การทดสอบมาตรวัดความคล้ายคลึงเชิงความหมาย สำหรับคำเหมือนและคำตรงข้ามภาษาไทย

รศ.ดร. พรฤดี เนติโสภากุล

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

Email: ponrudee@it.kmitl.ac.th

บทคัดย่อ

งานประยุกต์ของการประมวลผลข้อความมักจำเป็นต้องมีการวัดระยะห่างหรือความคล้ายคลึงกันของข้อความ ตัวอย่างเช่น การจัดกลุ่มประเภทข่าว การตรวจสอบการคัดลอกผลงานวิจัย เป็นต้น ในภาษาอังกฤษ มีการนำเสนอมาตรวัดความคล้ายคลึงเชิงความหมายระดับหน่วยคำหลายมาตรวัด ได้แก่ *Path Distance Similarity, Leacock Chodorow Similarity, Wu-Palmer Similarity, Resnik Similarity, Jiang-Conrath Similarity* และ *Lin Similarity* งานวิจัยนี้ มีวัตถุประสงค์เพื่อทดสอบว่า 1) สามารถประยุกต์ใช้มาตรวัดดังกล่าวกับการวัดความคล้ายคลึงกันในภาษาไทยหรือไม่ อย่างไร และ 2) หากประยุกต์ใช้ได้ วิธีการดังกล่าวจะให้ผลลัพธ์การวัดที่สมเหตุสมผลหรือไม่ การทดสอบเริ่มต้นจากการคัดเลือกคำไทยที่พบบ่อยจำนวนเริ่มต้น 99 คำ และคู่คำตรงข้ามภาษาไทยจำนวน 99 คู่ ในการตอบคำถามวิจัยที่หนึ่ง พบว่า ไม่สามารถเปรียบเทียบคำไทยสองคำโดยตรง แต่การเปรียบเทียบสามารถทำได้โดยทำการเชื่อมทรัพยากร *Thai wordnet* เข้ากับ *Princeton's Wordnet* ในการทดลองที่สอง ได้แบ่งชุดคำเป็นสองชุดคือ ชุดคำเหมือนและชุดคำตรงข้าม และหาค่าเฉลี่ยของมาตรวัดทั้งหก เพื่อทำการทดสอบทางสถิติแบบ *pair-wised t-test* สำหรับสมมุติฐานว่า ค่าความคล้ายคลึงในชุดคำเหมือนจะมากกว่าค่าความคล้ายคลึงในชุดคำตรงข้าม ผลการทดสอบพบว่า สามารถยอมรับสมมุติฐานดังกล่าวได้ที่ระดับนัยสำคัญ 0.05 หรือที่ความเชื่อมั่น 95% ดังนั้น จึงสรุปได้ว่า มาตรวัดเชิงความหมายสามารถประยุกต์ใช้วัดความคล้ายคลึงในคำภาษาไทยได้

คำสำคัญ – มาตรวัดความคล้ายคลึงข้อความ, ทรัพยากรคำศัพท์, การประมวลผลภาษาไทย

Abstract

Many of text processing applications usually need to measure text distance or text similarity. For example, text news clustering, research plagiarism checking. In English language, there are many semantic-based word similarity measurements. Those are Path Distance Similarity, Leacock Chodorow Similarity, Wu-Palmer Similarity, Resnik Similarity, Jiang-Conrath Similarity and Lin Similarity. This research's objective is to investigate whether: 1) Those measurements are applicable for Thai term

similarity measures or not and 2) In case they are applicable, the resultant evaluation values are valid or not. The investigation process begins with selection of 99 high frequent Thai terms and their 99 opposite terms. The answer to the first research question is that Thai word similarity can be measure only indirectly via Thai wordnet, by related terms to Princeton's wordnet. To answer the second research question, an experiment design has been conducted by constructing two sets of term-pairs: synonym term-pair set and antonym term-pair set. Average term-pair similarity values are obtained from those six semantic word similarity methods mentioned above. Using pair-wised t-test for hypothesis testing to test that average similarity value in synonym-pair is significantly greater than that in antonym-pair, we found that the hypothesis is true at the significant level 0.05 or at 95% confident interval. Therefore, we conclude that the semantic similarity measurements is applicable for Thai words.

Keyword– *Text Similarity Measurement; Lexical Resource, Thai Language Processing*

1. บทนำ

มาตรวัดความคล้ายคลึงกันของข้อความ (*Text Similarity Measurement)* เป็นงานย่อย (*Subtask)* ที่มีความสำคัญ ต่อการประมวลผลด้านต่างๆ ที่มีข้อความมาเกี่ยวข้องเป็นอย่างมาก ตัวอย่างงานที่จำเป็นต้องใช้มาตรวัดดังกล่าว เช่น (1) ใช้ในการปรับปรุงประสิทธิภาพของงานค้นหา (*Search Task)* หรืองานค้นคืนข่าวสาร (*Information Retrieval Task)* โดยการค้นหาข้อความที่มีความหมายเดียวกันหรือคล้ายคลึงกัน แต่ใช้คำหรือข้อความที่แตกต่าง [1] (2) ใช้ในการทำเหมืองข้อความ เช่น การจำแนกหมวดหมู่ หรือจัดกลุ่มประเภทข้อความจากแหล่งต่างๆ เช่น การจัดกลุ่มของข่าวสารออกเป็นข่าวเศรษฐกิจ ข่าวการเมือง ฯลฯ การจำแนกความคิดเห็นของลูกค้าจากสื่อสังคมออนไลน์ หรือการจำแนกข้อความที่แสดงความเห็นเชิงบวกออกจากข้อความที่แสดงความเห็นเชิงลบ หรือใช้ในงานสรุปย่อข้อความจากหลายเอกสาร [2] เป็นต้น (3) ใช้ในการตรวจสอบการคัดลอกวิทยานิพนธ์หรือผลงานวิจัย โดยตรวจสอบไปถึง ระดับการใช้คำในวลี เช่นคำเหมือนหรือคำทดแทนมาแทนคำในต้นฉบับ เป็นต้น นอกจากนี้ ยังมีงานอื่นๆ อีกจำนวนมากที่ต้องใช้มาตรวัดความคล้ายคลึงของข้อความเป็นงานย่อยในการทำงานให้ได้ผลลัพธ์ที่ดี [3]

