ต่าง|ประเทศ|
1	พันธุ์|ต่าง|ประเทศ|ที่|
1	ต่าง|ประเทศ|ที่|นิยม|
1	ประเทศ|ที่|นิยม|เลี้ยง|
1	ที่|นิยม|เลี้ยง|กัน|

ภาพที่ 4.3 ตัวอย่างการจับคู่ความสัมพันธ์ของคำ

2. Thai Linguistic Rules (ตรวจสอบกฎคำนามประสม) เป็นการนำคำที่ได้จากการจับคู่ความสัมพันธ์ของคำมาตรวจสอบกับกฎคำนามประสมทั้ง 13 กฎ ซึ่งจะได้ผลลัพธ์ดังภาพที่ 4.4 ตัวอย่างการตรวจสอบกฎคำนามประสม (Predit = ส่วน POS ของ swath ตรวจสอบกับกฎคำนามประสม, PredicLw = ส่วน POS ของ Lexitron ตรวจสอบกับกฎคำนามประสม)

ID	Word	Predic	PredicLw	F
1	หมู|พันธุ์|	NN	NN	
1	พันธุ์|ต่าง|	NO	NN	
1	ต่าง|ประเทศ|	ON	NN	
1	ประเทศ|ที่|	NO	NN	
1	ที่|นิยม|	OV	NN	
1	นิยม|เลี้ยง|	VV	NO	
1	เลี้ยง|กัน|	VS	ON	
1	หมู|พันธุ์|ต่าง|	NNO	NNN	
1	พันธุ์|ต่าง|ประเทศ|	NON	NNN	
1	ต่าง|ประเทศ|ที่|	ONO	NNN	
1	ประเทศ|ที่|นิยม|	NOV	NNN	
1	ที่|นิยม|เลี้ยง|	OVV	NNO	
1	นิยม|เลี้ยง|กัน|	VVS	NON	
1	หมู|พันธุ์|ต่าง|ประเทศ|	NNON	NNNN	
1	พันธุ์|ต่าง|ประเทศ|ที่|	NONO	NNNN	
1	ต่าง|ประเทศ|ที่|นิยม|	ONOV	NNNN	
1	ประเทศ|ที่|นิยม|เลี้ยง|	NOVV	NNNO	
1	ที่|นิยม|เลี้ยง|กัน|	OVVS	NNON	

ภาพที่ 4.4 ตัวอย่างการตรวจสอบกฎคำนามประสม

3. Permanent Compound Noun Analysis เมื่อผ่านกฎคำนามประสมแล้ว ขั้นตอนต่อไปจึงนำมาตรวจสอบคำพจนานุกรมทั้งสองฉบับดังภาพที่ 4.5 ตัวอย่างการตรวจสอบกับพจนานุกรม (Check_Lexitron = ตรวจสอบกลุ่มคำกับ Lexitron, Check_RadChaBudit = ตรวจสอบกลุ่มคำกับพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2542)

id	Word	Swatch	Lexitron	PredicSw	PredicLw	Check_Lexitron	Check_RadChaBudit	Score
1	หมู|พันธุ์|	T	T	NN	NN	0	0	50
2	พันธุ์|ต่าง|	N	T	NO	NN	0	0	50
3	ต่าง|ประเทศ|	N	T	ON	NN	1	0	100
4	ประเทศ|ที่|	N	T	NO	NN	0	0	50
5	ที่|นิยม|	N	T	OV	NN	0	0	50
6	นิยม|เลี้ยง|	T	N	VV	NO	0	0	50
7	หมู|พันธุ์|ต่าง|	N	T	NNO	NNN	0	0	50
8	พันธุ์|ต่าง|ประเทศ|	N	T	NON	NNN	0	0	50

ภาพที่ 4.5 ตัวอย่างการตรวจสอบกับพจนานุกรม

จาก 3 ขั้นตอนข้างต้นจะสามารถสรุปได้ว่า ต่างประเทศ เป็นคำนามประสม เนื่องจากพบในพจนานุกรม Lexitron และทำหน้าที่เป็นนาม ส่วนกลุ่มคำอื่นๆ ที่มี Score (คะแนน) ไม่ใช่ 100 จะต้องนำไปพิจารณาในส่วนการพิจารณาคำนามประสมเป็นลำดับถัดไป

