การจับคู่ประโยคที่ตรงกันในคลังข้อความขนานด้วยอนุกรมเวลา

นางสาวศิรินันท์ สินธุวาทิน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

SENTENCE ALIGNMENT IN PARALLEL TEXT CORPORA USING TIME SERIES

Miss Sirinun Sintuwatin

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2007

หัวข้อวิทยานิพนธ์	การจับคู่ประโยคที่ตรงกันในคลังข้อความขนานด้วยอนุกรมเวลา
โดย	นางสาวศิรินันท์ สินธุวาทิน
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.พรศิริ หมื่นไชยศรี)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

..... กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร.ทรงศรี ตั้งศรีไพโรจน์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ นครทิพย์ พร้อมพูล)

ศิรินันท์ สินธุวาทิน : การจับคู่ประโยคที่ตรงกันในคลังข้อความขนานด้วยอนุกรมเวลา. (SENTENCE ALIGNMENT IN PARALLEL TEXT CORPORA USING TIME SERIES) อ. ที่ปรึกษา : อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ, 106 หน้า.

ในปัจจุบันโปรแกรมประยุกต์ที่พัฒนาจากคลังข้อความขนานมีเพิ่มมากขึ้นเรื่อย ๆ โดยเฉพาะอย่างยิ่งในด้านการค้นคืนข้ามภาษา การแปลภาษาด้วยเครื่องและมนุษย์ และการประมวลผลภาษาธรรมชาติ ทำให้การประมวลผลคลังข้อความขนานกลายเป็นเรื่องที่นักวิจัยให้ความสนใจมากขึ้น ในงานวิจัยนี้นำเสนอกลวิธีในการจับคู่ประโยคที่ตรงกันในคลังข้อความขนานโดยใช้อนุกรมเวลาซึ่งจะเก็บข้อมูลเกี่ยวกับความถี่และตำแหน่งของคำที่ปรากฏในคลังข้อความขนานสองภาษาใด ๆ และทำการจับคู่คำโดยการวัดความเหมือนกันของอนุกรมเวลา วิธีนี้มีข้อดีคือ ไม่ต้องใช้ความรู้ทางภาษาศาสตร์ เช่น ไวยากรณ์ วากยสัมพันธ์ โครงสร้างประโยค และการแปลจากพจนานุกรม เป็นต้น อย่างไรก็ตาม แม้ว่าคำที่เป็นคำเดียวกันในคลังข้อความขนานหลายภาษามักจะมีความถี่และตำแหน่งของการปรากฏคล้ายกัน ทำให้สามารถจับคู่ประโยคโดยใช้คำเหล่านี้เป็นตัวบ่งชี้ได้ แต่ก็ยังมีคำอีกเป็นจำนวนมากที่ไม่สามารถจับคู่คำด้วยวิธีนี้ได้ จากการทดลองพบว่าวิธีนี้เป็นประโยชน์และให้ผลดีกับข้อความขนานขนาดสั้นประมาณ 1 หน้ามากกว่าข้อความขนาดยาว เมื่อทดลองกับข้อความขนาดสั้นโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันความถูกต้องเฉลี่ยคิดเป็น 58 เปอร์เซ็นต์

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ศิรินันท์ สินธุวาทิน
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2550

4971474421 : MAJOR COMPUTER SCIENCE

KEY WORD: PARALLEL TEXT PROCESSING / WORD ALIGNMENT / SENTENCE ALIGNMENT / TIME SERIES

SIRINUN SINTUWATIN : SENTENCE ALIGNMENT IN PARALLEL TEXT CORPORA USING TIME SERIES. THESIS ADVISOR : CHOTIRAT RATANAMAHATANA, Ph.D., 106 pp.

As applications based on parallel corpora (parallel text) has increasingly expanded, especially in the areas of cross-language informational retrieval, machine/human translation, natural language processing, and multilingual lexicography, parallel-text processing has become the heart of the development. In this research, we propose a novel sentence alignment technique. We exploit a notion of time series representation, recording the position and frequency of word appearance, without any requirement of any linguistic knowledge, e.g. grammar/syntax, sentence structure, dictionary lookup, etc. We align word by using similarity measurement and the result of word alignment will be subsequently used for sentence alignment. Our intuition lies in the belief that similar words in any multilingual parallel text should possess similar frequency and the position of word occurrences. However, the experiment results have revealed several limitations of the method, where its utility and effectiveness seem to work better with short parallel text about 1 page. The experiment result on short parallel text by using manhattan distance gives an accuracy of 58 percent.

Department ..Computer Engineering.....Student's signature.....

Field of study.....Computer Science..... Advisor's signature.....

Academic year .....2007.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี เนื่องมาจากความช่วยเหลืออย่างดียิ่งของท่าน อ.ดร.โชติรัตน์ รัตนามหัทธนะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลาให้คำปรึกษาแนะนำแนวทางเกี่ยวกับงานวิจัยอย่างดีตลอดมาจนเสร็จสมบูรณ์ และผู้วิจัยขอกราบขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ รศ.ดร.พรศิริ หมื่นไชยศรี ผศ.นครทิพย์ พร้อมพูล และ อ.ดร.ทรงศรี ตั้งศรีไพโรจน์ ที่ได้ให้คำแนะนำ ข้อคิดเห็น ข้อเสนอแนะ และแนวทางในการพัฒนางานวิจัย

ขอขอบคุณรัฐบาลสำหรับทุนอุดหนุนการศึกษา งบประมาณแผ่นดิน ประจำปี 2549

ขอขอบคุณ ดร.เทพชัย ทรัพย์นิธิ และนักวิจัยที่เนคเทค ที่ให้ข้อมูลเพื่อใช้ในการทดลองและให้คำแนะนำเป็นอย่างดี

