การศึกษาการแยกนัยความหมายของ *หัว* ในภาษาไทย โดยใช้วิธีการวิเคราะห์ความหมายแอบแฝง*

นัชชา ถิระสาโรช**
วิโรจน์ อรุณมานะกุล***

บทคัดย่อ

ในภาษามีคำจำนวนมากเป็นคำหลายความหมาย สำหรับมนุษย์แล้ว คำหลายความหมายไม่ได้เป็นอุปสรรคในการสื่อสารเนื่องจากผู้ส่งสารและผู้รับสารยังคงสามารถเข้าใจความหมายได้ตรงกัน แต่สำหรับคอมพิวเตอร์แล้ว การสอนให้คอมพิวเตอร์รู้จักความหมายของคำ รวมถึงรู้ว่าควรเลือกใช้ความหมายใดจากความหมายทั้งหมดของคำหลายความหมายเมื่ออยู่ในบริบทต่างๆ นั้นยังเป็นปัญหาอยู่และยังเป็นเรื่องที่มีการศึกษากันมาอย่างต่อเนื่องในภาษาต่างๆ รวมถึงภาษาไทย สำหรับบทความนี้ได้ศึกษาคำว่า *หัว* โดยใช้วิธีการวิเคราะห์ความหมายแอบแฝง โดยใช้คำบริบทตำแหน่งต่างๆ ในการช่วยแยกความหมาย ผลการศึกษาพบว่า คำบริบทที่อยู่ติดกับคำเป้าหมายและมีกรอบหน้าต่างหรือระยะห่างไม่มากช่วยให้ระบบแยกความหมายได้ดีกว่าคำบริบท

---

* บทความนี้เป็นส่วนหนึ่งของวิทยานิพนธ์ระดับดุษฎีบัณฑิตเรื่อง “การศึกษาการแยกนัยความหมายของคำในภาษาไทยโดยใช้วิธีการวิเคราะห์ความหมายแอบแฝง” ของผู้วิจัยชื่อแรก.

** นิสิตระดับดุษฎีบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ติดต่อได้ที่: pernn39@gmail.com

*** รองศาสตราจารย์ประจำภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ติดต่อได้ที่: awirote@gmail.com

ที่ใช้กรอบหน้าต่างมาก และคำบริบททางซ้ายช่วยให้ระบบแยกความหมายได้ดีกว่าบริบททางขวา เนื่องจากคำบริบททางซ้ายที่เป็นตัวช่วยบ่งชี้ความหมายมักปรากฏร่วมกับความหมายใดความหมายหนึ่ง ในขณะที่คำบริบททางขวาบางคำจะไปปรากฏร่วมกับความหมายอื่นด้วย รวมถึงบริบททางขวามีการจับคู่คำที่เป็นคำกับช่องว่างมากกว่าบริบททางซ้ายจึงส่งผลให้ประสิทธิภาพของระบบเมื่อใช้บริบททางขวาลดลง นอกจากนี้ การใช้แยกนัยความหมายในที่นี้ยังได้ผลไม่ดีนัก (ถูกต้อง 41.63%) สาเหตุน่าจะมาจากใช้เพียงรูปคำอย่างเดียวและจำนวนตัวอย่างที่ใช้ก็ไม่มากนัก

คำสำคัญ: คำหลายความหมาย; การแยกนัยความหมาย; การวิเคราะห์ความหมายแอบแฝง; ภาษาศาสตร์คอมพิวเตอร์

A Study of Word Sense Discrimination of /*Hua4*/ in Thai Using Latent Semantic Analysis*

Nutcha Tirasaroj**
Wirote Aroonmanakun***

Abstract

In language, a number of words are polysemous. For humans, polysemy is not a problem in communication as the sender and the receiver are able to understand the same meaning of multi-meaning words. However, teaching a computer to know the senses of a word and choose the appropriate meaning when the word is in different contexts is still a problem. The purpose of this paper was to study word sense discrimination of /hua4/ using Latent Semantic Analysis. Contexts are the clues that help discriminating the senses in this study. The results show that contexts in a small window size tend to help discriminate senses more than those in a large window size. Furthermore, the systems using left contexts are better than those using right contexts because many word clues that help

---

* This article is a part of the first author's dissertation title "A Study of Word Sense Discrimination in Thai Using Latent Semantic Analysis"

** Ph. D.'s student, Department of Linguistics, Faculty of Arts, Chulalongkorn University, e-mail: pernn39@gmail.com

*** Associate Professor, Department of Linguistics, Faculty of Arts, Chulalongkorn University, e-mail: awirote@gmail.com

indicate the senses of a word on the left are found in only one meaning while some word clues on the right are found in several meanings. Moreover, when word pairs of the form word-space and vice versa of right contexts are much more than those of left contexts, the efficiency of the systems is decreased. The result of word sense discrimination in this study is not high (accuracy 41.63%). This could result of the use of only word forms and the small quantity of training data.

Keywords: polysemy; word sense discrimination; Latent Semantic Analysis; Computational Linguistics

1. บทนำ

ปรากฏการณ์ที่คำหนึ่งคำสามารถมีได้หลายความหมายเป็นปรากฏการณ์ทางธรรมชาติของภาษาที่พบได้ทั่วไปในภาษาต่างๆ ทั่วโลก ซึ่งสำหรับมนุษย์แล้วแม้คำจำนวนมากจะมีหลายความหมาย แต่ก็ไม่ได้เป็นอุปสรรคในการสื่อสารระหว่างกันมากนัก ผู้ส่งสารและผู้รับสารยังคงสามารถเข้าใจความหมายได้ตรงกันโดยอาศัยคำบริบทรอบข้างของคำนั้นๆ เป็นสิ่งที่ช่วยให้รู้ว่าคำนั้นควรมีความหมายว่าอะไร เช่น คำว่า *ติด* ในบริบทดังต่อไปนี้ ในที่สุดลูกคนหนึ่งก็*ติดยา* กับ เราสามารถ*ติดโรค*จากสัตว์ได้ จากตัวอย่าง *ติด* ในสองบริบทนี้มีความหมายต่างกัน โดยรู้ได้จากคำบริบทในตัวอย่างแรก คำว่า *ยา* ทำให้รู้ว่า *ติด* นี้หมายถึง “ชอบอย่างขาดไม่ได้” และในตัวอย่างที่สอง ติด หมายถึง “ได้รับเชื้อโรค” โดยใช้คำบริบทคือ *โรคจากสัตว์* เป็นต้น แต่คำหลายความหมายนี้จะเป็นปัญหากับงานด้านการแปลหรือการทำพจนานุกรม (Ravin and Leacock, 2006) ในงานด้านการแปล หากนักแปลไม่เข้าใจความหมายของคำอย่างถ่องแท้แล้ว ว่าเมื่ออยู่ในบริบทใดควรแปลว่าอย่างไร ก็อาจทำให้เลือกแปลความหมายผิดได้ ส่วนการทำพจนานุกรมนั้น การจะแยกความหมายของคำหลายความหมายว่าควรมีกี่ความหมาย หรือการจัดกลุ่มความหมายว่าความหมายหนึ่งควรเป็นความหมายย่อยของอีกความหมายหนึ่งหรือไม่นั้น แม้แต่พจนานุกรมภาษาเดียวกันแต่ต่างเล่มกันยังให้ความหมายของคำไม่เท่ากัน หรือให้ความหมายต่างกัน ทั้งนี้ขึ้นอยู่กับหลักเกณฑ์ของผู้ทำพจนานุกรมเล่มนั้นๆ ว่ามีเกณฑ์การแยกความหมายอย่างไร (Ravin and Leacock, 2006) นอกจากนี้ ในการทำพจนานุกรมยังจำเป็นต้องอาศัยตัวอย่างข้อมูลการใช้ภาษาจริงจำนวนมากเพื่อนำมาใช้วิเคราะห์หาความหมาย ดังนั้นระยะเวลาในการทำพจนานุกรมเล่มหนึ่งๆ จึงใช้เวลาค่อนข้างมาก

ปัจจุบันเทคโนโลยีได้เข้ามามีบทบาทสำคัญในการช่วยแบ่งเบาการทำงานของมนุษย์แทบจะทุกภาคส่วน เช่น การใช้เครื่องจักรในโรงงานการผลิตต่างๆ เป็นต้น ในด้านภาษาก็เช่นเดียวกัน มีการสอนให้คอมพิวเตอร์เรียนรู้ภาษามนุษย์เพื่อนำมาช่วยงานต่างๆ เช่น การค้นหาข้อมูล การรู้จำคำสั่งเสียง รวมถึงการสอนให้คอมพิวเตอร์เรียนรู้การแยกความหมายของคำ ซึ่งหากคอมพิวเตอร์สามารถแยกความหมายของ