แม้ว่ามาตรวัดดังกล่าวจะมีความสำคัญอย่างยิ่ง แต่การวัดระยะห่างของข้อความก็ไม่สามารถทำได้อย่างตรงไปตรงมา ดังเช่นการวัดระยะห่างระหว่างจุดสองจุด เนื่องจากการใช้คำในภาษามีลักษณะที่ยืดหยุ่น คำหนึ่งคำมีหลายความหมาย คำเหมือนและคำตรงข้ามก็อาจมีบริบทการใช้งานที่ทดแทนกันตรงๆ ไม่ได้ทีเดียว นอกจากนี้ มาตรวัดบางตัวที่มีอยู่ก็อาจจะไม่เหมาะสมกับภาษาอื่นนอกจากภาษาอังกฤษ

โดยเฉพาะในภาษาไทย จากการสำรวจไม่พบงานวิจัยเฉพาะด้านในการวัดระยะห่างข้อความภาษาไทย แต่พบแหล่งทรัพยากรภาษาไทยที่สำคัญคือ *Thai Wordnet* ซึ่งสร้างจาก *Princeton's Wordnet (PWN)* ซึ่งเป็นทรัพยากรคำศัพท์ภาษาอังกฤษ มีผู้วิจัยหลายท่านได้สร้างมาตรวัดความคล้ายคลึงระดับหน่วยคำจากทรัพยากร *PWN Wordnet* ดังกล่าว มาตรวัดเหล่านี้จัดเป็นมาตรวัดความคล้ายคลึงเชิงความหมาย (*Semantic Similarity)*

เนื่องจากยังไม่มีผู้ทดลองใช้มาตรวัดดังกล่าวกับภาษาไทย ดังนั้น งานวิจัยนี้จึงได้เลือกใช้มาตรวัดต่างๆ เหล่านี้ มาทำการทดลองกับคำภาษาไทย เพื่อทดสอบความเหมาะสมของการใช้มาตรวัดเชิงความหมายที่มี

ส่วนที่ *2* ของบทความนี้กล่าวถึง *Wordnet* และมาตรวัดความคล้ายคลึงเชิงความหมายที่เกี่ยวข้องจำนวนหกมาตรวัด ส่วนที่ *3* ระบุคำถามวิจัยและการออกแบบการทดสอบ ส่วนที่ *4* แสดงผลลัพธ์การทดสอบ วิเคราะห์และสรุปผลการทดสอบ และส่วนที่ *5* สรุปและเสนอแนะงานวิจัยในอนาคต

2. Wordnet กับ มาตรวัดความคล้ายคลึงเชิงความหมาย

มาตรวัดความคล้ายคลึงเชิงความหมาย จะพิจารณาความคล้ายคลึงกันของเทอม (*Term Similarity*) จากความหมาย มักจะคำนวณได้จากทรัพยากรคำศัพท์ที่มีโครงสร้างแบบอนุกรมวิธาน (*Taxonomy*) เช่น *Wordnet* โดยมีผู้นำเสนอไว้หลายมาตรวัด การที่จะเข้าใจมาตรวัดดังกล่าว ต้องทำความเข้าใจกับโครงสร้างทรัพยากรคำศัพท์ที่ใช้คือ *Wordnet* ก่อน

2.1 โครงสร้างของทรัพยากร Wordnet

Wordnet เป็นทรัพยากรคำศัพท์ที่สำคัญในการประมวลผลภาษา ซึ่งได้ถูกสร้างขึ้นตั้งแต่ปี ค.ศ. 1985 [4] และถูกปรับปรุงมาตามลำดับ [5] ในปัจจุบัน แนวคิด *Wordnet* ได้ถูกนำไปใช้ในการสร้างทรัพยากรภาษาอื่นๆ จำนวนมาก ซึ่งมักสร้างจาก *Wordnet* ภาษาอังกฤษเป็นจุดเริ่มต้น ส่วนคำในภาษาอื่น จะออกแบบให้สามารถจับคู่กับคำใน *Wordnet* ภาษาอังกฤษ เท่าที่จะทำได้ แม้ว่าจะมีคำที่ไม่สามารถจับคู่ได้ด้วยก็ตาม

รูปศัพท์หนึ่งคำอาจมีหลายความหมาย และหลายบทบาท เพื่อจำแนกความหมายของคำเหล่านี้ ในโครงสร้างของ *Wordnet* จะแทนคำเป็น *Synset* ซึ่งย่อมาจาก *Synonym set* นั่นคือ หนึ่งความหมายของคำจะแทนด้วยหนึ่ง *Synset* โดยมีรูปแบบดังนี้

"*<word>.<pos>.<number>*"

เมื่อ

<word> คือ รากศัพท์ของคำ

<pos> แสดงบทบาทของคำในประโยค มีบทบาทสำคัญๆ ได้แก่ *NOUN (n), VERB (v), ADJECTIVE (a or s), ADVERB (r)*

<number> เป็นเลข *sense number* เรียงลำดับตั้งแต่ 01 โดยในภาษาอังกฤษจะเรียงตามความถี่ของคำที่เกิด

ตัวอย่างเช่น คำว่า '*success*' จะมี *sense* 4 แบบคือ

*Synset('success.n.*01*'),*

*Synset('success.n.*02*'),*

*Synset('success.n.*03*'),* และ

*Synset('achiever.n.*01*')]*

โดยแต่ละแบบจะมีนิยามและตัวอย่างการใช้งานที่แตกต่างกัน เช่น

*Success.n.*01 หมายถึง *'an event that accomplishes its intended purpose'* ใช้กล่าวถึงตัวความสำเร็จ เช่น *"let's call heads a success and tails a failure"*