ขอขอบคุณ พี่ธุรการภาคฯ ทุกคนที่ช่วยอำนวยความสะดวกในการทำงานและช่วยตักเตือนแนะนำสิ่งดี ๆ เสมอมา

สุดท้ายนี้ ขอกราบขอบพระคุณคุณพ่อคุณแม่ที่คอยเลี้ยงดู และสนับสนุนในด้านการศึกษาเป็นอย่างดีเสมอมา

สารบัญ

หน้า

หน้า

หน้า

สารบัญภาพ

หน้า

สารบัญตาราง

หน้า

หน้า

หน้า

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญ

ในปัจจุบันภาษาต่างประเทศได้เข้ามามีบทบาทต่อการสื่อสารในชีวิตประจำวันของมนุษย์มากขึ้นโดยเฉพาะภาษาอังกฤษ ปัญหาที่พบคือ คนไทยส่วนใหญ่ไม่มีความเชี่ยวชาญในภาษาอังกฤษ ทำให้ต้องเปิดพจนานุกรมประกอบการอ่านอยู่ตลอดเวลา ซึ่งต้องใช้เวลามาก ปัญหานี้สามารถลดได้โดยอาศัยระบบการแปลภาษาด้วยเครื่อง (Machine Translation) ซึ่งเป็นระบบที่นำคอมพิวเตอร์มาช่วยอำนวยความสะดวกในการแปลข้อความจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง (ซึ่งจะกล่าวถึงรายละเอียดในหัวข้อ 2.1.1) อย่างไรก็ตาม ผลที่ได้จากระบบการแปลภาษาด้วยเครื่องในปัจจุบันยังไม่เป็นที่น่าพอใจ จึงมีผู้คิดพัฒนาระบบดังกล่าวอยู่อย่างต่อเนื่อง

การประมวลผลคลังข้อความขนาน (Parallel Text Processing) เป็นพื้นฐานส่วนหนึ่งที่นำไปสู่การพัฒนาการแปลภาษาด้วยเครื่อง โดยให้เครื่องเรียนรู้การแปลจากประโยคตัวอย่าง ซึ่งได้มาจากกระบวนการจับคู่ประโยคที่ตรงกัน (Sentence Alignment) ในคลังข้อความขนาน โดยคลังข้อความขนาน (Parallel Text) คือข้อความในภาษาหนึ่งประกอบกับคำแปลของข้อความนั้นในภาษาอื่น [1]

การจับคู่ประโยคที่ตรงกันในคลังข้อความขนานไม่ใช่เรื่องง่าย [1] เนื่องจากการแปลข้อความในภาษาหนึ่งเป็นอีกภาษาหนึ่ง จะมีบางประโยคถูกเพิ่มเข้ามาเพื่อขยายความ และบางประโยคถูกละทิ้งไม่ทำการแปล อีกทั้งลำดับก่อนหลังของประโยคมักจะไม่ตรงกัน

จากการศึกษาพบว่าวิธีการจับคู่ประโยคที่ตรงกันในคลังข้อความขนาน ส่วนใหญ่จะอ้างอิงสองแนวความคิด [1] ได้แก่ 1. การพิจารณาความยาวของประโยค และ 2. การหาคำที่เป็นคู่กันแล้วจึงระบุประโยคที่ตรงกันโดยพิจารณาจากตำแหน่งคู่คำนั้น ๆ โดยคำที่เป็นคู่กันนั้นอาจได้มาจากพจนานุกรม การพิจารณาคำที่มาจากรากศัพท์เดียวกัน หรือจากการจับคู่คำที่ตรงกัน (Word Alignment) จากการศึกษาตัวอย่างคลังข้อความขนานภาษาอังกฤษและภาษาไทย พบว่าส่วนใหญ่ข้อความภาษาไทยจะยาวกว่า และมีจำนวนคำที่มากกว่าข้อความภาษาอังกฤษ อีกทั้งประโยคต้นฉบับ (ในที่นี้กำหนดให้ภาษาอังกฤษเป็นต้นฉบับ) ที่สั้นก็ไม่ได้คู่กับประโยคคำแปลที่สั้นเสมอไป เนื่องจากประโยคคำแปลบางประโยคใช้คำฟุ่มเฟือย ดังนั้นผู้วิจัยจึงคิดว่าแนวความคิดที่สองน่าจะเหมาะสมกับคลังข้อความขนานภาษาอังกฤษและภาษาไทยมากกว่า เนื่องจากถ้าทราบ

คำที่เป็นคู่กันในทั้งสองภาษาแล้วก็มีความเป็นไปได้มากที่จะจับคู่ประโยคที่ตรงกันได้โดยไม่ต้องคำนึงถึงความยาวของประโยค ซึ่งงานวิจัยนี้ศึกษาเกี่ยวกับการจับคู่คำที่ตรงกันและการจับคู่ประโยคที่ตรงกันในคลังข้อความขนานด้วยอนุกรมเวลา

อย่างไรก็ตาม นอกจากจะให้เครื่องเรียนรู้การแปลจากคู่ประโยคตัวอย่างแล้ว ยังสามารถนำคู่ประโยคตัวอย่างไปใช้ในการแปลโดยตรงได้ เช่น การแปลแบบ Translation Memory และ Example-based Machine Translation (ซึ่งจะกล่าวถึงรายละเอียดในหัวข้อ 2.1.1.2 และ 2.1.1.4 ตามลำดับ)