คำได้จะเป็นประโยชน์อย่างมากในงานด้านต่างๆ เช่น การแปลเอกสาร รวมถึงการทำพจนานุกรม เพราะจะสามารถช่วยแยกข้อมูลเบื้องต้นที่จะนำมาใช้ในการวิเคราะห์ออกเป็นแต่ละกลุ่มความหมายได้

ในภาษาอังกฤษ ความสนใจเรื่องความหมายของคำในทางคอมพิวเตอร์มีมานานแล้วตั้งแต่ช่วงปลายยุค ค.ศ.1940 (Agirre and Edmons, 2007) และมีการศึกษากันมาอย่างกว้างขวางทำให้มีทรัพยากรต่างๆ มากมายที่สามารถนำมาใช้ช่วยวิเคราะห์ความหมายของคำ เช่น รายการความหมายของคำ คลังข้อมูลที่มีการกำกับความหมายให้กับคำหรือกำกับข้อมูลอื่นๆ ให้กับคำ เช่น ข้อมูลโครงข่ายคำ (WordNet) เป็นต้น แต่สำหรับภาษาไทยจัดว่าเป็นภาษาที่ขาดแคลนทรัพยากร เพราะปัจจุบันคลังข้อมูลที่เปิดให้บุคคลทั่วไปใช้ได้ เช่น คลังข้อมูลภาษาไทยแห่งชาติฯ (Thai National Corpus: TNC) ยังไม่มีข้อมูลทางภาษาใดๆ เช่น หน้าที่ของคำ ความหมายของคำ เป็นต้น กำกับไว้ แม้ในภาษาไทยจะมีการศึกษาเรื่องความหมายของคำมาบ้างแล้วและอาจจะมีคลังข้อมูลที่ได้กำกับข้อมูลทางภาษาไว้ แต่นั่นก็เป็นคลังข้อมูลที่มีไว้สำหรับศึกษากันภายในองค์กรเท่านั้น ดังนั้นเมื่อไม่มีข้อมูลทางภาษาใดๆ มาใช้ช่วยในการวิเคราะห์ความหมาย เทคนิควิธีหนึ่งที่จะนำมาใช้ช่วยในการหาความหมายของภาษาที่ขาดแคลนทรัพยากรในทางคอมพิวเตอร์ คือ การหาความหมายของคำแบบอัตโนมัติจากบริบทที่ปรากฏในข้อมูล

แนวคิดเรื่องการหาความหมายของคำจากบริบทนี้ตรงกับแนวคิดของนักภาษาศาสตร์โครงสร้าง เช่น เฟิร์ธ (Firth, 1957) กล่าวว่า *"You shall know a word by the company it keeps!"* กล่าวคือ เราสามารถเดาความหมายของคำที่เราไม่เคยรู้ความหมายมาก่อนได้จากคำที่ปรากฏรอบข้าง และคำที่มีความหมายคล้ายกันมักจะปรากฏอยู่ในบริบทที่คล้ายกัน ดังคำกล่าวของแฮริส (Harris, 1968) ซึ่งเป็นนักภาษาศาสตร์โครงสร้างอีกคน ที่ว่า *"Words will occur in similar contexts if and only if they have similar meanings*." จากแนวคิดดังกล่าวแสดงให้เห็นว่าบริบทสามารถใช้ช่วยแยกความหมายของคำได้ โดยวัตถุประสงค์ของบทความนี้คือ เพื่อศึกษาว่าเราจะสามารถใช้บริบทช่วยคอมพิวเตอร์ในการจำแนกความหมายต่างๆ ของคำได้ถูกต้องมากน้อยเพียงใด โดยคำที่ใช้ในการศึกษาในครั้งนี้ คือ คำนามคำว่า

หัว โดยผู้วิจัยตั้งสมมติฐานว่าบริบททางขวาจะมีส่วนช่วยในการแยกความหมายของคำนามได้ดีกว่าบริบททางซ้าย เนื่องจากในภาษาไทยคำขยายคำนามส่วนใหญ่มักอยู่ทางขวา เช่น คำนาม *หนังสือ* ในตัวอย่างประโยค *หนังสือ*เล่มนี้ดี กับ *หนังสือ*เล่มนี้สวย จะเห็นว่าสิ่งที่ใช้ช่วยแยกความหมายของหนังสือคือ คำวิเศษณ์ *สวย* และ *ดี* ทางด้านขวา โดยคำว่า *สวย* ช่วยให้รู้ว่า *หนังสือ* หมายถึงรูปลักษณะภายนอก ในขณะที่ *ดี* ช่วยให้รู้ว่า *หนังสือ* หมายถึง เนื้อหาในเล่ม เป็นต้น

สำหรับงานวิจัยนี้ ใช้วิธีการวิเคราะห์ความหมายแอบแฝง (Latent Semantic Analysis: LSA) ในการช่วยแยกความหมาย ซึ่ง LSA มีแนวทางหลักว่าให้สกัดเอาคุณสมบัติที่ซ่อนอยู่ในความหมายของคำออกมา โดยมีความเชื่อว่าคำที่มีความหมายคล้ายกันจะปรากฏอยู่ในเอกสารหรือบริบทที่มีลักษณะคล้ายกัน ที่ผ่านมามีการนำ LSA ไปใช้ศึกษาเรื่องความกำกวมทางความหมายของคำ เนื่องจาก LSA สามารถหาความหมายของคำและข้อความที่คล้ายคลึงกันได้ถูกต้องค่อนข้างใกล้เคียงกับคำตอบที่มนุษย์ให้มา (Landauer et al., 1998) ในภาษาไทย สุนีย์ พงษ์พินิจภิญโญ และวันชัย ริ้วไพบูลย์ (Pongpinigpinyo and Rivepiboon, 2005) ได้ใช้ Latent Semantic Indexing (LSI) ให้เครื่องเรียนรู้แบบไม่มีผู้สอนในการแยกความกำกวมทางความหมายของคำ เปรียบเทียบประสิทธิภาพของระบบทั้งภาษาไทยและภาษาอังกฤษ ซึ่งผลการวิจัยพบว่าค่าความถูกต้อง (F-measure) ของระบบอยู่ที่ประมาณ 70%

2. ระเบียบวิธีวิจัย

คำว่า *หัว* ที่จะนำมาศึกษานี้เป็นส่วนหนึ่งของงานวิจัยของผู้วิจัยที่ศึกษาเปรียบเทียบประสิทธิภาพของระบบระหว่างระบบที่ใช้แยกความหมายของคำนามกับคำกริยา และระบบที่ใช้แยกความหมายของคำที่มีความหมายน้อยกับคำที่มีความหมายมาก คำที่มีความหมายน้อยของผู้วิจัย คือ คำที่มีจำนวนความหมายน้อยกว่า 5 ความหมาย และคำที่มีความหมายมาก คือ คำที่มีจำนวนความหมายมากกว่า 7 ความหมาย ซึ่งจำนวนความหมายนี้อิงจากจำนวนความหมายในพจนานุกรมฉบับราชบัณฑิตยสถาน

พ.ศ.2554 (ราชบัณฑิตยสถาน, 2556) โดยคำที่ใช้ศึกษาทั้งหมดมี 4 คำด้วยกัน คือ คำนาม 2 คำ ได้แก่ *เสียง* และ *หัว* และ คำกริยา 2 คำ ได้แก่ *บอก* และ *ติด* ผู้วิจัยคัดเลือกมาจากรายการคำ 5,000 คำแรกที่ปรากฏอยู่ในคลังข้อมูลภาษาไทยแห่งชาติฯ และคัดเลือกเฉพาะคำที่มีชนิดของคำชนิดเดียว เนื่องจากความถี่ที่แสดงในรายการคำ 5,000 คำนี้เป็นการรวบรวมตามการปรากฏของรูปคำเท่านั้น ไม่ได้นำชนิดของคำมาร่วมพิจารณาด้วย ดังนั้นเพื่อเป็นการประหยัดเวลาในการวิเคราะห์คัดแยกข้อมูลกระจายไปตามแต่ละชนิดของคำ ผู้วิจัยจึงเลือกคำที่มีชนิดของคำชนิดเดียว เพื่อให้รู้ปริมาณข้อมูลที่แน่นอน นอกจากนี้ปริมาณความหมายของคำนามและคำกริยาก็ต้องใกล้เคียงกัน เนื่องจากในงานวิจัยมีการเปรียบเทียบประสิทธิภาพของระบบระหว่างคำนามและคำกริยา ดังนั้นเพื่อป้องกันไม่ให้ปัจจัยเรื่องจำนวนความหมายมีผลต่อประสิทธิภาพของระบบ ผู้วิจัยจึงเลือกคำที่มีปริมาณความหมายใกล้เคียงกันให้มากที่สุด สำหรับในบทความนี้ผู้วิจัยจะกล่าวถึงเฉพาะคำว่า *หัว* เท่านั้น ในหัวข้อนี้จะแบ่งออกเป็น 3 ส่วนด้วยกัน คือ วิธีการรวบรวมข้อมูล วิธีการวิเคราะห์ความหมายของคำ และวิธีการประมวลผลข้อมูล