*Success.n.*02 หมายถึง *'an attainment that is successful'* ใช้กล่าวถึงการประสบความสำเร็จ เช่น *'his new play was a great success'*

ในด้านความคล้ายคลึงกัน *synset* ทั้งสี่จะมีค่าความคล้ายคลึงแตกต่างกัน ตามมาตรวัดที่เลือกใช้ เช่น หากเลือกใช้ *path similarity* จะได้

ค่าความคล้ายคลึงระหว่าง *success.n.*01 และ *success.n.*02 เท่ากับ 0.125

ค่าความคล้ายคลึงระหว่าง *success.n.*01 และ *success.n.*03 เท่ากับ 0.083

ค่าความคล้ายคลึงระหว่าง *success.n.*01 และ

*achiever.n.*01 เท่ากับ 0.1

ค่าความคล้ายคลึงระหว่าง *success.n.*02 และ *success.n.*03 เท่ากับ 0.067

ค่าความคล้ายคลึงระหว่าง *success.n.*02 และ *achiever.n.*01 เท่ากับ 0.077 และ

ค่าความคล้ายคลึงระหว่าง *success.n.*03 และ *achiever.n.*01 เท่ากับ 0.077

โดยค่ายิ่งมาก แสดงว่ามีความคล้ายคลึงกันมาก

การหาค่าความคล้ายคลึง มักใช้ความสัมพันธ์เชิงโครงสร้างใน *Wordnet* ในการคำนวณ โดยโครงสร้าง *wordnet* จัดเรียงเป็นลำดับชั้นตามความสัมพันธ์ต่างๆ โดยมีความสัมพันธ์ (*Relation)* ได้แก่

Hypernym – Hyponym ความสัมพันธ์เชิง *superclass-subclass* หรือที่รู้จักกันว่า *is-a taxonomy* เช่น *dog.n.*01 มี *hypernyms* คือ *'canine.n.*02*'* และ *'domestic_animal.n.*01*'* และมี *hyponyms* เป็นสุนัขพันธ์ต่างๆ เช่น *'basenji.n.*01*', 'dalmatian.n.*02*', 'poodle.n.*01*'* เป็นต้น

Meronyms - Holonyms เป็นความสัมพันธ์แบบ *part-of* ตัวอย่างเช่น หน้าต่างเป็นส่วนหนึ่ง (*meronym)* ของอาคาร และอาคารครอบคลุม (*holonym)* หน้าต่าง นอกจากนี้ ในคำกิริยา ยังมีความสัมพันธ์ต่างๆ ดังนี้

Troponym - เมื่อกิริยา *A* เป็นลักษณะย่อยของกิริยา *B* เช่น การเดินสวนสนามเป็นการเดินแบบหนึ่ง ดังนั้น การเดินสวนสนามเป็น *troponym* ของการเดิน

Entailment : เมื่อเกิดการกระทำ *A* จะเกิดการกระทำ *B* ด้วยเสมอ เช่น เมื่อคุณทานอาหารมื้อค่ำ แสดงว่าคุณกำลังทานอาหาร

นอกจากนี้ เราสามารถหาคำเหมือน (*synonym)* และคำตรงข้าม (*antonym)* ของ *synset* ได้ ถ้ามี

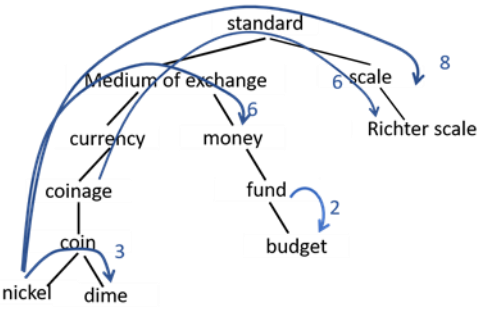
2.2 มาตรวัดความคล้ายคลึง

มาตรวัดความคล้ายคลึงเชิงความหมาย จากโครงสร้าง *Wordnet* จำนวนหกมาตรวัด [6][7] สามารถแบ่งได้เป็นสองกลุ่ม คือ กลุ่มที่วัดจากโครงสร้างเชิงลำดับชั้น ได้แก่ ได้แก่ *Path Distance Similarity, Leacock Chodorow Similarity, Wu-Palmer Similarity* และกลุ่มที่วัดจากเนื้อหาสารสนเทศ *(Information Content* หรือ *IC)* ได้แก่ *Resnik Similarity, Jiang-Conrath Similarity* และ *Lin Similarity*

2.2.1 มาตรวัดความคล้ายคลึงเชิงลำดับชั้น *[7]*

มาตรวัดสามตัวแรกคือ *Path Distance Similarity, Leacock Chodorow Similarity* และ *Wu-Palmer Similarity* เป็นกลุ่มมาตรวัดที่คำนวณจาก ความสัมพันธ์แบบ *is-a taxonomy (hypernym / hypnoym)* 1) โดย *Path Distance Similarity* จะคำนวณจากจำนวนโหนดในเส้นทางที่สั้นที่สุดที่เชื่อมเทอมทั้งสองในความสัมพันธ์แบบ *is-a taxonomy (hypernym / hypnoym)* และคืนค่าคะแนนตั้งแต่ 0 ถึง 1 มีสูตรคือ

$$pathsim = \frac{1}{pathlen(C1,C2)} \qquad (1)$$



รูปที่ 1 ตัวอย่างเพื่อคำนวณ *path similarity*

จากตัวอย่างในรูปที่ 1 สามารถคำนวณ *path similarity* ได้ดังนี้

simpath(fund, budget) = ½ = 0.5

simpath(nickel, dime) = 1/3 =0.34
simpath(nickel, currency) = 1/4 = 0.25
simpath(coinage, Richter scale) = 1/6 = 0.167
simpath(nickel, Richter scale) = 1/8 = 0.125