1.2 วัตถุประสงค์ของการวิจัย

ออกแบบกระบวนการที่ใช้ในการจับคู่ประโยคที่ตรงกันในคลังข้อความขนานภาษาอังกฤษและภาษาไทย โดยใช้อนุกรมเวลา และทำการวัดความคล้ายกันของอนุกรมเวลา พร้อมทั้งทดสอบและวิเคราะห์ว่ากระบวนการที่ออกแบบนี้มีความเหมาะสมสำหรับการจับคู่ประโยคมากน้อยเพียงใด

1.3 ขอบเขตของการวิจัย

1. คลังข้อความขนานที่ใช้ประกอบด้วยข้อความภาษาอังกฤษและภาษาไทย
2. ตัดคำในข้อความภาษาไทยโดยใช้โปรแกรม SWATH ซึ่งพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ [2]
3. ลดรูปคำศัพท์ในภาษาอังกฤษให้อยู่ในรูปทั่วไปหรือรากศัพท์โดยใช้โปรแกรม KTAGGER [3]
4. ทำการทดลองกับคลังข้อความขนานที่มีขนาดแตกต่างกัน
5. เปรียบเทียบผลการจับคู่ประโยคระหว่างวิธีที่ใช้อนุกรมเวลาและไม่ได้ใช้อนุกรมเวลา
6. จับคู่ประโยคโดยจับคู่ 1 ประโยคในภาษาอังกฤษกับ *N* วรรคในภาษาไทย

1.4 ขั้นตอนการวิจัย

1. หาคลังข้อความขนานของภาษาอังกฤษและภาษาไทยเพื่อนำมาใช้ในการวิจัย
2. หาชุดโปรแกรมที่เป็นประโยชน์ในการเตรียมข้อมูลภาษาอังกฤษและภาษาไทย
3. ออกแบบวิธีในการจัดเก็บข้อมูลคำ ประโยค และย่อหน้า
4. เตรียมข้อมูลก่อนการดำเนินการจับคู่คำ
5. ค้นคว้าหาวิธีการจับคู่คำโดยพิจารณาจากความถี่และตำแหน่งที่ปรากฏของคำ
6. พัฒนาโปรแกรมเพื่อจับคู่คำ

7. ทดสอบการจับคู่คำและวิเคราะห์หาผลการทดสอบ
8. ค้นคว้าหาวิธีการจับคู่ประโยค โดยพิจารณาจากคู่คำที่ได้
9. พัฒนาโปรแกรมเพื่อจับคู่ประโยค
10. ทดสอบการจับคู่ประโยคและวิเคราะห์ผลการทดสอบ
11. สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่ได้รับ

1. สามารถจับคู่คำและคู่ประโยคในคลังข้อความขนานได้โดยใช้อนุกรมเวลา และไม่ต้องใช้ความรู้ทางภาษาศาสตร์ ยกเว้นการลดรูปคำศัพท์ภาษาอังกฤษ
2. ได้คู่ประโยคตัวอย่างที่นำไปใช้สำหรับพัฒนาระบบการแปลภาษาด้วยเครื่องที่ใช้เทคนิคการแปลโดยอาศัยตัวอย่าง

1.6 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้ถูกแบ่งออกเป็น 6 บท ดังนี้คือ บทที่ 1 เป็นบทนำ บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง เช่น การแปลภาษาด้วยเครื่อง การลดรูปคำศัพท์ และการหาค่าเฉลี่ยของส่วนกลับลำดับชั้น เป็นต้น บทที่ 3 กล่าวถึงการดำเนินงานวิจัย โดยอธิบายเป็นขั้นตอนต่าง ๆ ทั้งการจับคู่คำและการจับคู่ประโยค บทที่ 4 จะกล่าวถึงข้อมูลเข้าและพารามิเตอร์ ส่วนในบทที่ 5 เป็นการทดลองและผลที่ได้จากการทดลองตามชุดการทดลองต่าง ๆ และท้ายสุดคือบทที่ 6 เป็นการสรุปผลการวิจัยและข้อเสนอแนะของงานวิจัย ซึ่งอาจจะเป็นประโยชน์ต่องานวิจัยอื่น ๆ ต่อไปในอนาคต

1.7 ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตีพิมพ์เป็นบทความทางวิชาการ (Sirinun Sintuwatin and Chotirat Ratanamahatana. Parallel Text Alignment Using Bursty Sequences. 2nd International Conference on Advances in Information Technology 2007 : pp.163-170.) ในงานประชุมวิชาการนานาชาติ ณ โรงแรมเอเชีย กรุงเทพมหานคร ระหว่างวันที่ 1-2 พฤศจิกายน 2550 ดังแสดงในภาคผนวก ค

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การแปลภาษาด้วยเครื่อง

ระบบการแปลภาษาด้วยเครื่อง เป็นแขนงหนึ่งของภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics) เป็นงานที่สนใจเกี่ยวกับการหาวิธีใช้โปรแกรมคอมพิวเตอร์เพื่อช่วยในการแปลข้อความหรือคำพูดจากภาษาหนึ่งเป็นอีกภาษาหนึ่ง ในอดีตมีผู้คิดหาแนวทางในการแปลไว้หลายแนวทาง ได้แก่

2.1.1.1 การแปลแบบคำต่อคำ (Word-by-word Translation) [4] เป็นการแปลโดยใช้พจนานุกรม ข้อดีคือง่ายต่อการแปล แต่มีข้อเสียคือประสิทธิภาพไม่ดีนัก เนื่องจากคำส่วนใหญ่มีได้หลายความหมาย จึงเป็นการยากที่คอมพิวเตอร์จะเลือกความหมายได้ถูกต้องตรงกับบริบท ตัวอย่างเช่น “My major in the university is computer science.” อาจแปลได้เป็น “*สาขาวิชา*ที่ฉันเรียนในมหาวิทยาลัยคือวิทยาการคอมพิวเตอร์” หรือแปลว่า “*แก่น*ของฉันในมหาวิทยาลัยวิทยาการคอมพิวเตอร์” ซึ่งประโยคแรกเป็นความหมายที่ถูกต้องตรงกับบริบทมากกว่า ส่วนประโยคที่สองเป็นตัวอย่างความหมายที่ได้จากโปรแกรมแปลภาษาที่มีอยู่ในปัจจุบัน