4.2.2 Semantic-based Compound Noun Analysis (การพิจารณาคำนามประสม)

กรณีที่ไม่พบคำนามประสมในพจนานุกรม ต้องพิจารณาความแตกต่างระหว่างคำนามประสมและนามวลี และประโยคประกอบด้วย 3 ขั้นตอนดังนี้

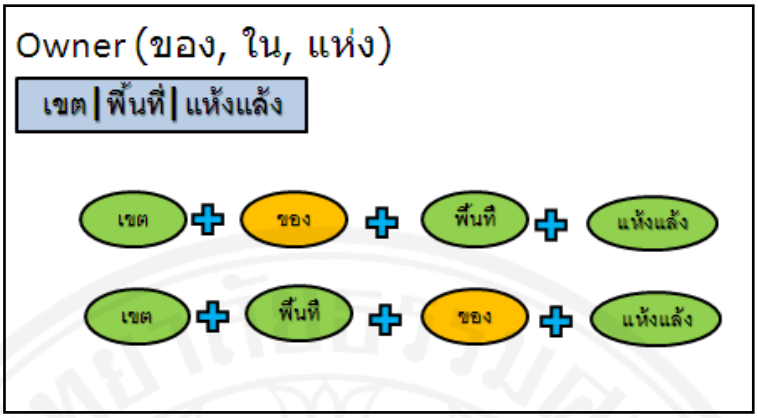
ประโยคที่ใช้ในการทดสอบระบบ คือ ภูมิอากาศในเขตพื้นที่แห้งแล้งของโลกจะมีหมูเป็นจำนวนน้อย โดยประโยคข้างต้นจะต้องผ่านการตรวจสอบคำนามประสมมาก่อนถึงจะนำส่วนกลุ่มคำที่เหลือมาเข้าสู่ขั้นตอนการพิจารณาคำนามประสมดังภาพที่ 4.6 โดยจะยกตัวอย่างการพิจารณาในส่วนของกลุ่มคำที่มีแถบสีเขียว เขต|พื้นที่|แห้งแล้ง|

id	Word	Swatch	Lexitron	PredicSw	PredicLw	Check_Lexitron	Check_RadChaBudit	IR	Score
1	ภูมิอากาศ|ใน|	T	N	NB	NO	0	0	1	0
2	เขต|พื้นที่|	T	T	NN	NN	1	0	1	100
3	พื้นที่|แห้งแล้ง|	T	T	NV	NV	0	0	1	0
4	แห้งแล้ง|ของ|	N	T	VB	VN	0	0	1	0
5	ของ|โลก|	N	T	BN	NN	0	0	1	0
6	โลก|จะ|	T	N	NV	NO	0	0	1	0
7	จะ|มี|	T	N	VV	OV	0	0	13	50
8	มี|หมู|	T	T	VN	VN	0	0	3	50
9	หมู|เป็น|	T	T	NV	NS	0	0	17	50
10	เป็น|จำนวน|	T	T	VN	SN	0	0	7	50
11	จำนวน|น้อย|	T	T	NS	NS	0	0	3	50
12	เขต|พื้นที่|แห้งแล้ง|	T	T	NNV	NNV	0	0	1	0
13	พื้นที่|แห้งแล้ง|ของ|	N	T	NVB	NVN	0	0	1	0
14	หมู|เป็น|จำนวน|	T	N	NVN	NSN	0	0	3	50