2.1 วิธีการรวบรวมข้อมูล

ตัวอย่างข้อมูลที่นำใช้มาจากคลังข้อมูลภาษาไทยแห่งชาติฯ มีจำนวน 2,006 ตัวอย่าง ข้อมูลที่ดึงออกมาจะนำมาตัดคำโดยใช้โปรแกรมตัดคำ Thai Word Segmentation 2.1 (วิโรจน์ อรุณมานะกุล, 2545) เนื่องจากงานวิจัยนี้จะใช้บริบทในการแยกความหมายของคำ ดังนั้นข้อมูลที่ไม่มีบริบททางด้านซ้ายหรือด้านขวา หรือมีคำปรากฏร่วมทางด้านซ้ายหรือขวาน้อยกว่า 7 คำจะไม่นำมาใช้วิเคราะห์ เช่น

|คอ|มัน|จึง|ไม่|หัก|ลง|เสียก่อน|เพราะ|น้ำหนัก|*หัว*|และ|ปาก|----------
----------|*หัว*|ฉัน|เกือบ|ทิ่ม|กับ|กระจก|เป็น|ครั้ง|ที่|สอง|

นอกจากนี้ กรณีที่ *หัว* เป็นส่วนหนึ่งของคำอื่น เช่น เป็นส่วนหนึ่งของชื่อเฉพาะหรือเป็นคำที่มีความหมายเป็นสำนวน ก็จะไม่นำมาใช้ในการวิเคราะห์เช่นกัน เช่น

|เพื่อน|ที่|ออฟฟิศ|ของ|พี่|นั่น|ล่ะ|ตัวดี| |เรื่อง|ที่|ไป|*หัว*|หิน|ถึง|ได้|ถูก|เอา|มา|เม้าท์|ด้วย|

|จำต้อง|พึ่งพา|ผู้ใหญ่|หรือ|ผู้|มี|อำนาจ|อย่าง|*โง*|*หัว*|ไม่|ขึ้น| |เขา|มอง|ว่า|คน|ที่|จะ|ได้|ดี|ต้อง|มี|

จากตัวอย่างแรก *หัว* เป็นส่วนหนึ่งของชื่ออำเภอ *หัวหิน* และ *หัว* ในตัวอย่างที่สองเป็นส่วนหนึ่งของสำนวน *โงหัวไม่ขึ้น* ซึ่งในตัวอย่างนี้หมายถึง “เลิกไม่ได้” กล่าวคือต้องพึ่งพาผู้ใหญ่หรือผู้มีอำนาจไปตลอด

2.2 วิธีการวิเคราะห์ความหมายของคำ

สำหรับคำว่า *หัว* ในพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ.2554 (ราชบัณฑิตยสถาน, 2556) พบว่าคำนาม *หัว* มี 2 คำด้วยกัน เป็นคำพ้องรูป แต่ในงานวิจัยนี้จะรวมความหมายของทั้ง 2 คำไว้ด้วยกัน เนื่องจากในงานวิจัยนี้เป็นการศึกษาเรื่องการแยกความหมายเท่านั้น คือสนใจว่ารูปคำหนึ่งมีความหมายอะไรได้บ้างเพื่อนำไปใช้ในการประมวลผลภาษาต่อไป สำหรับคอมพิวเตอร์แล้ว เมื่อพบรูปคำหนึ่ง สิ่งที่ต้องสามารถทำได้คือวิเคราะห์ได้ว่ารูปคำในบริบทนั้นๆ มีความหมายอะไร โดยไม่สนใจว่าเป็นความหมายที่มาจากคำพ้องรูปหรือมาจากคำหลายความหมาย ในการศึกษาความหมายของคำ เนื่องจากความหมายของคำที่ใช้จริงอาจมีเพิ่มเติมแตกต่างจากความหมายที่ระบุไว้ในพจนานุกรมและเราไม่ทราบมาก่อนว่าความหมายทั้งหมดของคำนั้นๆ มีได้กี่ความหมาย ผู้วิจัยจึงจะดึงข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติฯ ออกมาคราวละ 200 ตัวอย่าง และวิเคราะห์ความหมายจนกว่าจะไม่เจอความหมายใหม่ในข้อมูลที่ดึงออกมา 2 ครั้งติดกัน ซึ่งเมื่อวิเคราะห์ความหมายจากตัวอย่างข้อมูลที่ดึงออกมา พบว่ามีบางความหมายในพจนานุกรมที่ผู้วิจัยไม่พบในตัวอย่างข้อมูล และพบว่ามีบางความหมายของ *หัว* ที่ไม่ได้เขียนไว้ในพจนานุกรม ดังแสดงไว้ในตารางที่ 1

ในการวิเคราะห์ความหมายนี้ ผู้วิจัยใช้ทฤษฎีคำหลายความหมายอย่างมีหลักการ (principled polysemy) ของวิเวียน อีแวนส์ และแอนเดรีย ไทเลอร์ (Vyvyan

Evans and Andrea Tyler, 2003) มาช่วยในการวิเคราะห์ ซึ่งเกณฑ์ที่ใช้ในการวิเคราะห์คำหลายความหมายของทฤษฎีนี้มี 3 เกณฑ์ด้วยกัน ได้แก่ เกณฑ์ทางความหมาย คือ ความหมายที่จะแยกออกไปอีกความหมายจะต้องมีความหมายใหม่เพิ่มขึ้นมา เกณฑ์ด้านการอธิบายมโนทัศน์ คือ ความหมายที่ต่างกันจะมีรูปแบบมโนทัศน์ที่ต่างกัน ซึ่งสามารถแสดงให้เห็นได้ในระดับภาษา เช่น มีคำปรากฏร่วมที่แตกต่างกัน เป็นต้น และเกณฑ์ทางไวยากรณ์ คือ ความหมายที่ต่างกันจะมีรูปแบบโครงสร้างทางไวยากรณ์ที่ต่างกันไปด้วย ในการนำเสนอผลการวิเคราะห์ ผู้วิจัยจะนำเสนอผลของความหมายที่ได้เลย เนื่องจากเป็นส่วนสำคัญที่จะนำไปใช้จัดกลุ่มข้อมูลและทดลองในขั้นต่อไป แต่อย่างไรก็ตาม ผู้วิจัยจะแสดงรายละเอียดวิธีการวิเคราะห์ความหมายของ หัว ตามแนวทฤษฎีนี้มา 1 ความหมายโดยเลือกความหมายใหม่ที่เพิ่มจากพจนานุกรม เพื่อให้ผู้ที่สนใจการวิเคราะห์ความหมายตามแนวทฤษฎีนี้ได้เข้าใจแนวทางในการวิเคราะห์

ความหมายของ หัว ที่วิเคราะห์ได้ในงานวิจัยนี้มีทั้งสิ้น 11 ความหมาย ซึ่งต่างจากงานของวิภารักษ์ กนกรัตนนุกูล (Kanokrattananukul, 2001) ที่วิเคราะห์ความหมายของคำว่า หัว ได้ทั้งหมด 20 ความหมาย ทั้งนี้อาจเนื่องมาจากคลังข้อมูลที่ใช้เป็นคนละคลังกัน วิภารักษ์ใช้คลังข้อมูลข่าวหนังสือพิมพ์กรุงเทพธุรกิจ และขั้นตอนการวิเคราะห์ความหมายก็ต่างจากผู้วิจัย โดยวิภารักษ์ได้ใช้วิธีการดึงข้อมูลออกมาตามจำนวนที่ต้องการก่อน แล้วค่อยมาวิเคราะห์ความหมาย ในขณะที่ผู้วิจัยดึงข้อมูลมาวิเคราะห์คราวละไม่มาก จำนวน 200 ตัวอย่าง จึงมีแนวโน้มที่จะไม่พบความหมายของ หัว ที่มีตัวอย่างข้อมูลน้อย เช่น ความหมายว่า “ด้านหนึ่งของเงินปลีกด้านหัวคู่กับด้านก้อย” ซึ่งในงานของวิภารักษ์พบข้อมูลเพียง 8 ตัวอย่างเท่านั้น ในงานของผู้วิจัยได้กำหนดขั้นตอนวิธีการวิเคราะห์ความหมายว่าจะวิเคราะห์ความหมายจนกว่าจะไม่เจอความหมายใหม่ในข้อมูลที่ดึงออกมา 2 ครั้งติดกัน ดังนั้นถึงแม้ภายหลังจะเจอความหมายนี้อยู่บ้างก็จะไม่นำมารวมไว้ด้วย