2.1.1.2 หน่วยความจำสำหรับการแปล (Translation Memory : TM) [5] เป็นฐานข้อมูลชนิดหนึ่งที่ออกแบบมาเพื่อช่วยผู้แปล โดยฐานข้อมูลนี้จะเก็บข้อความต้นฉบับกับความหมายของข้อความนั้นในอีกภาษาหนึ่ง โดยข้อความนั้นอาจจะเป็นคำ วลี ประโยคหรือย่อหน้า การแปลทำได้โดยค้นหาสิ่งที่ต้องการแปลจากฐานข้อมูลและแปลตามตัวอย่างนั้น ข้อดีได้แก่ ลดต้นทุนในการแปลระยะยาว และการแปลจะสอดคล้องกัน ตัวอย่างเช่น ประโยคต้นฉบับที่เหมือนกันจะแปลเป็นประโยคความหมายเดียวกันตลอดทั้งเอกสาร ข้อเสียคือ มีความเป็นไปได้น้อยที่วลีหรือประโยคที่ต้องการแปลจะเหมือนกับตัวอย่างที่มีในฐานข้อมูล

2.1.1.3 การแปลโดยอาศัยกฎไวยากรณ์ (Rule-based Machine Translation : RBMT) [6] เป็นการแปลที่ต้องมีการวิเคราะห์โครงสร้างประโยคและความหมาย ข้อดีคือ สามารถวิเคราะห์ทั้งโครงสร้างและความหมายของประโยคได้อย่างลึกซึ้ง ข้อเสียได้แก่ การเขียนกฎให้ครอบคลุมทั้งหมดเป็นไปได้ยาก และกฎที่เขียนขึ้นอาจขัดแย้งกัน

2.1.1.4 การแปลโดยอาศัยตัวอย่าง (Example-based Machine Translation : EBMT) [7] เป็นการแปลโดยการแบ่งข้อความที่ต้องการแปลออกเป็นส่วน ๆ และแปลแต่ละส่วนโดยอาศัยตัวอย่างการแปลของคน ตัวอย่างนี้จะเก็บอยู่ในฐานข้อมูล วิธีนี้จะคล้ายกับวิธีหน่วยความจำสำหรับการแปลแต่วิธีหน่วยความจำสำหรับการแปลจะไม่มีการแบ่งข้อความที่ต้องการแปลออกเป็นส่วน ๆ วิธีนี้มีข้อดีคือ อาศัยตัวอย่างการแปลของคน ทำให้การแปลมีคุณภาพสูง ข้อเสียได้แก่ ประสิทธิภาพในการแปลขึ้นอยู่กับขนาดของฐานข้อมูล และขึ้นอยู่กับอัลกอริทึมในการจับคู่

2.1.1.5 การแปลโดยอาศัยสถิติ (Statistical Machine Translation : SMT) [8] จะมีการสร้างโมเดลทางสถิติจากคลังข้อความขนาน เพื่อช่วยในการแปล ข้อดีได้แก่ สามารถแก้ปัญหาเรื่องความกำกวมของคำศัพท์ และสำนวน ใช้ทรัพยากรมนุษย์น้อย ข้อเสียคือ ไม่สามารถจัดการเกี่ยวกับเรื่องโครงสร้างประโยคได้

2.1.1.6 การแปลแบบผสมผสานเทคนิค (Hybrid Machine Translation) [9] เป็นการนำหลาย ๆ แนวทางที่กล่าวมาแล้วมาประยุกต์ใช้ร่วมกัน

การแปลภาษาด้วยเครื่องยังมีปัญหาเรื่องประสิทธิภาพในการแปลอยู่มาก เนื่องจากเป้าหมายของการแปลโดยเครื่องคือ การแปลให้ใกล้เคียงกับการใช้คนแปลมากที่สุด ซึ่งนับว่าเป็นเรื่องที่ยาก เพราะเครื่องไม่สามารถแสดงความรู้สึกนึกคิดได้เหมือนคน ทำให้การแปลที่ได้ไม่สละสลวย สำหรับการแปลภาษาอังกฤษ-ไทย ในปัจจุบันยังมีปัญหาอยู่หลายประการ [10] ได้แก่ (ตัวอย่างการแปลจากเว็บไซต์ www.suparsit.com ซึ่งพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ)

1. การเลือกใช้คำแปลไม่ตรงกับบริบท เช่น “Until recently, the company was in the red.” ควรจะแปลว่า “จนกระทั่งเร็ว ๆ นี้บริษัทอยู่ในภาวะขาดทุน” แต่แปลได้เป็น “จนกระทั่งเมื่อเร็ว ๆ นี้บริษัทอยู่ในแดง”
2. การวางคำขยายผิดตำแหน่ง เช่น “He lives quite near here.” ควรแปลว่า “เขาอยู่ใกล้ที่นี่ทีเดียว” แต่แปลได้เป็น “เขาอยู่ทีเดียวใกล้นี่”
3. การเพิ่มคำหรือวลีที่ไม่จำเป็น เช่น “Overall sales exceeded our expectations due to the improving economy.” ควรจะแปลว่า “ยอดขายทั้งหมดเกินความคาดหมายเพราะว่าเศรษฐกิจดีขึ้น” แต่แปลได้เป็น “การขายที่ทั้งหมดกำลังเกินสิ่งที่คาดหวังในอนาคตของเราเพราะว่าเศรษฐกิจดีขึ้นทำให้”
4. การหายไปของคำบางคำ เช่น “He is standing still.” ควรจะแปลว่า “เขายืนนิ่ง” แต่แปลได้เป็น “เขานิ่ง”