ส่วน *Leacock Chodorow Similarity (lch similarity)* [8] ปรับปรุงจาก *Path distance similarity* โดยคำนวณจากระยะทางที่สั้นที่สุดของเทอมสองเทอม เหมือนใน *Path distance* และความลึกสูงสุดของเทอมทั้งสอง โดยใช้สูตร

$$lchsim = -log(\frac{p}{2d}) \quad (2)$$

เมื่อ *p* คือ *shortest path length* และ
*d = 2 * max_depth*

ส่วน *Wu-Palmer Similarity (wup similarity)* [9] คำนวณจาก ความลึกของเทอมสองเทอมในอนุกรมวิธาน (*Taxonomy)* และเทอมบรรพบุรุษร่วมกันที่ใกล้ที่สุดของเทอมทั้งสอง (*Least Common Subsumer - LCS)* สูตรคือ

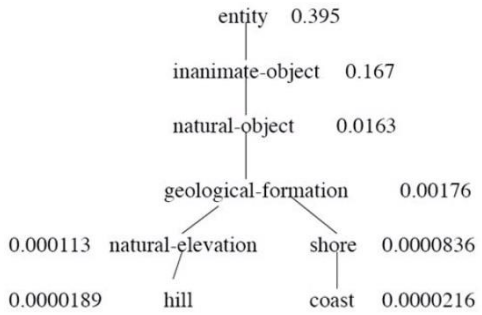
$$wupsim = \frac{2*depth}{len1+len2} \quad (3)$$

เมื่อ *depth* คือระยะทางจากโหนดรากถึงโหนด *LCS* และ *len1* และ *len2* คือระยะทางจากโหนดรากถึงโหนดเทอมที่ *1* และเทอมที่ *2* ตามลำดับ

2.2.2 มาตรวัดความคล้ายคลึงเชิงเนื้อสารสนเทศ

ค่า *Information Content (IC)* เป็นการวัดความเฉพาะเจาะจงของ *concept* ค่าที่มากหมายถึงมีความเฉพาะเจาะจงมาก ขณะที่ค่าที่น้อยลงจะมีความเป็นทั่วไปมากกว่า

ค่า *Information Content (IC)* เป็นค่าที่ได้จากการนับความถี่ของการเกิดเทอมนั้น ๆ ใน *corpus* ดังนั้น หากใช้ *corpus* ต่างกัน ค่า *IC* ก็จะต่างกันไปด้วย เมื่อพบเทอมใดๆ ใน *corpus* ค่าความถี่สำหรับเทอมนั้น รวมทั้งเทอมบรรพบุรุษของเทอมนั้นใน *wordnet taxonomy* จะเพิ่มขึ้นทุกครั้ง ในกรณีที่เทอมที่พบไม่มีการระบุ *sense* ไว้ ค่าความถี่จะถูกหารกระจายเท่าๆ กันสำหรับทุกๆ *sense* ตัวอย่างเช่น ถ้าเทอม *s* เกิดขึ้น 42 ครั้งใน *corpus* และเทอมนั้นมีจำนวน *sense* ทั้งหมด 6 แบบ เทอมนั้นและบรรพบุรุษของเทอมนั้นจะมีค่าความถี่เพิ่มขึ้น 7 ครั้ง [10]



รูปที่ *2* แสดงตัวอย่างของค่าความน่าจะเป็นที่คำนวณจากความถี่ใน *corpus*

นิยามของ *IC* เป็นดังนี้

$$IC(s) = -\log P(s) \quad (4)$$

เมื่อ *P(s)* คือความน่าจะเป็นที่จะเกิดเทอม *s* ใน *corpus* ตามวิธีการนับความถี่ที่กล่าวมาข้างต้น ดังแสดงตัวอย่างในรูปที่ *2*

Resnik Similarity, Jiang-Conrath Similarity [11] และ *Lin Similarity* [12] จะคำนวณความคล้ายคลึงเชิงเนื้อสารสนเทศ *(Information Content -IC)* ของโหนด *LCS* ของเทอมทั้งสอง มาตรวัดทั้งสามนี้ จะสามารถคำนวณได้เฉพาะ *noun* และ *verb* และคำนวณได้เฉพาะกับคู่คำที่มี *pos* เดียวกันเท่านั้น ถ้าไม่ตรงตาม

เงื่อนไขหรือไม่มีค่าจะคืนค่า *"None"*

สำหรับค่าความคล้ายคลึงของ *Resnik* นั้น จะคิดจากค่า *IC* ของ *LCS(s1,s2)* เท่านั้น เมื่อ *s1* และ *s2* เป็นเทอมที่ต้องการหาความคล้ายคลึง ดังสูตรนี้

$$ressim(s1, s2) = IC(LCS(s1, s2)) \quad (5)$$

ตัวอย่างเช่น จาก ค่าความคล้ายคลึงแบบ *Resnik* ของเทอม *dog* กับ *cat* จาก *brown corpus* เท่ากับ *7.911* และจาก *genesis corpus* เท่ากับ *7.204* ตามลำดับ จะเห็นว่าเมื่อคำนวณจาก *corpus* ที่แตกต่างกัน ค่าความคล้ายคลึงที่ได้จะไม่เท่ากัน

```
>>> dog.res_similarity(cat, brown_ic)  #
doctest: +ELLIPSIS
7.911...
>>> dog.res_similarity(cat, genesis_ic)  #
doctest: +ELLIPSIS
7.204...
```

รูปที่ *3 Resnik similarity* ของคำว่า *dog* กับ *cat* จาก *brown corpus* และ *genesis corpus*