ภาพที่ 4.6 ตัวอย่างคำที่ใช้ในการพิจาณา

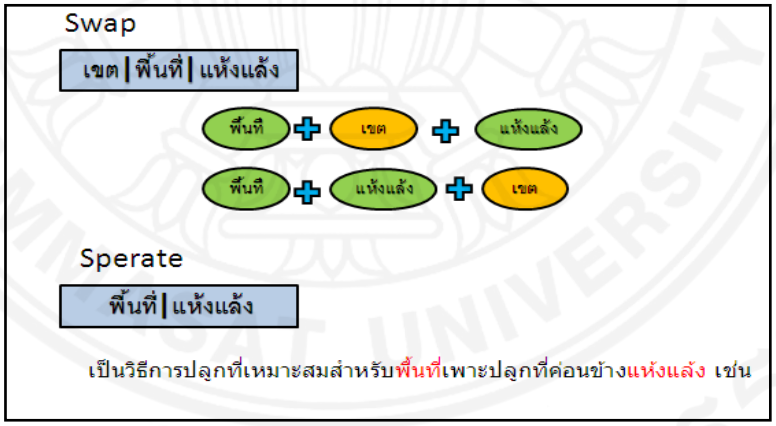
1. superordinate-subordinate ขั้นตอนการตรวจสอบจะใช้ความหมายของ subordinate เป็นตัวตั้งจากนั้นจึงนำ superordinate เข้าไปค้นหาในความหมายของ subordinate ว่ามี superordinate อยู่ในความหมายของ subordinate นั้นหรือไม่ ถ้าพบว่าเป็นส่วนหนึ่งของ subordinate และ ทั้ง superordinate และ subordinate มี Part of Speech เป็นคำนามจะถือว่าเป็นคำนามประสม จากกลุ่มคำที่นำมาทดสอบจะไม่พิจารณา superordinate-subordinate เนื่องจากคำที่นำมาทดสอบไม่เข้าเงื่อนไขของ superordinate-subordinate (สามารถดูคำอธิบายเพิ่มเติมได้จากบทที่ 3)

2. อรรถศาสตร์ เป็นการอ้างอิงเรื่องโครงสร้างการแสดงความเป็นเจ้าของ (Owner) โดยนำคำที่ต้องการค้นหาเติมคำว่า “ของ” ลงไปจากนั้นจึงนำไปค้นในคลังข้อมูล ถ้ามีความเป็นเจ้าของจะถูกคัดออก จากกลุ่มคำที่นำมาทดสอบสามารถพิจารณาลักษณะทางอรรถศาสตร์ได้ดังภาพที่ 4.7



ภาพที่ 4.7 โครงสร้างการแสดงความเป็นเจ้าของ

3. ลำดับการเรียงคำ เป็นการนำคำที่ต้องการค้นหามาทดลองสลับที่หรือแยกคำ จากนั้นจึงนำไปค้นหาในคลังข้อมูลเพื่อใช้ในการตรวจสอบ จากกลุ่มคำที่นำมาทดสอบสามารถพิจารณาเรื่องของลำดับการเรียงคำได้ดังภาพที่ 4.8 ลำดับการเรียงคำ (Swap = การสลับที่, Separate = การแยกคำ) กรณีที่พบว่าสามารถสลับที่หรือแยกคำได้จะถูกคัดออก



ภาพที่ 4.8 ลำดับการเรียงคำ

จาก 3 ขั้นตอนข้างต้นจึงสามารถสรุปได้ว่า เขตพื้นที่แห้งแล้ง เป็นคำนามประสมเนื่องจากผ่านเงื่อนไขของการพิจารณาคำนามประสม

4.2.3 ตารางสรุปผลการทดลอง

จากการทดสอบประโยค 100 ประโยคจากคำทั้งหมด 1,013 คำสามารถสรุปออกมาเป็นได้ดังตารางที่ 4.1 โดยสามารถอธิบายได้ดังนี้

4.2.3.1 Permanent Compound Noun Analysis

1. Table1 เป็นส่วนตรวจสอบกับพจนานุกรม

2. Permanent เป็นส่วนของการตรวจสอบกับฐานข้อมูลความถาวรที่ได้มาจากการเรียนรู้ในการทดสอบระบบ

4.2.3.2 Semantic-based Compound Noun Analysis

1. sup-sub เป็นส่วนการพิจารณา superordinate-subordinate

2. Check3 เป็นส่วนของการพิจารณาลักษณะทางอรรถศาสตร์ และลำดับการเรียงคำ

ตารางที่ 4.1

ตารางสรุปผลการทดลอง

วิธีการ	ที่แยกได้	ถูกต้อง	แยกไม่ถูกต้อง	แยกได้เพิ่ม	ตรวจสอบ
Table1	57	41	12	4	130
sup-sub	11	3	5	3	130
Permanent	12	12	0	0	130
Check 3	94	16	51	27	130
รวม	174	72	68	34	130
Recall	55.38				
Precision	41.38				