ตารางที่ 1

ความหมายของคำว่า หัว จากพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ.2554 จากงานวิจัยของวิภารักษ์ กนกรัตนนุกูล (Kanokrattananukul, 2001) และจากงานวิจัยนี้

ความหมาย	พจนานุกรม	Kanokrattananukul	งานวิจัยนี้
1. ส่วนบนสุดของร่างกายของคนหรือสัตว์	✓	✓	✓
2. ส่วนของพืชพันธุ์บางอย่างตอนที่อยู่ใต้ดิน	✓	✓	✓
3. ส่วนแห่งสิ่งของบางอย่างที่อยู่ข้างหน้า หรือข้างต้น หรือแรกเริ่ม เช่น หัวเรือ หัวถนน	✓	✓	✓
4. ส่วนที่อยู่ตรงข้ามกับหางหรือท้าย เช่น หัวแถว หัวเรือ	✓	✓	✓
5. ส่วนแห่งสิ่งของที่เป็นยอด เช่น หัวฝี, ส่วนแห่งสิ่งของที่ยื่นเด่นออกไป เช่น หัวแหลม หัวสะพาน	✓	✓	✓
6. ในการเล่นปั้นแปะหรือโยนหัวโยนก้อย เรียกสมมุติด้านหนึ่งของเงินปลีกว่า ด้านหัวคู่กับด้านก้อย	✓	✓	×
7. ส่วนที่เป็นแก่นสาร เช่น หัวยา หัวเหล้า	✓	✓	×
8. ส่วนเริ่มต้นที่เป็นวงของตัวหนังสือ	✓	×	×
9. ช่วงแรกเริ่มของเวลา เช่น หัวปี หัววัน หัวค่ำ	✓	✓	×
10. สติปัญญา, ความสามารถพิเศษ, ความคิดริเริ่ม เช่น เด็กคนนี้มีหัวทางดนตรี	✓	✓	✓
11. ผู้ที่มีความคิดหนักไปทางใดทางหนึ่ง เช่น หัวกฎหมาย	✓	✓	✓
12. ปัญญา, ความคิด เช่น หัวดี, หัวไว	✓	✓	✓
13. ผม	×	✓	✓
14. นามนัยแทนสิ่งที่กล่าวถึงทั้งหมด	×	✓	✓
15. ข้อความสำคัญของเรื่องหรือข่าว	×	✓	✓
16. ส่วนของอุปกรณ์หรือเครื่องมือ	×	✓	✓
17. อารมณ์	×	✓	×
18. ผู้นำ, หัวหน้า	×	✓	×
19. ส่วนต้นของสิ่งพิมพ์	×	✓	×
20. ชื่อ, ชื่อเรื่อง	×	✓	×

จากตารางที่ 1 ความหมายที่ 1-9 คือ ความหมายของ *หัว (1)* และความหมายที่ 10-12 คือ ความหมายของ *หัว (2)* ในพจนานุกรม ส่วนความหมายที่ 13-16 คือ ความหมายที่ได้จากการวิเคราะห์เพิ่มโดยผู้วิจัย และความหมายที่ 17-20 คือความหมายเพิ่มเติมที่พบในการศึกษาของวิภารักษ์ กนกรัตนนุกูล (Kanokrattananukul, 2001) สำหรับความหมายที่ 3 กับ 4 ของ *หัว (1)* ผู้วิจัยเห็นว่ามีความหมายที่มีความใกล้เคียง ซ้ำซ้อนกัน กล่าวคือ ความหมายที่ 3 ซึ่งหมายถึง “ส่วนแห่งสิ่งของบางอย่างที่อยู่ข้างหน้า หรือข้างต้น หรือแรกเริ่ม” ในความเป็นจริงส่วนที่อยู่ข้างหน้าข้างต้นก็ย่อมเป็นส่วนที่อยู่คนละด้านกับส่วนที่อยู่ด้านท้ายอยู่แล้ว เมื่อนำไปแยกเป็นอีกความหมายหนึ่งเป็นความหมายที่ 4 ว่า “ส่วนที่ตรงข้ามกับหางหรือท้าย” ก็ย่อมทำให้เกิดความหมายซ้ำซ้อนกัน ดังจะเห็นได้จาก ในพจนานุกรมให้ตัวอย่างของความหมายที่ 3 และ 4 เป็นคำๆ เดียวกัน คือ *หัวเรือ* ด้วยเหตุผลดังกล่าวผู้วิจัยจึงจะรวมความหมายที่ 3 และ 4 เข้าไว้ด้วยกันเป็นหนึ่งความหมาย ในส่วนของความหมายที่พบตามพจนานุกรม (ความหมายที่ 1-12) ที่พบในงานวิจัยนี้นั้น ผู้วิจัยจะไม่ได้แสดงรายละเอียดการวิเคราะห์ แต่จะแสดงเฉพาะความหมายใหม่ที่ผู้วิจัยวิเคราะห์เพิ่มขึ้นมา ได้แก่ ความหมายที่ 13-16 และจะแสดงวิธีการวิเคราะห์ความหมายโดยใช้ทฤษฎีคำหลายความหมายอย่างมีหลักการในความหมายที่ 13 ดังนี้

ความหมายที่ 13 จะพบว่ามีข้อมูลจำนวนหนึ่งที่ *หัว* ใช้ในความหมายว่า “ผม” เช่น

|มี|คน|อายุ|มาก|กว่า|ป้า|มา|ฟัง|อีก|เพราะ|ผู้ชาย|คน|นี้|เขา|*หัว*|ขาว|แล้ว|คือ|ผม|หงอก|ผม|ขาว|เกือบ|หมด|ทั้ง|*หัว*|

|ทีนี้|จะ|ทำ|ยังไง|ล่ะ| |เด็ก|*หัว*|จุก| |คน|เป็น|เจ้าของ|เรือ|ถาม|อย่าง|ขวาง|ๆ|

หากวิเคราะห์ตามทฤษฎีคำหลายความหมายอย่างมีหลักการแล้ว ด้วยเกณฑ์ทางความหมาย ความหมายนี้ขยายมาจากความหมายพื้นฐานที่ว่า “ส่วนบนสุดของร่างกายของคนหรือสัตว์” โดย *ผม* เป็นอวัยวะหนึ่งเป็นขนที่ขึ้นอยู่บนศีรษะ จะเห็นว่า *หัว* ในที่นี้มีลักษณะเป็นนามนัยที่เชื่อมโยงความหมายส่วนใหญ่-ส่วนย่อย (whole-part) กล่าวคือ *หัว* ไม่ได้หมายถึงส่วนบนของร่างกายมนุษย์ทั้งหมดแต่หมายถึงเฉพาะส่วนที่เป็นผมเท่านั้น

ส่วนของเกณฑ์ด้านการอธิบายมโนทัศน์และเกณฑ์ทางไวยากรณ์ แม้ว่าโครงสร้างนามวลีของ *หัว* ในความหมายนี้จะเหมือนกับโครงสร้างนามวลีของ *หัว* ในความหมายอื่นๆ คือ *หัว* + วิเศษณ์ เช่น *หัวดี หัวโต* แต่ด้วยมโนทัศน์ที่ต่างจากความหมายอื่น ทำให้คำปรากฏร่วมกับ *หัว* ในความหมายนี้จะต่างจากความหมายอื่น โดยคำวิเศษณ์ที่ปรากฏร่วมด้วยมักเป็นความหมายเกี่ยวกับสี เช่น ดำ แดง ทอง เป็นต้น หรือเป็นคำวิเศษณ์ที่แสดงลักษณะเป็นเส้นที่สามารถพันกันได้ เช่น ฟู ยุ่ง เกรียน เป็นต้น สถานการณ์ที่ผู้พูดใช้ *หัว* ในความหมายนี้มักใช้ในบริบทที่ไม่เป็นทางการและพบในภาษาพูดมากกว่าภาษาเขียน *หัว* ในความหมายนี้ สามารถใช้คำว่า *ผม* แทนที่ได้โดยไม่เปลี่ยนความหมาย เช่น ผู้ชายคนนี้เขา*ผม*ขาว หรือ เด็ก*ผม*จุก เป็นต้น