2.1.2 **การลดรูปคำศัพท์** (Stemming)

การลดรูปคำศัพท์ เป็นกระบวนการในการลดรูปการผันคำ (Inflection) ให้กลับมาอยู่ในรูปทั่วไป (Stem) หรืออยู่ในรูปรากศัพท์ (Root) [11]

รากศัพท์ คือ หน่วยเล็กที่สุดของคำและมีความหมายในตัวเอง เช่น รากศัพท์ของคำว่า ‘destabilized‘ คือ ’stabil-‘

รูปทั่วไป คือ รากศัพท์รวมกับหน่วยคำที่เปลี่ยนหน้าที่ (Derivational Morphemes) แต่ไม่รวมหน่วยคำในการผันคำ (Inflectional Elements) เช่น รูปทั่วไปของคำว่า ‘destabilized’ คือ ‘destabilize’ จะเห็นได้ว่าประกอบด้วยรากศัพท์และหน่วยคำที่เปลี่ยนหน้าที่ได้แก่ ‘de-‘ และ‘–ize’ แต่ไม่รวมหน่วยคำในการผันคำ ได้แก่ ‘–(e)d’

การเปลี่ยนคำ (Derivation) เป็นการสร้างคำใหม่โดยการเติมหน่วยคำเปลี่ยนหน้าที่ ตัวอย่างการเปลี่ยนคำแสดงดังตารางที่ 2.1

ตารางที่ 2.1 การเปลี่ยนคำ

หน่วยคำเปลี่ยนหน้าที่	คำเดิม	คำใหม่	ตัวอย่าง
-ly	คำคุณศัพท์	กริยาวิเศษณ์	slow → slowly
-ness	คำคุณศัพท์	คำนาม	slow → slowness
-ize	คำคุณศัพท์	กริยา	modern → modernize
-al	คำนาม	คำคุณศัพท์	nation → national
-fy	คำนาม	กริยา	glory → glorify
-able	กริยา	คำคุณศัพท์	drink → drinkable
-ance	กริยา	คำนาม	apply → appliance

การผันคำ (Inflection) เป็นการเติมหน่วยคำในการผันคำเข้าไปในคำศัพท์เพื่อใช้แสดงข้อมูลทางไวยากรณ์ได้แก่ จำนวน (Number) บุรุษ (Person) กาล (Tense) และเพศ (Gender) ซึ่งบางคำผันโดยการเปลี่ยนรูปแทนการเติมหน่วยคำ ตัวอย่างการผันคำแสดงในตารางที่ 2.2 และ2.3

ตารางที่ 2.2 การผันคำแบบเติมหน่วยคำ

ชนิดของการผัน	หน่วยคำที่เติมเพื่อการผันคำ	ตัวอย่าง
ผันตามจำนวน	-s และ -es	dog → dogs
ผันตามกาล	-ed และ ing	I work → I'm working
ผันตามบุรุษ	-s และ -es	I walk → He walks
ผันเพื่อการเปรียบเทียบขั้นกว่าและขั้นสุด	-er และ –est	long → longer → longest

ตารางที่ 2.3 การผันคำแบบเปลี่ยนรูป

ชนิดของการผัน	ตัวอย่างการเปลี่ยนรูป
ผันตามจำนวน	foot → feet
ผันตามกาล	go → went → gone
ผันตามเพศ	actor → actress
ผันเพื่อการเปรียบเทียบขั้นกว่าและขั้นสุด	bad → worse → worst

2.1.3 **คำหยุด** (Stop Words)

คำหยุดเป็นกลุ่มคำที่มักจะถูกกรองออกก่อนจากการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เนื่องจากคำหยุดถูกมองว่าเป็นสัญญาณรบกวน (Signal Noise) ที่ทำให้ความสามารถในการประมวลผลลดลง ตัวอย่างเช่น ในโปรแกรมค้นหา (Search Engine) ถ้านำข้อความทั้งหมดที่ผู้ใช้ป้อนเข้ามา ไปทำการค้นหาในฐานข้อมูล ผลที่ได้จากการค้นหาจะไม่ตรงกับความต้องการเท่าที่ควร เนื่องจากคำหยุดมักจะเป็นคำที่พบบ่อยในทุกประโยคหรือทุกเอกสาร คำเหล่านี้ไม่มีประโยชน์ที่จะใช้ค้นหาหรือพิจารณา คำหยุดมักจะมีความหมายในตัวเองเพียงเล็กน้อยและไม่เฉพาะเจาะจง จึงไม่เหมาะที่จะใช้เป็นคำสำคัญในการค้นหา ตัวอย่างคำหยุดในภาษาอังกฤษ เช่น 'a' 'of' 'the' 'I' 'it' 'you' และ 'and' การกรองเอาคำหยุดออกจะช่วยให้การค้นหาชัดเจนมากขึ้นว่าผู้ใช้ต้องการค้นหาเกี่ยวกับเรื่องอะไร ผลที่ได้จากการค้นหาจะตรงกับความต้องการมากขึ้น อีกทั้งยังทำให้จำนวนดัชนีที่ใช้ในการค้นหาลดลงด้วย [11]

ในภาษาอังกฤษมีคำหยุดเป็นจำนวนหลายร้อยคำ (แสดงในภาคผนวก ก) โดยในงานวิจัยนี้จะขอเรียกคำหยุดกลุ่มนี้ว่า กลุ่มคำหยุดทั่วไป [12] งานแต่ละอย่างอาจสนใจกลุ่มคำหยุดต่างกัน