4.3 วิธีวัดผลการทดลอง

4.3.1 ตรวจสอบโดยผู้เชี่ยวชาญ

รายชื่อผู้เชี่ยวชาญ

1. คุณ นิศาชล ลันวงษา ปริญญาตรี วิทยาศาสตรสิ่งแวดล้อม และประกาศนียบัตร อาชีพ ครู

2. คุณ ณัฐพล พันดวง ปริญญาตรี ครุศาสตรเอกภาษาไทย อาชีพ ครู

3. คุณ จุฑามาศ นกน้อย ปริญญาตรี ครุศาสตรบัณฑิต อาชีพ ครู

1. Permanent Compound Noun Analysis ส่วนการตรวจสอบคำนามประสมใช้ผู้เชี่ยวชาญจำนวนสามท่าน มาช่วยตรวจสอบคำนามประสมที่ได้ออกมาจากระบบ ซึ่งพบว่าระบบสามารถตรวจสอบคำนามประสมได้ถูกต้อง 60 คำ

2. Semantic-based Compound Noun Analysis ส่วนการพิจารณาคำนามประสมใช้ผู้เชี่ยวชาญจำนวนสามท่าน มาช่วยตรวจสอบคำนามประสมที่ได้ออกมาจากระบบ ซึ่งพบว่าระบบสามารถตรวจสอบคำนามประสมได้ถูกต้อง 49 คำ

4.3.2 การวัดผลการทดลอง

การวัดผลความถูกต้องของการคัดแยกคำนามประสมออกจากประโยคจะใช้ค่าความครบถ้วนและค่าความแม่นยำในการวัดผลการทดสอบระบบโดยได้ผลลัพธ์ดังนี้

คำที่ถูกต้อง 72

คำที่ใช้ในการทดสอบทั้งหมด 130

คำที่พบทั้งหมด 174

ค่าความครบถ้วน (Recall)

72/130 * 100 = 55.38 %

ค่าความแม่นยำ (Precision)

72/174*100 = 41.38 %

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการทดลอง

จากผลการทดลองจากประโยค 100 ประโยค จากขั้นตอนการตรวจสอบจากคำนามประสมและการพิจารณาคำนามประสมระบบสามารถจับกลุ่มคำได้มากสุด 4 คำ โดยอ้างอิงตามโครงสร้างคำนามประสมของงานวิจัยฉบับนี้ ซึ่งจะจับกลุ่มคำแต่ละคำ เริ่มต้นตั้งแต่ 2 คำไปจน 4 คำ จากการทดลองสามารถตรวจสอบคำนามประสมที่มีความยาวมากสุดได้ถึง 4 คำ และสามารถสรุปผลการทดลองออกเป็นดังนี้

5.1.1 ส่วนที่ 1 ตรวจสอบคำนามประสม

สามารถแยกคำนามประสมออกมาจากประโยคได้ 57 คำและมีความถูกต้อง 41 คำ ไม่ถูกต้อง 12 คำและพบคำนามประสมเพิ่มขึ้นมาอีก 4 คำ

5.1.2 ส่วนที่2 ส่วนพิจารณาคำนามประสม

5.1.2.1 Superordinate-Subordinate

สามารถแยกคำนามประสมออกจากประโยคได้ 11 คำและมีความถูกต้อง 3 คำ ไม่ถูกต้อง 5 คำและพบคำนามประสมเพิ่มขึ้นมาอีก 3 คำ

5.1.2.2 Permanent

สามารถแยกคำนามประสมออกจากประโยคได้ 12 คำและมีความถูกต้อง 12 คำ

5.1.2.3 ตรวจสอบการสลับที่การแยกคำและการแสดงความเป็นเจ้าของ

สามารถคัดแยกคำนามประสมออกจากประโยคได้ 94 คำ มีความถูกต้อง 16 คำไม่ถูกต้อง 51 คำและพบคำนามประสมเพิ่มขึ้นมาอีก 27 คำ