Jiang-Conrath Similarity (jcn similarity) และ *Lin Similarity (lin similarity)* เป็นการปรับปรุงค่าความคล้ายคลึงแบบ *Resnik* ให้มีการคำนึงถึงค่า *IC* ของเทอม *s1* และ *s2* ด้วย โดยมีสูตรดังนี้

$$jcnsim(s1, s2) = \frac{1}{IC(s1)+IC(s2)-2*IC(LCS(s1,s2))} \quad (6)$$

$$linsim(s1, s2) = \frac{2*LCS(s1,s2)}{IC(s1)+IC(s2)} \quad (7)$$

2.3 ทรัพยากรคำศัพท์ Wordnet ภาษาไทย

Thai wordnet พัฒนาภายใต้โครงการ *Asian wordnet framework* [13] และใน [14] วิธีการพัฒนาเริ่มจากการพยายาม *align* คำไทยเข้ากับ *synset* ใน *Princeton wordnet (PWN)* โดยใช้ดิคชั่นนารีไทย-อังกฤษ ดังนั้นทุกเทอมไทยใน *Thai wordnet* จะมี *synset* ภาษาอังกฤษกำกับไว้ด้วย การพัฒนาทำแบบอัตโนมัติโดยใช้กฎ หลังจากนั้น จะทำการเพิ่มคำศัพท์ด้วยมือ ผลลัพธ์ได้คำศัพท์ไทย 24,457 เทอม จาก *PWN* 207,010 เทอม หรือประมาณ 12 เปอร์เซ็นต์เท่านั้น

3. คำถามวิจัยและการออกแบบการทดสอบ

3.1 คำถามวิจัย

การนำมาตรวัดเชิงความหมายระดับคำดังกล่าว มาทดสอบกับภาษาไทย โดยมีคำถามวิจัย *R1* และ *R2* ดังต่อไปนี้

R1: มาตรวัดค่าความคล้ายคลึงกันของเทอมจากทรัพยากร *Wordnet* สามารถใช้กับเทอมภาษาไทยได้หรือไม่

R2: มาตรวัดดังกล่าว จะให้ค่าความคล้ายคลึงที่สมเหตุสมผลหรือไม่ นั่นคือ ให้ค่าความคล้ายคลึงของคู่คำเหมือน มากกว่าคู่คำตรงข้าม หรือไม่

งานวิจัยนี้ ได้ออกแบบการทดสอบมาตรวัดในคำถามวิจัย *R2* โดยกำหนดคู่คำเหมือน (*synonym)* และคู่คำตรงข้าม (*antonym)* และมีสมมุติฐานว่า มาตรวัดจะใช้งานได้ เมื่อค่าความคล้ายคลึงของคู่คำเหมือนมากกว่าคู่คำตรงข้าม

ให้ *th* เป็นเทอมภาษาไทย

th-syn เป็นเทอม *synonym* ของ *th* และ

th-ant เป็นเทอม *antonym* ของ *th*

ดังนั้น หากสมมุติฐานที่ต้องการทดสอบเป็นจริง จะได้ว่า

Similarity(th, th-syn) ≥ Similarity(th, th-ant)

การทดสอบเชิงสถิติสำหรับ *R2*

ในเชิงสถิติ จะสุ่มคำไทยมาอย่างน้อย 30 คำ และวัดค่าความคล้ายคลึงคู่คำเหมือนและคู่คำตรงข้ามของคำนั้น เพื่อทำการทดสอบแบบ *two sample paired test* โดยใช้ *t-test statistics* ในการทดสอบ โดยมีสมมุติฐานที่ต้องการทดสอบดังนี้

$H0 : \mu1 \leq \mu2$

$H1 : \mu1 > \mu2$

เมื่อ $\mu 1$ เป็นค่ากลางประชากรที่ 1 คือค่าความคล้ายคลึงของคู่คำเหมือน และ $\mu 2$ เป็นค่ากลางประชากรที่ 2 คือค่าความคล้ายคลึงของคู่คำตรงข้าม หากสามารถ *reject H*0 นั่นคือ *accept H*1 จะกล่าวได้ว่า ค่าความคล้ายคลึงของคู่คำเหมือนมีค่ามากกว่าคู่คำตรงข้ามอย่างมีนัยสำคัญ และสามารถเขียน *R*2 ใหม่ดังนี้

*R*2: ค่าความคล้ายคลึงกันของคู่คำเหมือน (*synonym)* จะมีค่ามากกว่าค่าความคล้ายคลึงกันของคู่คำตรงข้าม (*antonym)* อย่างมีนัยสำคัญทางสถิติหรือไม่

3.2 การออกแบบการทดสอบ

1. เลือกคำไทยที่มีคู่คำตรงข้ามจากเว็บไซต์ต่างๆ จำนวน 99 คำ โดยคำเหล่านั้นต้องเป็นคำที่มีความถี่ในการใช้งานที่เหมาะสม และไม่ใช่คำหยุด (*stop words)* โดยตัดคำที่มีความถี่มากหรือน้อยเกินไปออก จากนั้นนำเข้าสู่การทดสอบ

2. การหาคู่คำตรงข้าม ให้หาคำตรงข้ามภาษาไทยจากเว็บไซต์ต่างๆ ก่อน โดยควรเป็นคำที่มีความถี่ในการใช้งานที่เหมาะสม จากนั้นใช้ *Thai wordnet library* ในการวัดความคล้ายคลึง โดยใช้มาตรวัดทั้ง 6 ตัว ดังที่กล่าวมาแล้ว

3. คำนวณค่าความคล้ายคลึงจากค่าเฉลี่ยค่าคล้ายคลึง มีข้อสังเกตคือ *synset* ใน *Thai wordnet* จะเป็นเทอมภาษาอังกฤษเท่านั้น โดยหนึ่งเทอมไทยจะมี *synset* ภาษาอังกฤษหลายเทอม และเทอมเหล่านั้นก็สามารถแปลกลับเป็นคำไทยได้หลายคำ นอกจากนี้ การหาค่า *similarity* จากมาตรวัดทั้งหก ใน *Thai wordnet* สามารถทำได้กับคู่คำภาษาอังกฤษเท่านั้น ดังนั้น จึงออกแบบการคำนวณค่า *similarity* ดังนี้ สำหรับเทอม *th* ภาษาไทย ที่มี *synset syn*1, *syn*2, …, *synn* ให้คิด *similarity* สำหรับแต่ละมาตรวัด สำหรับทุกคู่คำ (*syn*1, *syn*2), …., *(synn*-1, *synn)* จากนั้นหาค่าเฉลี่ยของทุกคู่คำสำหรับมาตรวัดนั้นๆ ทำจนครบทุกมาตรวัด จากนั้นหาค่าเฉลี่ย *similarity* ของทั้ง 6 มาตรวัด