ความหมายที่ 14 *หัว* เป็นนามนัยหมายถึงตัวบุคคลนั้นๆ ซึ่งจะต่างจากความหมายที่ 13 โดยในความหมายนี้ *หัว* เป็นนามนัยที่เชื่อมโยงความหมายแบบส่วนย่อย-ส่วนใหญ่ (part-whole) ดังตัวอย่างเช่น

|ก็|พูด|เรื่อง|นี้|ขึ้น|มา| |หรือ|เขา|คิด|จะ|เฉด|*หัว*|ฉัน|เร็ว|ๆ| |นี้| |เพราะ|ทน|ไม่|ได้|

|อีก|ทั้ง|รายได้|ต่อ|*หัว*|ก็|ปรากฏ|ว่า|สูง|ขึ้น|เป็น|สอง|เท่า|

จากตัวอย่าง *หัว* ในที่นี้ไม่ตรงกับความหมายใดในความหมายที่ 1-12 แต่ *หัว* ในที่นี้หมายถึง ตัวบุคคล โดยใช้ *หัว* ซึ่งเป็นอวัยวะหนึ่งแทนส่วนอื่นๆ ที่เหลือทั้งหมด เช่น ตัวอย่างแรก “เฉด” หมายถึง การขับไล่ไสส่ง ในที่นี่ “เฉดหัว” ไม่ได้หมายถึงการขับไล่เฉพาะส่วนที่เป็นหัว แต่หมายถึงการขับไล่บุคคลนั้นๆ ให้ออกไป เป็นต้น

ในความหมายที่ 15 มักจะพบในตัวอย่างประโยคที่เกี่ยวกับสื่อสิ่งพิมพ์ต่างๆ เช่น

|ถ้อยคำ|บรรยาย|พาด|*หัว*|หนังสือพิมพ์|ทุก|วัน|ว่า| |บริษัท|นั้น| |ซื้อ|ตัว| |ผู้จัดการ|

|รายงาน|โดย|ต่าง|โปรย|*หัว*|ข่าว|ตรงกัน|ว่า| |กระทรวง|มหาด||ไทย|ของ|อังกฤษ|ห้าม|

จากตัวอย่างข้างต้น จะเห็นว่า *หัว* ในที่นี้ หมายถึง ส่วนสำคัญ ซึ่งหัวเรื่องของข่าวใช้ดึงดูดความสนใจของผู้อ่าน ความหมายนี้จะใกล้เคียงกับความหมายที่ 7 ของ *หัว* คือ “ส่วนที่เป็นแก่นสาร” แต่ผู้วิจัยแยกความหมายนี้ออกมา เนื่องจาก *หัว* ในความหมายนี้มีรูปแบบคำปรากฏร่วมที่ค่อนข้างเฉพาะ โดยมักปรากฏร่วมกับคำกริยา “พาด” “จั่ว” “โปรย” และกับคำนามที่มีความหมายเกี่ยวข้องกับสื่อสิ่งพิมพ์ต่างๆ เช่น ข่าว หนังสือพิมพ์ เรื่องราว เป็นต้น ดังจะเห็นได้จากตัวอย่างที่ยกมาข้างต้น

อีกความหมายหนึ่งที่พบของคำว่า *หัว* คือ ส่วนของอุปกรณ์หรือเครื่องมือ สาเหตุที่วิเคราะห์ให้มีความหมายนี้เพิ่มขึ้นมา เนื่องจากมีข้อมูลจำนวนหนึ่งที่ *หัว* จะตามด้วยคำกริยาแสดงอาการหรือการกระทำ (action verb) เช่น ฉีด อ่าน เผา ดูด เป็นต้น แสดงให้เห็นว่า *หัว* เป็นอุปกรณ์หรือเครื่องมือบางอย่างที่มีหน้าที่หรือการใช้งานตามคำกริยาที่ตามมา เช่น

|จะ|ต้อง|เตรียม|อะไหล่|สำรอง|ของ|*หัว*|เผา|ประเภท|นี้|ไว้|อยู่|ตลอด|เวลา|

|ลม|อัด|จะ|ระบาย|ออก|ไป|ทาง|*หัว*|ฉีด| |ทำให้|ความ|ดัน|ของ|ลม|อัด|ภายใน|ท่อ|ลด|ลง|

จากตัวอย่าง จะเห็นว่า *หัว* ใน “หัวเผา” และ “หัวฉีด” ต่างหมายถึงชิ้นส่วนของอุปกรณ์ที่ทำหน้าที่ใช้ “เผา” และใช้ “ฉีด” โดยปกติ *หัว* ถือเป็นอวัยวะส่วนสำคัญของมนุษย์และสัตว์เพราะเป็นส่วนที่ใช้สั่งการให้อวัยวะต่างๆ ทำงาน เมื่อนำคำว่า *หัว* มาใช้แทนชิ้นส่วนของอุปกรณ์ *หัว* จึงเปรียบเหมือนเป็นส่วนสำคัญของอุปกรณ์ที่จะบอกให้รู้ว่าเครื่องมือหรืออุปกรณ์นั้นๆ ใช้เพื่อทำอะไร เช่น “หัวเผา” ก็ทำให้รู้ว่าอุปกรณ์นี้ใช้สำหรับเผาอะไรบางอย่าง เป็นต้น

สรุปความหมายของคำว่า *หัว* และจำนวนตัวอย่างข้อมูลของแต่ละความหมายที่ใช้ในงานวิจัยนี้ แสดงไว้ในตารางที่ 2 ดังนี้

ตารางที่ 2

ความหมายของคำว่า หัว และจำนวนตัวอย่างที่ใช้ในงานวิจัยนี้

ความหมาย	จำนวนตัวอย่าง
1. ส่วนบนสุดของร่างกายของคนหรือสัตว์	999
2. ส่วนของพืชพันธุ์บางอย่างตอนที่อยู่ใต้ดิน	70
3. ส่วนแห่งสิ่งของบางอย่างที่อยู่ข้างหน้า หรือข้างต้น หรือแรกเริ่ม หรือส่วนที่อยู่ตรงข้ามกับหางหรือท้าย เช่น หัวเรือ หัวถนน หัวแถว	250
4. ส่วนแห่งสิ่งของที่เป็นยอด เช่น หัวฝี, ส่วนแห่งสิ่งของที่ยื่นเด่นออกไป เช่น หัวแหลม หัวสะพาน	158
5. สติปัญญา, ความสามารถพิเศษ, ความคิดริเริ่ม เช่น เด็กคนนี้มีหัวทางดนตรี	12
6. ผู้ที่มีความคิดหนักไปทางใดทางหนึ่ง เช่น หัวกฎหมาย	56
7. ปัญญา, ความคิด เช่น หัวดี, หัวไว	189
8. ผม	87
9. นามนัยแทนสิ่งที่กล่าวถึงทั้งหมด	106
10. ข้อความสำคัญของเรื่องหรือข่าว	10
11. ส่วนของอุปกรณ์หรือเครื่องมือ	69
รวม	2006

2.3 วิธีการประมวลผลข้อมูล

หลังจากนำข้อมูลไปผ่านโปรแกรมตัดคำแล้ว ข้อมูลทั้งหมดจะนำมาลบสัญลักษณ์ ตัวเลข เครื่องหมายวรรคตอนต่างๆ เช่น เครื่องหมายคำพูด วงเล็บ เป็นต้นออก เนื่องจากเครื่องหมายต่างๆ เหล่านี้จะไม่นำมาใช้พิจารณาเป็นคำปรากฏร่วมกับคำว่า *หัว* จากนั้นผู้วิจัยกำกับความหมายที่วิเคราะห์ได้จากหัวข้อ 2.2 ในข้อมูล เพื่อนำมาใช้ประเมินความถูกต้องของระบบหลังจากที่ระบบได้ประมวลผลแล้ว ตัวอย่างการกำกับข้อมูล เช่น

|ผม|ผู้|ซึ่ง|ต้อง|พลอย|เบรค|รถ|*หัว*| *1* คะมำ|ไป|ด้วย| |เพื่อ|ให้|ได้|มุม|ที่|รอดพ้น|

|ชีวาตม์|กับ|กิ่ง|คำ|นั่ง|ใกล้|บรูซ|กับ|คริส|ทาง|*หัว*| *3* โต๊ะ| |ไกล|จาก|คน| ทาง|ปลาย|โต๊ะ|พอสมควร|

การกำกับนี้ ผู้วิจัยจะกำกับเลขความหมายไว้ด้านหลังคำเป้าหมาย (target word) จากตัวอย่าง *หัว* ในตัวอย่างแรก ตรงกับความหมายที่ 1 ซึ่งหมายถึง ส่วนบนสุดของร่างกายของคนหรือสัตว์ และ *หัว* ในตัวอย่างที่สอง ตรงกับความหมายที่ 3 คือ ส่วนแห่งสิ่งของบางอย่างที่อยู่ข้างหน้า หรือข้างต้น หรือแรกเริ่ม หรือส่วนที่อยู่ตรงข้ามกับหางหรือท้าย