5.1.3 การวัดผลการทดลอง

5.1.3.1 การวัดผลจากผู้เชี่ยวชาญ

จากการทดลองโดยใช้การวัดผลการทดลองจากผู้เชี่ยวชาญสามารถสกัดคำนามประสมออกมาได้จากประโยค 130 คำ ซึ่งพบในพจนานุกรม 74 คำ และไม่พบในพจนานุกรม 56 คำ

5.1.3.2 การวัดผลจากการทดลอง

จากการทดลองโดยใช้การสกัดคำนามประสมออกมาจากผู้เชี่ยวชาญ สามารถคัดคำนามประสมออกมาจากประโยค 174 คำ และ ตรงกับการวัดผลจากผู้เชี่ยวชาญจำนวน 72 คำ และตรงกับคำในพจนุกรม 41 คำ และได้เพิ่มมากกว่าผู้เชี่ยวชาญ 4 คำ

5.2 การอภิปรายผลการทดลอง

ภาษาไทยเป็นภาษาที่มีความซับซ้อนทั้งทางด้านไวยากรณ์ ความหมาย รวมถึงลำดับการเรียงคำซึ่งภาษาไทยจะเขียนติดต่อกันเป็นสายอักขระ โดยไม่มีเครื่องหมายเว้นวรรค ตอนแสดงการแบ่งคำเหมือนภาษาอังกฤษ ทำให้เป็นความยากในการศึกษาและพัฒนาในเรื่องของการที่จะทำให้คอมพิวเตอร์สามารถคำนวณและวิเคราะห์คำแต่ละคำที่ประกอบอยู่ในประโยคได้ ซึ่งจากการทดลองสามารถสรุปปัญหาที่ได้มาจากการทดลองได้ดังนี้

5.2.1 ความผิดพลาดที่เป็นปัญหาก่อนการประมวลผลขั้นต้น

1. ปัญหาจากการแบ่งคำ เนื่องจากความซับซ้อนของคำในภาษาไทย ซึ่งภายในประโยคของภาษาไทย ประกอบด้วยหน่วยย่อย ภายในประโยคที่ประกอบด้วยคำหลายประเภท ทำให้ในบางกรณี การหาโครงสร้างของคำนามประสม ค่อนข้างมีความซับซ้อนและเกิดความคลุมเครือในการกำหนดขอบเขตของคำดังภาพที่ 5.1 ซึ่งโปรแกรมแบ่งคำไม่สามารถแบ่งคำออกมาได้เป็น หยวก|กล้วย

id	Word	Swatch	Lexitron	PredicSw	PredicLw	Check_Lexitron	Check_RadChaBudit	IR	Score
7	หยวก|กล้วย|	T	N	NVN	OVN	1	0	4	100

ภาพที่ 5.1 ความคลุมเครือในการกำหนดขอบเขตของคำ

2. ปัญหาจากโปรแกรมแบ่งคำไม่สามารถแบ่งคำออกเป็นคำย่อยๆ ได้ ทำให้การคัดแยกคำนามประสมมีความผิดพลาดตามไปด้วยจากภาพที่ 5.2 ความต้องการของระบบ คือคำว่า สัตว์|เลี้ยงแต่โปรแกรมแบ่งคำมองคำว่าสัตว์เลี้ยงเป็นคำหนึ่งคำ

id	name	POSW	ConvertSW	POSLex	ConvertLW
1	หมู	NCMN	N	N	N
2	เป็น	VSTA	V	ADJ	S
3	สัตว์เลี้ยง	NCMN	N	N	N
4	ที่	PREL	O	N	N

ภาพที่ 5.2 ไม่สามารถแบ่งคำออกเป็นคำย่อยๆได้

5.2.2 ลักษณะปัญหาต่างๆที่เกิดขึ้นในการทดลอง

1. แบ่งคำนามประสมปะปนกับคำอื่น จากการทดลองพบว่าโปรแกรมแบ่งคำแบ่งคำจากประโยคโดยมีคำนามประสม สัตว์เลี้ยง ติดไปกับคำอื่นทำให้ไม่สามารถคัดแยกคำนามประสมออกจากประโยคได้ดังตารางที่ 5.1