เช่น กลุ่มคำหยุดสำหรับคลังข้อความรอยเตอร์อาร์ซีวีหนึ่ง (Reuters-RCV1 : Reuters Corpus Volume 1) มี 25 คำ และกลุ่มคำหยุดที่ใช้ในกูเกิล [13] มี 35 คำ เป็นต้น ดังนั้นในงานวิจัยนี้จะใช้กลุ่มของคำหยุดเป็นพารามิเตอร์ตัวหนึ่งในการทดลอง โดยเลือกใช้กลุ่มคำหยุดสองแบบ ได้แก่ กลุ่มคำหยุดทั่วไปและกลุ่มคำหยุดที่ใช้ในกูเกิล ส่วนภาษาไทยเนื่องจากเป็นภาษาที่มีการใช้คำฟุ่มเฟือยและกลุ่มคำหยุดภาษาไทยมีจำนวนไม่มากนัก [14] (แสดงในภาคผนวก ก) จึงกำจัดทั้งหมด

2.1.4 **คำหน้าที่** (Function Words) **และคำเนื้อหา** (Content Words)

คำหน้าที่หรือคำไวยากรณ์เป็นคำที่มีความหมายในตัวเองเพียงเล็กน้อยหรือมีความหมายกำกวม แต่เป็นคำที่ใช้แสดงความสัมพันธ์ทางไวยากรณ์กับคำอื่น ๆ ในประโยค หรือเป็นคำที่แสดงถึงทัศนคติ ท่าทาง และอารมณ์ของผู้พูด [15] ตัวอย่างของคำหน้าที่ได้แก่

- คำนำหน้านาม (Article) : a an the
- คำสันธาน (Conjunction) : and or but after both…and either…or
- คำกริยานุเคราะห์ (Auxiliary Verb) : is am are was were do does
- คำสรรพนาม (Pronoun) : I you we they he she it him her
- คำบุพบท (Preposition) : on in at

ส่วนคำที่ตรงข้ามกับคำหน้าที่คือคำเนื้อหา ซึ่งเป็นคำที่มีความหมายเฉพาะเจาะจงไม่กำกวม ในพจนานุกรมสามารถระบุความหมายเฉพาะของคำเนื้อหาได้ [15] แต่สำหรับคำหน้าที่ในพจนานุกรมจะบอกถึงความหมายทั่วไปไม่เฉพาะเจาะจงหรือจะอธิบายการใช้ทั่วไปของคำนั้นมากกว่า เช่น คำว่า “an” ในพจนานุกรมอธิบายว่า “คำกำกับนามที่ขึ้นต้นด้วยเสียงสระ บ่งว่าเป็นเอกพจน์” แต่สำหรับหนังสือไวยากรณ์จะสามารถอธิบายการใช้คำหน้าที่ได้อย่างละเอียด และคำหยุดส่วนใหญ่มักจะเป็นคำหน้าที่

2.1.5 **การแปลงข้อมูลให้เป็นบรรทัดฐาน** (Data Normalization)

การแปลงข้อมูลให้เป็นบรรทัดฐานเป็นการปรับค่าของข้อมูลให้มีขอบเขตอยู่ในช่วงเล็กลง เช่น อยู่ในช่วง -1.0 ถึง 1.0 หรือ ช่วง 0.0 ถึง 1.0 เป็นต้น วิธีการแปลงข้อมูลให้เป็นบรรทัดฐานที่นิยมใช้กันอย่างแพร่หลายได้แก่ การแปลงตามค่าต่ำสุด-สูงสุด (Min-Max Normalization) การแปลงตามค่าคะแนนมาตรฐานซี (Z-Score Normalization) และการปรับมาตราทศนิยม (Decimal Scaling) [16] ในงานวิจัยนี้ เราไม่ทราบค่าสูงสุดและต่ำสุดของข้อมูลอนุกรมเวลา

ดังนั้นจึงเลือกใช้การแปลงข้อมูลให้เป็นบรรทัดฐานด้วยค่าคะแนนมาตรฐานซี เนื่องจากการแปลงข้อมูลด้วยวิธีนี้ใช้ได้ดีในกรณีที่ไม่ทราบค่าสูงสุดและต่ำสุดของข้อมูล

2.1.5.1 <u>การแปลงตามค่าคะแนนมาตรฐานซี</u> เป็นการแปลงค่าข้อมูลโดยปรับการกระจายของข้อมูลให้มีค่าเฉลี่ยเท่ากับ 0 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 1 การคำนวณหาค่าคะแนนมาตรฐานซีหาได้จากสูตร

$$Z_i = \frac{X_i - Mean}{SD}$$

โดยที่ Z_i คือ ค่าบรรทัดฐานของอนุกรมเวลา X ที่ตำแหน่ง i

$Mean$ คือ ค่าเฉลี่ยเลขคณิตของอนุกรมเวลา X

SD คือ ค่าเบี่ยงเบนมาตรฐานของอนุกรมเวลา X

2.1.6 **ข้อมูลอนุกรมเวลา** (Time Series Data)

ข้อมูลอนุกรมเวลาเป็นข้อมูลประเภทหนึ่ง ซึ่งประกอบด้วยลำดับของค่าหรือเหตุการณ์ที่เปลี่ยนแปลงตามเวลา โดยค่านั้นจะวัดที่ระยะห่างของเวลาที่เท่ากันหรือไม่เท่ากันก็ได้ เช่น ค่าของดัชนีหุ้นที่เปลี่ยนแปลงไป เป็นต้น ข้อมูลอนุกรมเวลาได้ถูกนำไปใช้ในหลายงาน เช่น การศึกษาเกี่ยวกับความผันผวนของตลาดหุ้นในแต่ละวัน การทดลองด้านวิทยาศาสตร์ การรักษาทางการแพทย์ เป็นต้น เหมืองข้อมูลอนุกรมเวลาสามารถนำมาใช้ได้หลายด้าน เช่น การวิเคราะห์แนวโน้ม (Trend Analysis) การวัดความเหมือนกัน (Similarity Measurement) การทำเหมืองเพื่อหาแบบอย่าง (Pattern Mining) [16]