ในกรณีคำตรงข้าม ก็ทำเช่นเดียวกัน โดยให้ *th_ant* เป็นคำตรงข้ามภาษาไทยของ *th* และ *ant*1, *ant*2, …, *antn* เป็น *synset* ภาษาอังกฤษของ *th_ant* เราเริ่มจากหาค่าเฉลี่ยของทุกคู่คำ (*syn*1, *ant*1), *(syn*1, *ant*2) …., *(synn, antn*-1), *(synn, antn)* ของแต่ละมาตรวัด จากนั้นนำทั้งหกค่ามาหาค่าเฉลี่ยของทั้ง 6 มาตรวัด

4. ผลลัพธ์การทดสอบ วิเคราะห์และสรุปผลการทดสอบ

4.1 ผลลัพธ์การทดสอบมาตรวัดสำหรับภาษาไทย

จากการทดลองใช้งาน พบว่าไม่สามารถเปรียบเทียบเทอมไทยสองเทอมได้โดยตรงเหมือนเทอมภาษาอังกฤษ เนื่องจาก ในการพัฒนา *Thai wordnet* ไม่ได้สร้างโดยตรงจากเทอมและความสัมพันธ์ภาษาไทย แต่สร้างโดยการเชื่อมต่อกับ *PWN* ดังนั้น เทอมไทยจึงไม่มี *synset* ภาษาไทย แต่จะแสดงเป็น *synset* ภาษาอังกฤษเท่านั้น ดังแสดงในรูปที่ 4 แสดงให้เห็น *synset* ของคำไทยคือ *'หนึ่ง'* ว่ามี *synset* ภาษาอังกฤษ *4* คำ

```
>>> a=thaiwordnet()
>>> print(a.synsets('หนึ่ง'))
[Synset('one.s.05'), Synset('one.s.04'), Synset('one.s.01'), Synset('one.n.01')]
```

รูปที่ *4* แสดง *synset* ของคำว่า 'หนึ่ง' ซึ่งตรงกับ 4 *synsets* ในภาษาอังกฤษ

จาก

รูปที่ *5* แสดงการหาค่า *similarity* ของคำว่า 'วิ่ง' กับคำเหมือนและคำตรงข้าม คือคำว่า 'หยุด' โดย

บรรทัดที่ 4 และบรรทัดที่ 7 แสดงรายการ *synsets* คำเหมือนและคำตรงข้ามตามลำดับ

บรรทัดที่ 8-14 แสดงค่าความคล้ายคลึงเมื่อเทียบกับตัวเอง ซึ่ง *path similarity* และ *wup similarity* จะให้ค่าเป็น 1 ตามนิยาม ส่วน *lch similarity* ให้ค่าค่อนข้างสูงตามสูตร

บรรทัดที่ 15-23 แสดงค่าความคล้ายคลึงของคำในรายการคำเหมือนด้วยกัน ที่ไม่ใช่ตัวเอง ส่วนบรรทัดที่ 24-36 แสดงค่าความคล้ายคลึงของคำในรายการคำเหมือนและรายการคำตรงข้าม

บรรทัดที่ 37-45 แสดงให้เห็นว่า คำที่ไม่คล้ายกันมักจะไม่มีค่า *similarity* คือคืนค่าเป็น *None* เช่น ในบรรทัดที่ 44 ที่วัด *similarity* ของคู่คำ (*w1='small.a.01', w3='large.a.01'*) พบว่า คืนค่าเป็น *None*

```
1)  >>> th = thaiwordnet()
2   // synsets of วิ่ง
3)  >>> syns = th.synsets('วิ่ง')
4)  [Synset('run.v.29'), Synset('run.v.01')]
5)  // synsets of หยุด
6)  >>> ants = th.synsets('หยุด')
7)  [Synset('cut.v.36'), Synset('stop.v.05'),
Synset('halt.v.01'), Synset('stop.v.01'), Synset('cheese.v.01')]
8)  // similarity of the word itself
9)  >>> print(syns[0].path_similarity(syns[0]))
10) 1.0
11) >>> print(syns[0].wup_similarity(syns[0]))
12) 1.0
13) >>> print(syns[0].lch_similarity(syns[0]))
14) 3.258096538021482
15) // similarity of the word and its synonyms
16) >>> print(syns[0].path_similarity(syns[1]))
17) 0.2
18) >>> print(syns[0].wup_similarity(syns[1]))
19) 0.25
20) >>> print(syns[1].wup_similarity(syns[0]))
21) 0.25  22)  >>>
print(syns[0].lch_similarity(syns[1]))
23) 1.6486586255873816
24) // similarity of its synonyms and antonyms
25) >>> print(syns[0].path_similarity(ants[0]))
26) 0.1111111111111111
27) >>> print(syns[0].lch_similarity(ants[0]))
28) 1.0608719606852628
29) >>> print(syns[0].wup_similarity(ants[0]))
30) 0.2
31) >>> print(syns[0].lch_similarity(ants[1]))
32) 1.6486586255873816
33) >>> print(syns[0].lch_similarity(ants[0]))
34) 1.0608719606852628
35) >>> print(syns[1].lch_similarity(ants[3]))
36) 1.6486586255873816
37) // return None if no similarity applied
38) >>> w1 = th.synset('small.a.01')
39) >>> w2 = th.synset('run.v.29')
40) >>> print(w1.path_similarity(w2))
41) None
42) >>> w3 = th.synset('large.a.01')
43) Synset('large.a.01')
44) >>> print(w1.lch_similarity(w3))
45) None
```