ตารางที่ 5.1

การแบ่งคำจากประโยคโดยมีคำนามประสมติดไปกับคำอื่น

Word	Predic	Rule	Check	PredicLw	RSwatch	RLexi
หมูเป็น	NV			NS	T	T
เป็นสัตว์เลี้ยง	VN			SN	T	T
สัตว์เลี้ยงที่	NO			NN	F	T
ที่รู้จัก	OV			NV	F	T

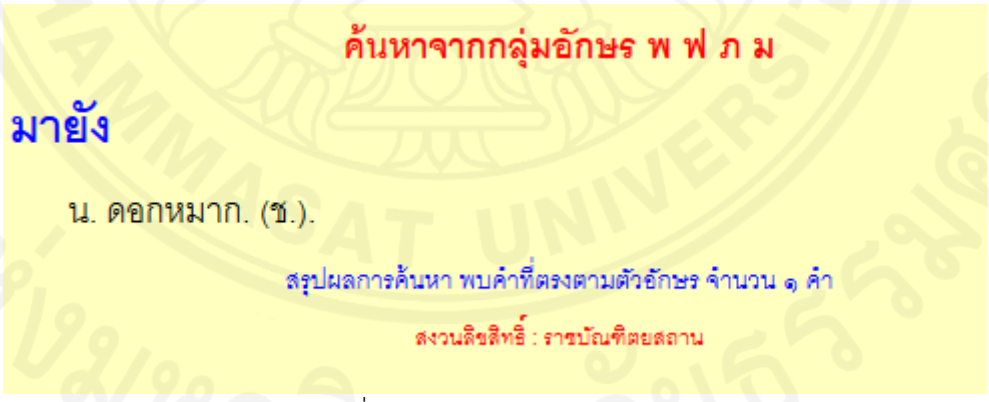
2. โปรแกรมแบ่งคำแบ่งคำไม่ถูกต้อง จากการทดลองพบว่าโปรแกรมแบ่งคำแบ่งคำจากประโยคได้ไม่ถูกต้อง ทำให้ไม่สามารถคัดแยกคำนามประสมออกจากประโยคได้ถูกต้องดังแสดงในตารางที่ 5.2 ซึ่งโปรแกรมแบ่งคำไม่สามารถแบ่งคำออกมาได้เป็น พันธุ์|ดูร็อก

ตารางที่ 5.2

แบ่งคำไม่ถูกต้อง

Word	Predic	Rule	Check	PredicLw	RSwatch	RLexi
พันธุ์ลาร์จไวต์	NN			NO	T	F
ลาร์จไวต์พันธุ์	NN			ON	T	F
พันธุ์ดู	NV			NV	T	T
ดูร็	VN			VO	T	F
ร็อก	NN			ON	T	F
พันธุ์ลาร์จไวต์พันธุ์	NNN			NON	T	F

3. คำที่ไม่ใช่คำนามประสม จากการทดลองพบว่ามีการพบคำที่ไม่ใช่คำนามประสมและเป็นไปตามเงื่อนไขของการตรวจสอบคำนามประสมและการพิจารณาคำนามประสม เนื่องจากภาษาไทยมีความซับซ้อนมาก ทำให้ในบางกรณีจึงไม่สามารถคัดแยกคำนามประสมออกจากประโยคได้หรือคัดแยกออกมาได้ถูกต้องดังภาพที่ 5.3



ภาพที่ 5.3 คัดแยกออกมาไม่ถูกต้อง

4. จากการทดลองเรื่องการสลับที่ การแยกคำ และความเป็นเจ้าของ ระบบคัดแยกออกมาได้ไม่ถูกต้องดังตารางที่ 5.3

ตารางที่ 5.3

สลับที่แบ่งคำไม่ถูกต้อง

สลับที่,แยกคำ,ความเป็นเจ้าของ		
คำที่ตรงกัน	ลำดับคำอื่น	คำอื่น
	16	ลูกผสม|ต่างๆ|
	17	ยัง|นำ|
	18	นำ|เอา|