การวัดความเหมือนกันของข้อมูลอนุกรมเวลา สามารถวัดได้จากฟังก์ชันระยะห่าง โดยหาระยะห่างระหว่างจุดที่อยู่ในลำดับ ลำดับคู่ใดที่คำนวณแล้วได้ผลรวมของระยะห่างน้อยที่สุดจะถือว่ามีความเหมือนกันมากที่สุด ฟังก์ชันระยะห่างที่นิยมใช้ได้แก่

1. ระยะห่างแมนฮัตตัน (Manhattan Distance) สูตรที่ใช้คือ

$$D(X,Y) = \sum_{i=1}^{n} |X_i - Y_i|$$

โดยที่ $D(X,Y)$ คือ ระยะห่างแมนฮัตตันระหว่างอนุกรมเวลา $X = X_1X_2...X_n$ และ $Y = Y_1Y_2...Y_n$

2. ระยะห่างยุคลิเดียน (Euclidean Distance) สูตรที่ใช้คือ

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

โดยที่ $D(X,Y)$ คือ ระยะห่างยุคลิเดียนระหว่างอนุกรมเวลา $X = X_1X_2...X_n$ และ $Y = Y_1Y_2...Y_n$

3. ไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping) เป็นอัลกอริทึมที่ใช้วัดความเหมือนกันระหว่างสองลำดับที่อาจจะมีความต่างกันในเรื่องของเวลาและความเร็ว เช่น ใช้ตรวจจับรูปแบบการเดินที่เหมือนกันได้แม้ว่าคนจะเดินเร็วหรือช้า เป็นต้น ซึ่งการคำนวณแบบนี้จะใช้เวลามาก เนื่องจากเป็นกำหนดการพลวัต (Dynamic Programming) นิยามของไดนามิกไทม์วอร์ปปิงแสดงได้ดังนี้ [17]

ให้ข้อมูล $X = X_1X_2...X_n$ และ $Y = Y_1Y_2...Y_m$ โดยจะสามารถนิยามระยะทาง ไดนามิกไทม์วอร์ปปิงเป็นสมการเวียนเกิดได้ดังต่อไปนี้

$$\begin{aligned} DTW(X,Y) &= \gamma(n,m) \\ \gamma(i,j) &= D(X_i,Y_j) + \min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\} \\ D(X_i,Y_j) &= (Y_i - X_j)^2 \end{aligned}$$

โดยที่ γ คือ ระยะทางสะสม

D คือ ฟังก์ชันหาระยะทางระหว่างจุดสองจุด และ

$1 \le i \le n$ และ $1 \le j \le m$

2.1.7 การหาค่าเฉลี่ยของส่วนกลับลำดับชั้น (Mean Reciprocal Rank)

ค่าส่วนกลับลำดับชั้น (Reciprocal Rank) คือ ค่าเศษหนึ่งส่วนลำดับชั้น [11] แสดงได้ดังสูตรต่อไปนี้ โดย i = ลำดับชั้น มีค่าตั้งแต่ 1 ถึง n

$$\text{Reciprocal Rank} = \frac{1}{i}$$

ค่าเฉลี่ยของส่วนกลับลำดับชั้นเป็นค่าสถิติสำหรับประเมินการประมวลผลใด ๆ ที่ให้ผลลัพธ์หลายค่าโดยแต่ละค่าเรียงตามลำดับความน่าจะเป็นของความถูกต้อง คำนวณได้โดยใช้สูตรดังนี้

$$\text{Mean Reciprocal Rank} = \frac{\sum_{i=1}^{n}\left(\frac{1}{i}\times \text{Number of elements in rank}_i\right)}{\text{Number of all elements}}$$

สำหรับงานวิจัยนี้ ค่าส่วนกลับลำดับชั้นจะถูกนำไปใช้ในการให้คะแนนคู่ประโยค ซึ่งจะกล่าวถึงรายละเอียดในหัวข้อ 3.5.1 ส่วนค่าเฉลี่ยของส่วนกลับลำดับชั้นจะใช้ในการคำนวณความถูกต้องของการจับคู่คำ ซึ่งจะกล่าวถึงรายละเอียดในหัวข้อ 3.4.2

2.2 งานวิจัยที่เกี่ยวข้อง

ในปัจจุบันมีงานวิจัยเกี่ยวกับการหาวิธีจับคู่ข้อความอยู่มากมาย ทั้งการจับคู่ในระดับคำ วลี ประโยค และย่อหน้า การจับคู่โดยใช้ความยาวของประโยคก็เป็นวิธีหนึ่ง เช่น ประโยคที่สั้นมีแนวโน้มที่จะจับคู่กับประโยคความหมายที่สั้น เป็นต้น ตำแหน่งของประโยคก็มีความสำคัญ เช่น ประโยคแรกมีโอกาสน้อยมากที่จะจับคู่กับประโยคสุดท้ายในข้อความอีกภาษาหนึ่ง [1, 18] การจับคู่ประโยคโดยอาศัยความยาวและตำแหน่งของประโยคมีข้อดีคือ ไม่จำเป็นต้องอาศัยความรู้ทางโครงสร้างและความหมายของภาษา อย่างไรก็ตาม การอาศัยความยาวของประโยคในการจับคู่จะเหมาะสำหรับภาษาที่มีความยาวใกล้เคียงกันแต่ไม่เหมาะกับภาษาที่มีความยาวแตกต่างกันมาก ๆ เช่น ภาษาอังกฤษและภาษาไทย ส่วนใหญ่ข้อความภาษาไทยจะมีจำนวนคำมากกว่าข้อความภาษาอังกฤษเพราะภาษาไทยมักใช้คำฟุ่มเฟือย