รูปที่ *5* **การคำนวณค่าความคล้ายคลึงสำหรับคู่คำไทย**

ข้อสังเกตคือ ในบรรดาคำทั่วๆ ไป หากสุ่มคู่คำสองคู่มา โดยส่วนใหญ่จะไม่พบค่า *similarity* เลย ปัจจัยนี้ อาจส่งผลต่อการเปรียบเทียบค่าความคล้ายคลึงของคู่คำเหมือนและคู่คำตรงข้าม นอกจากนี้ หากผิดเงื่อนไขบางประการของมาตรวัด เช่น มาตรวัด *res similarity, jcn similarity* และ *lin similarity* คือ คำทั้งสองต้องมี *pos* เดียวกัน ไม่เช่นนั้น โปรแกรมจะเกิด *Error* คือไม่ทำงาน

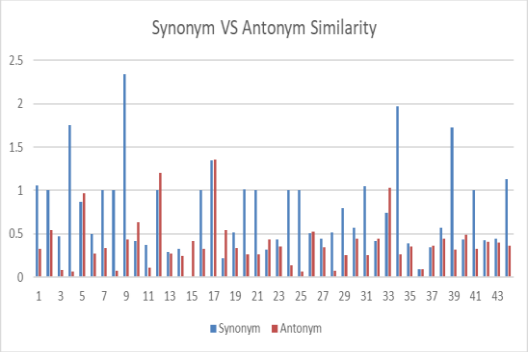
ดังนั้น คำตอบสำหรับ *R1* มาตรวัดค่าความคล้ายคลึงกันของเทอมจากทรัพยากร *Wordnet* สามารถใช้กับเทอมภาษาไทยได้หรือไม่ คำตอบคือ ใช้ได้ แต่ไม่ใช่โดยตรง เนื่องจากการเปรียบเทียบ *similarity* ทั้งหกมาตรวัด จะรับ *parameter* เป็น *synset* สองเทอม และทรัพยากร *Thai wordnet* คืนค่าเป็น *synset* เป็นภาษาอังกฤษเท่านั้น ดังนั้น จึงต้องวัดทางอ้อม โดยนำค่า *synset* ภาษาอังกฤษ มาหาค่าความคล้ายคลึงแทน

มีข้อสังเกตเพิ่มเติมคือ โดยปรกติ *synset* ของคำมักมีมากกว่าหนึ่ง ดังนั้น คู่คำและค่าความคล้ายคลึงสำหรับเทอมหนึ่งๆ จึงมีจำนวนมาก และหากสุ่มค่าคำใดๆ มาคู่หนึ่ง มาวัดค่า *similarity* จากมาตรวัดทั้งหก มักจะพบว่าไม่มีค่าความคล้ายคลึงคืนกลับมา

4.2 ผลลัพธ์การทดสอบค่าความคล้ายคลึงสำหรับคำเหมือนและคำตรงข้ามภาษาไทย

จากคู่คำจำนวน 99 คำ พบว่ามีคู่คำที่มาตรวัดคืนค่าเป็น *None* สำหรับคู่คำเหมือน หรือ คู่คำตรงข้ามจำนวน

มากคือ 55 คำ ซึ่งจะทำให้ไม่สามารถเปรียบเทียบค่าเพื่อทดสอบสมมุติฐานได้ จึงตัดคำเหล่านั้นออกไป เหลือคำที่ใช้ทดสอบ 44 คำ รูปที่ *6* แสดงกราฟค่าความคล้ายคลึงสำหรับคู่คำเหมือนและคู่คำตรงข้าม จะพบว่าคำเหมือนมีค่าความคล้ายคลึงมากกว่าคำตรงข้ามอยู่ 32 คำจาก 44 คำ คิดเป็น 72.7%



รูปที่ *6* ความคล้ายคลึงของคู่คำเหมือนและคู่คำตรงข้ามภาษาไทย 44 คำ

การทดสอบสมมุติฐานด้วย *paired data t-test* แสดงได้ดังตารางที่ 1 ค่ากลางของความคล้ายคลึงคู่คำเหมือน เท่ากับ 0768 (*variance* 0.25) มากกว่าค่ากลางของความคล้ายคลึงคู่คำตรงข้ามคือ 0.390 (*variance* 0.079) ผลการทดสอบสมมุติฐาน พบว่า ค่า *t-statistics* มีค่า 4.5 ซึ่งมากกว่า ค่า *t-critical* คือ 1.68 ในการทดสอบแบบ *one-tail* ที่ระดับนัยสำคัญ 0.05 ได้ค่า $p = .0000254$ ซึ่งน้อยกว่า 0.05 ดังนั้น จึง *reject H0* : $\mu 1 \leq \mu 2$ และยอมรับ $H1 : \mu 1 > \mu 2$

ตารางที่ 1 Paired data t-test for mean

	Synonym	*Antonym*
Mean	0.7683625	0.39042128
Variance	0.24815581	0.078957676
Observations	44	44
Pearson Correlation	0.06063214	
Hypothesized Mean Difference	0	
df	43	
t Stat	4.50165585	
P(T<=t) one-tail	2.5402E-05	
t Critical one-tail	1.6810707	
P(T<=t) two-tail	5.0804E-05	
t Critical two-tail	2.0166922	

นั่นคือ ค่าความคล้ายคลึงของคู่คำเหมือน มากกว่าค่าความคล้ายคลึงของคู่คำตรงข้ามอย่างมีนัยสำคัญที่ 0.05 หรือที่ความเชื่อมั่น 95%

ดังนั้น จึงสามารถสรุปสำหรับสมมุติฐาน *R2* ได้ว่ามาตรวัดทั้งหก สามารถจำแนกความแตกต่างระหว่างค่าความคล้ายคลึงของคู่คำเหมือนและคู่คำตรงข้ามได้อย่างมีนัยสำคัญ