5. การวิเคราะห์ superordinate-subordinate ยังมีความผิดบางอยู่ในบางกรณี เนื่องจากคำ superordinate-subordinate สามารถใช้ตรวจสอบได้ในบางกรณีและมีอีกหลายกรณีที่ยังเกิดความผิดพลาดซึ่งสืบเนื่องมาจากความซับซ้อนของภาษาไทยดังตารางที่ 5.4

ตารางที่ 5.4

Superordinate-Subordinate (sup-sub) แบ่งคำไม่ถูกต้อง

sup-sub		
id	Word	Sup-Sub
1	ประเทศ|เดนมาร์ก|	1
3	ของ|ซาก|	1
4	ที่|ดี|	1
9	รส|ดี|	1
11	หมู|ประเภท|	1

5.3 สรุปผลการวิจัย

จากงานวิจัยฉบับนี้ทำให้สามารถ เรียนรู้วิธีการพัฒนาอัลกอริทึมเพื่อใช้สำหรับในเรื่องของการสกัดคำนามประสมออกจากประโยคในภาษาไทย ซึ่งจากผลการทดลองของการพิจารณาคำนามประสม ทำให้สามารถสกัดคำประสมออกมาจากประโยคได้อยู่นอกเหนือจากส่วนที่อยู่ในพจนานุกรมและได้คำประสมชนิดอื่นออกมารวมอยู่ด้วย จากผลลัพธ์ดังกล่าวทำให้สามารถนำไปใช้เป็นแนวทางในการเรียนรู้ เพื่อศึกษาในเรื่องของการสกัดคำประสมออกจากประโยคชนิดอื่นๆ ได้ ในอนาคตและควรจะหาวิธีการที่เหมาะสม เพื่อมาปรับปรุงในเรื่องของการแบ่งคำและการคัดแยกคำ เนื่องจากโปรแกรมการสกัดคำนามประสมออกจากประโยคยังไม่สามารถคัดแยกคำออกมาได้ถูกต้องตามลักษณะปัญหาต่างๆ ที่อธิบายในส่วนข้างต้น การนำไปพัฒนาต่อจึงจำเป็นต้องให้ความสำคัญในเรื่องของการแบ่งคำเพื่อเป็นส่วนเสริมทำให้ระบบมีประสิทธิภาพเพิ่มมากขึ้น

รายการอ้างอิง

หนังสือและบทความในหนังสือ

อัญชลี สิงห์น้อย, คำนามประสม : ศาสตร์และศิลป์ในการสร้างคำไทย, (กรุงเทพมหานคร : สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2548)

สุนันท์ อัญชลีนุกูล, ระบบคำภาษาไทย, (กรุงเทพมหานคร : โครงการเผยแพร่ผลงานทางวิชาการ, คณะอักษรศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย, 2546)

วิทยานิพนธ์

Downing and Levi (2520). *On the creation and use of English compound nouns Language.*

ณัฐกานต์ เพ่งผล (2545). *การวิเคราะห์นามวลีภาษาโดยอ้างอิงข้อมูลสถิติและข้อสนเทศทางภาษา. วิศวกรรมสาสตร์มหาบัณฑิต* (วิศวกรรมคอมพิวเตอร์). มหาวิทยาลัยเกษตรศาสตร์.

C. Hansakunbutheung, A. Thangthai, C. Wutiwiwatchai and R.Siricharoenchai (2548). *Learning Methods and Features for Corpus Based*. NECTEC, Thailand.

Kanyanut Kriengket, KritKo Sawat, Sunant Anchaleenukul (2550). *A Computation Linguistics Study of Compound Nouns in Thai*. NECTEC, Thailand.

กัญญาณัฐ เกรียงเกตุ (2550). *การศึกษาคำนามแสดงอุปกรณ์ด้านวิทยาศาสตร์แนวภาษาศาสตร์คอมพิวเตอร์.* อักษรศาสตรมหาบัณฑิต. จุฬาลงกรณ์มหาวิทยาลัย.