สำหรับวิธีการที่ผู้วิจัยจะใช้ในการประมวลผลนี้ คือ วิธีการวิเคราะห์ความหมายแอบแฝง (Latent Semantic Analysis: LSA) ของเดียร์เวสเตอร์ และคณะ (Deerwester et al., 1990) ซึ่ง LSA มีแนวทางหลักว่าให้สกัดเอาคุณสมบัติที่ซ่อน (latent) อยู่ในความหมายของคำออกมา ทั้งนี้โดยมีความเชื่อว่าคำที่มีความหมายคล้ายกันจะปรากฏอยู่ในเอกสารที่มีลักษณะคล้ายกัน ในที่นี้เอกสารก็เปรียบเป็นบริบทที่ความหมายคล้ายกันจะปรากฏอยู่ในบริบทที่คล้ายกัน จากแนวคิดของ LSA ผู้วิจัยจะดึงคำบริบทที่ปรากฏร่วมกับคำเป้าหมายมาประมวลผล โดยในการดึงคำบริบทนี้ จะดึงเอาคำบริบททางด้านซ้ายอย่างเดียว คำบริบททางด้านขวาอย่างเดียว และคำบริบททั้งด้านซ้ายและด้านขวาที่อยู่ติดกับคำเป้าหมาย โดยจะกำหนดกรอบหน้าต่างหรือระยะห่างของคำบริบทจากคำเป้าหมายว่าจะใช้กี่คำ เช่น การดึงคำบริบท 3 คำทางด้านซ้ายของคำเป้าหมาย

|ผม|ผู้|ซึ่ง|ต้อง|พลอย|เบรค|รถ|*หัว*| *1* คะมำ|ไป|ด้วย| |เพื่อ|ให้|ได้|มุม|ที่| รอดพ้น|

จากตัวอย่าง จะได้คำบริบทมา 3 คำ ได้แก่ พลอย เบรค และรถ ซึ่งจะดึงคำมาเช่นนี้จากทุกตัวอย่าง เมื่อได้คำบริบทมาทั้งหมดแล้ว ก็จะนำมาหาค่าความถี่ของการปรากฏร่วมกัน เช่น จากตัวอย่างข้างต้นจะนับความถี่ ได้เป็น พลอย-เบรค ปรากฏร่วมกัน 1 ครั้ง พลอย-รถ 1 ครั้ง และ เบรค-รถ 1 ครั้ง ดังแสดงไว้ในตารางที่ 3

ตารางที่ 3

การนับความถี่ของคำปรากฏร่วม

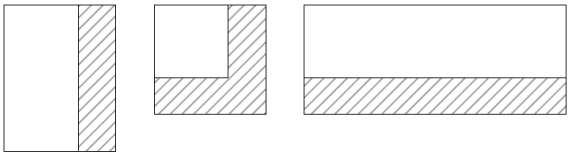
คำ	รถ	เบรค	พลอย
รถ	0	0	0
เบรค	1	0	0
พลอย	1	1	0

จากตัวอย่างตารางที่ 3 เมื่อได้ความถี่ของคำปรากฏร่วมทั้งหมด ค่าทั้งหมดจะแสดงออกมาในรูปของเมทริกซ์ของคำปรากฏร่วมที่สมมาตรกัน โดยแถวแรกและคอลัมน์แรกจะแสดงรายการคำปรากฏร่วมทั้งหมดที่พบในข้อมูล ซึ่งคำที่อยู่ในแถวและคอลัมน์จะเหมือนกันดังตารางที่ 3 และข้อมูลคือความถี่หรือจำนวนครั้งที่คำปรากฏร่วมกัน ซึ่งเมื่อนับการปรากฏร่วมกันทั้งหมดของคำต่าง ๆ แล้วจะได้เมทริกซ์ขนาดใหญ่ออกมา จากนั้นจะแปลงค่าความถี่เป็นค่าสถิติ log-likelihood[1] (Dunning, 1993) เนื่องจากจำนวนความถี่ที่นับมานั้นเป็นข้อมูลดิบจะยังไม่เห็นความสัมพันธ์ของคำที่ปรากฏร่วมกันมากนัก คำบางคำแม้ปรากฏร่วมกันน้อยครั้ง แต่มีความสัมพันธ์กันมาก เพราะปรากฏร่วมกันตลอด ในขณะที่บางคำแม้ปรากฏร่วมกันมาก แต่มีความสัมพันธ์กันน้อย เนื่องจากไปปรากฏร่วมกับคำอื่นมากเช่นกัน เช่น จากข้อมูล เบรค-รถ ปรากฏร่วมกัน 1 ครั้ง แปลงเป็นค่า log-likelihood ได้ 10.83 ในขณะที่ เริ่ม-ใช้ ปรากฏร่วมกัน 1 ครั้งเช่นกัน แต่ได้ค่า log-likelihood 3.41 ทั้งนี้เพราะ “เบรค” และ “รถ” ไม่ได้ปรากฏร่วมกับคำอื่นมากนัก ในขณะที่ “เริ่ม” กับ “ใช้” จะปรากฏร่วมกับคำอื่นด้วย เช่น เริ่ม-จาก เริ่ม-ทำ การ-ใช้ ที่-ใช้ เป็นต้น

เมื่อได้ค่าทั้งหมดแล้ว ก็จะนำไปลดขนาดของเมทริกซ์ลง ซึ่งการลดขนาดเมทริกซ์นี้ถือเป็นแนวคิดสำคัญของ LSA เพราะเป็นการตัดเอาส่วนที่ไม่จำเป็นออกไป เหลือเพียงคุณสมบัติที่สำคัญเท่านั้น การลดขนาดของเมทริกซ์ใช้วิธีการทางคณิตศาสตร์

[1] ค่าสถิติ Log-likelihood แสดงค่าความสัมพันธ์ของคำปรากฏร่วมว่ามีความสัมพันธ์มากน้อยเพียงใด คำนวณจากความถี่ของการปรากฏร่วมกันของคำ เทียบกับเมื่อคำนั้นไปปรากฏร่วมกับคำอื่น

Singular Value Decomposition[2] (SVD) ที่จะแยกองค์ประกอบของเมทริกซ์จากเมทริกซ์เดียวเป็น 3 เมทริกซ์ ได้แก่ เมทริกซ์ของแถวแนวนอน เมทริกซ์ของคอลัมน์แนวตั้ง และเมทริกซ์แนวทแยง ดังภาพที่ 1



ภาพที่ 1. ลักษณะการแยกองค์ประกอบของเมทริกซ์ด้วย SVD (Clarke, 2007)

จากภาพ ส่วนที่ใช้คือส่วนที่เป็นสีขาวเท่านั้น ดังนั้นขนาดของเมทริกซ์เมื่อผ่านขั้นตอนการแยกองค์ประกอบด้วย SVD แล้วขนาดของเมทริกซ์จะลดลง โดยส่วนที่ตัดออกนั้นถือว่าเป็นส่วนที่ไม่มีประโยชน์หรือเป็น noise ของข้อมูล จากนั้นค่าที่ได้จากการแยกองค์ประกอบนี้จะนำมารวมเข้าด้วยกัน ได้เป็นเมทริกซ์ที่มีค่าใหม่ ซึ่งค่าใหม่นี้ SVD ได้ไปคำนวณหาค่าจากการปรากฏร่วมกับคำอื่นๆ ว่ามีความสัมพันธ์กับคำอื่นอย่างไร ซึ่งลักษณะเช่นนี้คือการแสดงความสัมพันธ์ที่ซ่อนอยู่ออกมา จากนั้นผู้วิจัยหาค่ากลางของคำบริบทในแต่ละตัวอย่างโดยการนำค่าของคำบริบทมาบวกกัน เมื่อได้ผลรวมมาทั้งหมดแล้ว ก็จะมาหาค่าความคล้ายคลึงกันของแต่ละตัวอย่าง โดยใช้ K-means clustering algorithm[3] (Lloyd, 1982) โดยวิธีการนี้จะให้กำหนดจำนวนขึ้นมา K กลุ่ม ในที่นี้คือ K ความหมาย ระบบจะคำนวณหาจุดศูนย์กลางของแต่ละกลุ่ม แล้วจากนั้นจะจัดให้ค่าของตัวอย่างที่ใกล้จุดศูนย์กลางของกลุ่มไหนมากที่สุดเข้าอยู่

---

[2] Singular Value Decomposition (SVD) เป็นเทคนิคทางคณิตศาสตร์เพื่อลดขนาดของเมทริกซ์ โดยแยกองค์ประกอบของเมทริกซ์เป็น 3 เมทริกซ์ คือ เมทริกซ์แถว เมทริกซ์คอลัมน์ และเมทริกซ์แนวทแยง เพื่อเก็บค่าขนาดเมทริกซ์ จากนั้นลดขนาดของเมทริกซ์ทั้ง 3 ลง แล้วจึงค่อยวิเคราะห์ย้อนไปหาเมทริกซ์ดั้งเดิม โดยส่วนที่ตัดออกไปคือส่วนที่ไม่สำคัญของข้อมูล