จากการวิเคราะห์เบื้องต้น ในกลุ่มคำที่มีค่าความคล้ายคลึงของคำเหมือนน้อยกว่าคู่คำตรงข้ามนั้น พบว่าในกลุ่มคำเหล่านี้ บางคำมีค่าเฉลี่ยความคล้ายคลึงของคำเหมือนและคำตรงข้ามที่ใกล้เคียงกัน เช่นคำว่า สวม-ถอด มีค่าความคล้ายคลึงของคำเหมือนและคำตรงข้ามที่ 0.345492121 และ0.361495308 ตามลำดับ นอกจากนี้ พบว่า คู่คำตรงข้ามมีบางมิติที่คล้ายคลึงกันอย่างมาก เช่น เด็ก-ผู้ใหญ่ แม่-พ่อ ไป-มา เป็นต้น อย่างไรก็ตาม การทำ *error analysis* จำเป็นต้องทำการวิเคราะห์เชิงลึกต่อไป

5. ข้อสรุปและข้อเสนอแนะ

งานวิจัยนี้ ได้ทดสอบมาตรวัดระยะห่างกับภาษาไทย โดยเลือกใช้มาตรวัดระยะห่างเชิงความหมายในระดับคำ จำนวนหกมาตรวัด จากทรัพยากรคำศัพท์ *wordnet* และ *Thai wordnet* เพื่อตอบคำถามวิจัย 2 ข้อ คือ มาตรวัดดังกล่าว สามารถนำมาใช้กับภาษาไทยได้หรือไม่ และมาตรวัดดังกล่าว สามารถวัดระยะห่างข้อความอย่างสมเหตุสมผลหรือไม่

จากการทดลอง พบว่า มาตรวัดทั้งหก ได้แก่ *Path Distance Similarity, Leacock Chodorow Similarity, Wu-Palmer Similarity, Resnik Similarity, Jiang-Conrath Similarity* และ *Lin Similarity* สามารถใช้วัดค่าความคล้ายคลึงสำหรับคู่ *synset* สองคู่ โดยหากคู่ *synset* ไม่มีความคล้ายคลึงที่สามารถวัดได้ มาตรวัดจะคืนค่า *None* เนื่องจาก *Thai*

wordnet มี *synset* เป็นเทอมภาษาอังกฤษ การทดลองจึงออกแบบการวัดคู่ *synset* ภาษาอังกฤษของเทอมไทย ดังนั้น จึงตอบคำถามวิจัยที่ 1 ได้ว่า มาตรวัดสามารถใช้งานกับภาษาไทยได้โดยอ้อม ผ่าน *synset* ของคำไทยใน *Thai wordnet*

ส่วนคำถามวิจัยที่สอง ได้ออกแบบการทดลองแบบ *paired data t-test* โดยตั้งสมมุติฐานว่า หากมาตรวัดมีความสมเหตุสมผล ค่าความคล้ายคลึงของคู่คำเหมือนควรต้องมีค่ามากกว่าคู่คำตรงข้าม การทดสอบสมมุติฐานได้ผลลัพธ์ว่า ค่าความคล้ายคลึงของคู่คำเหมือนมากกว่าค่าความคล้ายคลึงของคู่คำตรงข้าม อย่างมีนัยสำคัญที่ 0.05 หรือ ที่ความเชื่อมั่น 95%

โดยสรุป การวัดระยะห่างข้อความเชิงความหมายสำหรับภาษาไทยโดยตรง ยังไม่สามารถทำได้เนื่องจากทรัพยากรคำศัพท์สร้างมาจากทรัพยากรภาษาอังกฤษ นอกจากนี้ และคู่คำส่วนใหญ่ มาตรวัดเชิงความหมายคืนค่า *None* มาให้ ดังนั้น ในภาษาไทย จึงควรมีการสร้างมาตรวัดเชิงสถิติเพิ่มเติม เพื่อใช้ในการวัดระยะห่างข้อความสำหรับภาษาไทยโดยเฉพาะ

กิตติกรรมประกาศ

บทความนี้เป็นส่วนหนึ่งของโครงการวิจัยเรื่อง “การสืบค้นมาตรวัดระยะห่างของข้อความ” ซึ่งได้รับการสนับสนุนทุนวิจัยจากแหล่งเงินทุนวิจัยเงินรายได้คณะเทคโนโลยีสารสนเทศ สจล. ประจำปีงบประมาณ 2560

เอกสารอ้างอิง

[1] Nitesh, P., Gyanchandani, M., & Wadhvani, R. (2015, June). A Review on Text Similarity Technique used in IR and its Application. International Journal of Computer Applications (0975 – 8887), 120(9).

[2] Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de França Silva, G., Simske, S. J., & Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. Expert Systems with Applications, 41(13), 5780-5787.

[3] Chainapaporn, P., & Netisopakul, P. (2013). Word similarity algorithm for merging Thai herb information from heterogeneous data sources. International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 159-163). York Jarkata, Indonesia: IEEE.

[4] Miller, G. B. (1990). WordNet: An online lexical database. . Int. J. Lexicograph, 3(4), 235–244.

[5] Miller, G. A. (1995). "WordNet: a lexical database for English.". Communications of the ACM , 38(11), 39-41.

[6] Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration papers at HLT-NAACL 2004 (pp. 38-41). Association for Computational Linguistics.

[7] Yang, D., & Powers, D. M. (2005, January). Measuring semantic similarity in the taxonomy of WordNet. In Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38 (pp. 315-322). Australian Computer Society, Inc..

[8] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet

similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

[9] Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces, New Mexico.

[10] Pederson, T. (2010). Information Content Measures of Semantic Similarity. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (pp. 329-332). Los Angeles, CA: Association for Computational Linguistics.

[11] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In C. S. Mellish (Ed.), Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95). 1, pp. 448–453. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

[12] Lin, D. (1998, July). An information-theoretic definition of similarity. In Icml (Vol. 98, No. 1998, pp. 296-304).

[13] Sornlertlamvanich, V., Potipiti, T., Wutiwiwatchai, C., & Mittrapiyanuruk, P. (July 31-August 04, 2000). The State of the Art in Thai Language Processing. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, (pp. 1-2). Saarbrücken, Germany. doi:10.3115/1075218.1075296

[14] Thoongsup, S. R. (2009). Thai WordNet construction. Proceedings of the 7th workshop on Asian language resources (pp. 139-144). Association for computaional linguistics