ณรงค์กรรณ รอดทรัพย์ (2554). *โครงสร้าง และ วากยสัมพันธ์ของนามวลีภาษาไทยในป้ายรณรงค์หาเสียงเลือกตั้งสมาชิกสภาผู้แทนราษฎร.* วารสารปาริชาต (ฉบับพิเศษ), ผลงานวิจัยจากการประชุมวิชาการ ครั้งที่ 22.

สื่ออิเล็กทรอนิกส์

Edufind, “Parts of speech of Compound nouns,” Accessed October 28, 2014, http://www.edufind.com/english-grammar/compound-nouns/

โรงเรียนมหิดลวิทยานุสรณ์, “เอกสารประกอบการเรียน รายวิชา ท๔๐๑๐๕ หลักภาษาไทยในชีวิตประจำวัน,” สืบค้นเมื่อวันที่ 28 ตุลาคม 2557, http://www.mwit.ac.th/~saktong/learn6/79.pdf

Trueplookpanya, “หลักภาษาไทย,” สืบค้นเมื่อวันที่ 28 ตุลาคม 2557, http://www.trueplookpanya.com/new/cms_detail/knowledge/1673-00/

Guessing the Meanings of Compound Words, “คำนามประสม” สืบค้นเมื่อวันที่ 28 ตุลาคม 2557, https://compoundwords4.wordpress.com/2012/01/02/คำนามประสม-compound-nouns/

ภาคผนวก

ภาคผนวก ก
ประวัติผู้เชี่ยวชาญคนที่ 1

ชื่อ	นางสาวนิศาชล ลันวงษา
วันเดือนปีเกิด	วันที่ 27 ธันวาคม พ.ศ. 2529
วุฒิการศึกษา	ปีการศึกษา: 2551 วิทยาศาสตรสิ่งแวดล้อม และประกาศนียบัตรอาชีพครู มหาวิทยาลัยศิลปากร
อาชีพ	ครู ที่ โรงเรียนจันทร์กระจ่าง

ภาคผนวก ข
ประวัติผู้เชี่ยวชาญคนที่ 2

ชื่อ	นายณัฐพล พันดวง
วันเดือนปีเกิด	วันที่ 19 พฤศจิกายน พ.ศ. 2531
วุฒิการศึกษา	ปีการศึกษา: 2557 ครุศาสตรเอกภาษาไทย มหาวิทยาลัยรามคำแหง
อาชีพ	ครู ที่ โรงเรียนจันทร์กระจ่าง

ภาคผนวก ค
ประวัติผู้เชี่ยวชาญคนที่ 3

ชื่อ	นางสาวจุฑามาศ นกน้อย
วันเดือนปีเกิด	วันที่ 7 ธันวาคม พ.ศ. 2531
วุฒิการศึกษา	ปีการศึกษา: 2555 ครุศาสตรบัณฑิต มหาวิทยาลัยจุฬาลงกรณ์
อาชีพ	ครูอาสา ที่ วัดพุทธานุสรณ์

ประวัติผู้เขียน

ชื่อ	นางสาวณิชชา บำรุง
วันเดือนปีเกิด	วันที่ 10 เมษายน พ.ศ. 2530
วุฒิการศึกษา	ปีการศึกษา: 2551 เทคโนโลยีบัณฑิต (เทคโนโลยีสารสนเทศ) มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปราจีนบุรี
ตำแหน่ง	Senior Programmer บริษัท กรุงไทยธุรกิจลีสซิ่ง จำกัด

ผลงานทางวิชาการ

ณิชชา บำรุง, และ รัชฎา คงคะจันทร์. (ธันวาคม 2558). ระบบอัตโนมัติสำหรับสกัดคำนามประสมในประโยคภาษาไทย. The Eleventh International Symposium on Natural Language Processing (SNLP 2016), พระนครศรีอยุธยา.

ประสบการณ์ทำงาน	2557-ปัจจุบัน : Senior Programmer บริษัท กรุงไทยธุรกิจลีสซิ่ง จำกัด 2552-2556 : Programmer บริษัท สุพรีมโพรดักส์ จำกัด