[3] K-Means clustering algorithm เป็นวิธีการหนึ่งในการจัดกลุ่มข้อมูลโดยไม่มีผู้สอนที่ง่ายที่สุด

ในกลุ่มนั้น ในการศึกษาครั้งนี้ ผู้วิจัยได้กำหนดให้จัดกลุ่มข้อมูลตัวอย่างโดยให้ K เท่ากับ 11 ตามจำนวนความหมายที่วิเคราะห์ได้ จากนั้นจึงประเมินผลการจัดกลุ่มตัวอย่างว่าแต่ละตัวอย่างอยู่ในกลุ่มความหมายถูกต้องตามที่ผู้วิจัยได้วิเคราะห์ไว้หรือไม่

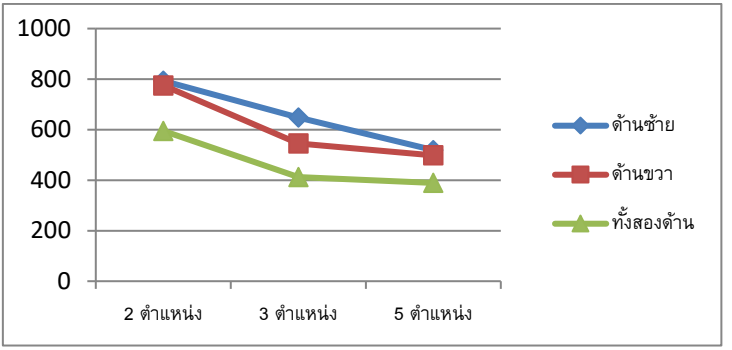
3. ผลการศึกษา

ประสิทธิภาพของระบบเมื่อใช้คำบริบทด้านซ้าย ด้านขวา และทั้งด้านซ้ายและขวาของ *หัว* ด้วยกรอบหน้าต่างหรือระยะห่างที่แตกต่างกันได้แสดงไว้ในตารางที่ 4 ดังนี้

ตารางที่ 4

ค่าความถูกต้องของระบบเมื่อใช้บริบทและกรอบหน้าต่างที่แตกต่างกัน

กรอบหน้าต่าง / บริบท	2 ตำแหน่ง		3 ตำแหน่ง		5 ตำแหน่ง	
	จำนวน	%	จำนวน	%	จำนวน	%
ด้านซ้าย	793	39.53	648	32.30	519	25.87
ด้านขวา	775	38.63	545	27.17	498	24.83
ด้านซ้ายและขวา	595	29.66	412	20.54	389	19.39



ภาพที่ 2. ค่าความถูกต้องของระบบเมื่อใช้บริบทและกรอบหน้าต่างที่แตกต่างกัน

จากตารางที่ 4 จะเห็นว่าเมื่อใช้กรอบหน้าต่างที่กว้างขึ้น ประสิทธิภาพของระบบในการแยกข้อมูลตัวอย่างออกเป็นแต่ละกลุ่มความหมายลดลง และบริบททางซ้ายสามารถช่วยแยกกลุ่มความหมายได้ดีกว่าบริบทอื่นๆ ในทุกๆ ตำแหน่ง ซึ่งไม่ตรงกับสมมติฐานที่ตั้งไว้ว่าบริบททางขวาน่าจะมีส่วนช่วยในการแยกความหมายของคำนามได้ดีกว่าบริบททางซ้าย เมื่อผู้วิจัยได้ทดสอบประสิทธิภาพของระบบโดยการใช้คำบริบททั้งสองด้านด้วยระยะห่าง 1 ตำแหน่งหรือหมายถึงคำที่อยู่ติดกับ *หัว* ทั้งด้านซ้ายและขวา ผลปรากฏว่า ช่วยให้ระบบสามารถแยกกลุ่มความหมายได้ดีกว่าแบบอื่น โดยสามารถแยกได้ถูกต้อง 835 ตัวอย่าง คิดเป็น 41.63% ผลที่ได้เป็นข้อพิสูจน์ว่าการใช้บริบทน้อยมีแนวโน้มจะช่วยให้ระบบสามารถแยกกลุ่มตัวอย่างได้ถูกต้องมากกว่าการใช้บริบทมาก และบริบทที่อยู่ติดกับคำเป้าหมายช่วยให้ระบบสามารถแยกกลุ่มความหมายได้ดีกว่าคำบริบทที่ห่างออกไป ดังจะเห็นได้จากประสิทธิภาพของระบบที่ใช้คำบริบท 1 ตำแหน่งด้านซ้ายและขวาให้ผลดีกว่าแบบอื่น ซึ่งหากเปรียบเทียบจากกราฟในภาพที่ 2 จะเห็นว่า เมื่อเป็นคำบริบทตั้งแต่ 2 ตำแหน่งขึ้นไป การใช้บริบททั้งสองด้านจะทำให้ประสิทธิภาพของระบบลดลงกว่าการใช้คำบริบทเพียงด้านซ้ายหรือด้านขวาเท่านั้น

สาเหตุที่เมื่อใช้กรอบหน้าต่างที่กว้างขึ้นแล้วทำให้ประสิทธิภาพของระบบลดลงนั้น เมื่อพิจารณาส่วนของคำปรากฏร่วมจะพบว่าเมื่อใช้คำมากขึ้น ความหลากหลายของคำก็จะมากขึ้นตามไปด้วย เมื่อเป็นเช่นนี้ระบบจะไม่เห็นความสัมพันธ์ของบริบทกับคำเป้าหมายชัดเจน เนื่องจากเมื่อมีคำมาก นั่นหมายถึงคู่คำปรากฏร่วมก็จะมีมากขึ้นเช่นกัน ส่งผลให้การกระจายตัวของข้อมูลสูง

เมื่อพิจารณาคำที่ปรากฏร่วมกับ *หัว* ในข้อมูลแล้ว พบว่า คำที่มักปรากฏร่วมกับ *หัว* ในความหมายใดความหมายหนึ่ง จะมีส่วนช่วยให้ระบบสามารถแยกความหมายนั้นออกจากความหมายอื่นได้ เช่น คำกริยา “คลุม” “ลูบ” “ส่าย” พบว่าจะปรากฏนำหน้าคำนาม *หัว* ในความหมายที่ 1 เพียงความหมายเดียว หรือคำนาม “เข่า” “เรือ” จะปรากฏต่อจาก *หัว* เสมอในความหมายที่ 3 และไม่พบปรากฏร่วมกับ *หัว* ในความหมายอื่น หรือหากมีปรากฏร่วมกับ *หัว* ในความหมายอื่น คำบริบทนั้นก็จะเกิดร่วมด้วยน้อยเมื่อเทียบสัดส่วนการปรากฏร่วมกับ *หัว* ในความหมายที่คำนั้น

บ่งชี้อยู่ เช่น คำบุพบท “ใน” พบว่าปรากฏนำหน้า *หัว* ทั้งหมด 75 ตัวอย่าง โดยปรากฏร่วมกับ *หัว* ในความหมายที่ 7 72 ตัวอย่าง และปรากฏร่วมกับ *หัว* ในความหมายอื่นเพียง 3 ตัวอย่างเท่านั้น เป็นต้น

จากผลการทดลองที่พบว่าบริบทด้านซ้ายทำให้ประสิทธิภาพของระบบดีกว่าการใช้บริบทด้านขวานั้น เมื่อเปรียบเทียบคำบริบทที่ปรากฏร่วมกับ *หัว* ทางด้านซ้ายและขวาแล้ว จะเห็นว่าคำบ่งชี้ที่จะสามารถช่วยแยกความหมายของ *หัว* ได้ส่วนใหญ่จะอยู่ทางด้านขวา เช่น *ผัก* ปรากฏตามหลัง *หัว* ในความหมายที่ 2 *เข่า มุม* ในความหมายที่ 4 *แดง* ในความหมายที่ 8 หรือ *ฉีด* ในความหมายที่ 11 เป็นต้น ในขณะที่คำที่ปรากฏร่วมกับ *หัว* ทางด้านซ้ายจะค่อนข้างกระจาย มีเพียงบางความหมายเท่านั้นที่เห็นได้ชัดเจน ดังเช่นความหมายที่ 1 และ 7 ที่ได้ยกตัวอย่างไปก่อนหน้านี้ แต่สาเหตุที่ทำให้บริบททางด้านซ้ายดีกว่านั้น เป็นเพราะจำนวนตัวอย่างข้อมูลที่ใช้ในความหมายที่ 1 มีมากกว่าความหมายอื่น โดยความหมายที่ 1 มีประมาณครึ่งหนึ่งของข้อมูล และเมื่อเรียงความถี่ของคำที่ปรากฏร่วมกับ *หัว* ในความหมายที่ 1 จะพบว่าคำบริบทที่ปรากฏทางซ้ายที่มีความถี่ในอันดับต้นๆ ของความหมายที่ 1 มักเป็นคำที่ปรากฏร่วมกับ *หัว* ในความหมายนี้เสมอ เช่น คลุม ลูบ ส่าย ท่วม เหนือ เป็นต้น ในขณะที่คำบริบททางด้านขวา มีบางคำที่จะไปปรากฏร่วมกับ *หัว* ในความหมายอื่นด้วย เช่น คำว่า *ของ* พบว่าปรากฏร่วมกับ *หัว* ในความหมายที่ 1 มี 35 ตัวอย่างจากจำนวนทั้งหมด 66 ตัวอย่าง เป็นต้น นอกจากนี้ช่องว่างหรือการเว้นวรรคที่ผู้วิจัยไม่ได้ตัดออกไปจากข้อมูลก็น่าจะเป็นปัจจัยหนึ่งที่มีผลต่อประสิทธิภาพของระบบเช่นกัน เนื่องจากเมื่อเรียงความถี่ของคำปรากฏร่วมกับ *หัว* จะพบว่า ช่องว่างมีความถี่มากที่สุดในทุกตำแหน่ง และบริบททางขวามีช่องว่างมากกว่าบริบททางซ้าย เมื่อเป็นเช่นนี้จึงมีการจับคู่คำบริบทที่เป็นคำกับช่องว่างจำนวนมาก ทำให้ระบบไม่เห็นความสัมพันธ์ระหว่างคำ และเมื่อปริมาณของช่องว่างของบริบททางด้านขวามีมากกว่าทางด้านซ้ายจึงส่งผลให้ระบบที่ใช้บริบททางด้านขวาเพียงอย่างเดียวมีประสิทธิภาพลดลงตามไปด้วย

แม้ผลการทดลองจะออกมาว่าบริบททางซ้ายจะช่วยแยกความหมายของ *หัว* ได้ดีกว่าบริบททางขวา นั่นไม่ได้หมายความว่าบริบททางซ้ายจะเหมาะสมมากกว่า

ในการช่วยแยกความหมายของคำนาม เนื่องจากในงานวิจัยนี้ใช้เพียงรูปคำเท่านั้นในการแยกความหมาย ดังนั้นสิ่งที่ระบบจะสามารถนำมาใช้ช่วยแยกความหมายได้จึงมีเพียงความถี่ของคำที่ปรากฏร่วมกันและความสัมพันธ์ระหว่างคำที่คำนวณจากความถี่ของคำที่ปรากฏร่วมกันเทียบกับเมื่อ 2 คำนั้นไปปรากฏร่วมกับคำอื่นเท่านั้น ซึ่งเมื่อวิเคราะห์ความหมายของ *หัว* จริงๆ แล้ว บริบททางขวายังคงมีบทบาทในการช่วยแยกความหมาย เพราะคำที่ช่วยบ่งชี้ความหมายส่วนใหญ่มักเป็นบริบททางขวา เช่น ในความหมายที่ 2 ที่ว่า “ส่วนของพืชพันธุ์บางอย่างตอนที่อยู่ใต้ดิน” คำด้านขวาของ *หัว* จะเป็นชื่อของพืช เช่น มัน เผือก เป็นต้น หรือความหมายว่า “ส่วนแห่งสิ่งของที่เป็นยอด” คำที่ปรากฏร่วมทางขวา ได้แก่ ฝี เข่า เป็นต้น

4. สรุปและอภิปรายผลการศึกษา

จากผลการศึกษาพบว่าบริบททางด้านซ้ายของข้อมูลช่วยให้ระบบสามารถแยกกลุ่มข้อมูลออกเป็นแต่ละกลุ่มความหมายได้ดีกว่าบริบททางขวา แม้ว่าคำปรากฏร่วมทางด้านขวาจะช่วยบ่งชี้ในการแยกความหมายได้ดีกว่าบริบททางซ้าย แต่ในความหมายที่ 1 ซึ่งมีมากเป็นครึ่งหนึ่งของตัวอย่างทั้งหมด กลับพบว่า มีคำที่มีความถี่สูงที่ปรากฏร่วมกับ *หัว* ทางด้านขวาบางคำจะไปปรากฏร่วมกับ *หัว* ในความหมายอื่นด้วย ในขณะที่บริบททางซ้ายจะเป็นคำที่มักปรากฏร่วมกับ *หัว* ในความหมายนี้เท่านั้น นอกจากนี้ ช่องว่างที่มีเป็นจำนวนมากทางด้านขวาก็น่าจะมีผลทำให้ประสิทธิภาพของระบบลดลงเช่นกัน

อย่างไรก็ตาม ในงานวิจัยนี้ใช้ปริมาณตัวอย่างน้อย เพียงแค่ 2,006 ตัวอย่างเท่านั้น และมีบางความหมายที่มีตัวอย่างเพียง 10-12 ความหมาย ทำให้ระบบมีตัวอย่างเรียนรู้น้อยเกินไป ไม่เห็นความแตกต่างระหว่างความหมายต่างๆ และในงานวิจัยนี้ใช้เพียงการพิจารณารูปคำเท่านั้น ไม่ได้ใช้ข้อมูลทางภาษาอื่น เช่น หน้าที่ของคำ ในการนำมาช่วยระบบแยกกลุ่มความหมาย จึงส่งผลให้ระบบไม่สามารถแยกกลุ่มความหมายได้ดีเท่าที่ควร เพราะคำบางคำสามารถเป็นได้หลายหน้าที่ เมื่อปรากฏ

ร่วมกับความหมายหนึ่งอาจเป็นหน้าที่หนึ่ง แต่เมื่อปรากฏร่วมกับอีกความหมายหนึ่งก็อาจจะเป็นหน้าที่อื่นได้แม้มีรูปคำเดียวกัน เมื่องานวิจัยนี้พิจารณาเพียงรูปคำเท่านั้นก็จะไม่เห็นความแตกต่างในเรื่องนี้ ดังนั้นสิ่งที่น่าจะศึกษาต่อไปคือ ทดลองใช้ข้อมูลที่มีการกำกับข้อมูลทางภาษา รวมถึงเพิ่มปริมาณตัวอย่างที่ใช้ให้มากขึ้น และทดลองลบส่วนที่เป็นช่องว่างหรือเว้นวรรคในข้อมูล เพื่อให้การจับคู่คำปรากฏร่วมเป็นคำกับคำทั้งหมด ให้ระบบได้เห็นความสัมพันธ์ของคำจริงๆ ซึ่งน่าจะมีผลให้ระบบสามารถแยกความหมายได้ดียิ่งขึ้น

รายการอ้างอิง

ราชบัณฑิตยสถาน. (2556). *พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ.2554*. กรุงเทพฯ: นานมีบุ๊คส์พับลิเคชั่นส์.

วิโรจน์ อรุณมานะกุล. (2545). *Thai Word Segmentation*. Retrieved from http://pioneer.chula.ac.th/~awirote/resources/thai-word-segmentation.html

Agirre, E. and Edmonds, P. (2007). Introduction. In Agirre, E. and Edmonds, P. (eds.). *Word Sense Disambiguation Algorithms and Application*.

Clarke, D. (2007). *Context-theoretic Semantics for Natural Language an Algebraic Framework* (Doctor of Philosophy's thesis). University of Sussex, Brighton.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41*(6), pp. 391-407.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1).

Evans, V. and Tyler, A. (2003). *Towards a Theory of Principled Polysemy: The Case of In*. ICLC 2003.

Firth, J.R. (1957). *Papers in Linguistics (1934-1951)*. London: Oxford University Press.

Harris, Z. (1968). *Mathematical Structures of Language*. New York; Krieger.

Kanokrattananukul, W. (2001). *Word Sense Disambiguation in Thai Using Decision List Collocation* (Master of Arts Degree Thesis, Linguistics) Chulalongkorn University, Bangkok.

Landauer, T.K., Laham, D., and Foltz, P. (1998). Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. In Report, M.I., Jordan, M.J., Kearns & S.A. Sollar (eds.). *Advances in Neural Information Processing Systems 10*. Cambridge: MIT Press.

Lloyd, S.P. (1982). Least squares quantization in PCM. In *IEEE Transactions on Information Theory, 28 (2)*: 129-137.

Pongpinigpinyo, S. and Rivepiboon, W. (2005). Distributional Semantic Approach to Thai Word Sense Disambiguation, In *International Journal of Computational Intelligence Vol. 2 No.3 2005*.

Ravin, Y. and Leacock, C. (2006). Polysemy: An Overview. In Ravin, Y. and Leacock, C. (eds.) *Polysemy: Theoretical and Computational Approaches*. New York: Oxford University Press.