นอกจากนี้ยังมีการเสนอวิธีการจับคู่อีกแบบอื่น ๆ ได้แก่ การใช้คำร่วมเชื้อสาย (Cognate) โดยคำร่วมเชื้อสายอาจจะเป็นสัญลักษณ์หรือคำที่สะกดใกล้เคียงกันในสองภาษาเช่น คำว่า “langage” ในภาษาฝรั่งเศสและ “language” ในภาษาอังกฤษ [19] เทคนิคนี้จะเหมาะสมกับภาษาที่อยู่ในตระกูลเดียวกัน แต่ไม่เหมาะกับภาษาที่อยู่ต่างตระกูลกัน เช่น ภาษาอังกฤษกับภาษาไทย ภาษาอังกฤษกับภาษาจีน และภาษาอังกฤษกับภาษาญี่ปุ่น เป็นต้น เนื่องจากคำในภาษาเหล่านี้ใช้ตัวอักษรคนละแบบ

อีกวิธีหนึ่งที่ใช้ในการจับคู่ข้อความคือ การใช้คำที่ระบุถึงสิ่งเดียวกัน (Anchor Word) ซึ่งคำที่ระบุถึงสิ่งเดียวกันอาจจะได้มาจากการใช้พจนานุกรม [20] หรือจากกระบวนการจับคู่คำ [21] โดยเทคนิคที่ใช้ในการจับคู่คำเช่น Stochastic Inversion Transduction Grammars [22] และการใช้สถิติร่วมกับความรู้ทางภาษาศาสตร์ [23] เป็นต้น

จากการศึกษางานวิจัยพบว่าการใช้สถิติจะเหมาะสำหรับกรณีที่คำศัพท์มีความถี่ในการปรากฏมาก ๆ และยังใช้ได้ดีแม้ว่าจะตัดคำผิด แต่มีข้อจำกัดคือไม่เหมาะกับคำที่มีความถี่ในการปรากฏน้อย ในขณะที่การใช้พจนานุกรมจะไม่มีปัญหาเกี่ยวกับเรื่องความถี่ในการปรากฏ แม้ว่าคำนั้นจะปรากฏเพียงครั้งเดียวก็สามารถจับคู่โดยใช้พจนานุกรมได้ แต่การใช้พจนานุกรมมีข้อจำกัดเนื่องจากเราไม่ทราบบริบทของคำศัพท์ ทำให้เป็นการยากที่จะเลือกความหมายได้ตรงกับบริบท เพราะแต่ละคำสามารถมีความหมายได้หลายอย่าง และถ้าตัดคำผิดก็อาจทำให้หาคำศัพท์นั้นในพจนานุกรมไม่พบ

ความถูกต้องในการจับคู่ข้อความในคลังข้อความขนาน นอกจากจะขึ้นอยู่กับอัลกอริทึมในการจับคู่แล้ว ยังขึ้นอยู่กับปัจจัยอื่น ๆ ด้วยเช่น ความถูกต้องในการตัดคำ ความถูกต้องในการตัดประโยค เนื่องจากบางภาษาเช่น ภาษาไทย มีการเขียนคำต่อเนื่องกันไปไม่มีการเว้นวรรคระหว่าง

คำ อีกทั้งไม่มีเครื่องหมายจบประโยคเหมือนภาษาอังกฤษ ดังนั้นถ้าประสิทธิภาพในการตัดคำและตัดประโยคไม่ดีพอ ก็ย่อมมีผลกระทบต่อการจับคู่ข้อความในคลังข้อความขนาน และสำหรับภาษาไทยมีงานวิจัยที่เกี่ยวข้องกับการตัดคำอยู่เป็นจำนวนมาก [24, 25] แต่งานวิจัยเกี่ยวกับการตัดประโยคยังมีไม่มากนัก [26, 27]

ส่วนงานวิจัยด้านการจับคู่ข้อความในคลังข้อความขนานภาษาอังกฤษและภาษาไทย ยังมีไม่มากนัก ตัวอย่างเช่น Asanee Kawtrakul และ Prachya Boonkwan [28] นำเสนอวิธีการจับคู่คำและจับคู่วลี โดยอาศัยปัจจัยสองส่วนได้แก่ ข้อมูลวงกว้าง (Global Information) และข้อมูลท้องถิ่น (Local Information) สำหรับข้อมูลวงกว้างพิจารณาว่าคำในภาษาอังกฤษ ควรจะมีการกระจายตัวตลอดทั้งเอกสารเหมือนกับคำในภาษาไทยที่จะเป็นคู่กัน โดยใช้เคเวกเตอร์อัลกอริทึม (K – Vector Algorithm) ทำการสร้างเวกเตอร์ที่แทนการกระจายความถี่ของแต่ละคำในภาษาอังกฤษและภาษาไทย หลังจากนั้นจึงหาความเหมือนกันของเวกเตอร์ของคำภาษาอังกฤษหนึ่งคำกับเวกเตอร์ของคำภาษาไทยทุกคำ โดยใช้สมการโคไซน์ (Cosine) ซึ่งผลที่ได้จะเป็นกลุ่มของคำภาษาไทยที่มีความเป็นไปได้ว่าจะเป็นคู่กับคำภาษาอังกฤษที่กำลังพิจารณา (Candidate Word Pairs) และเพื่อให้การจับคู่คำมีความถูกต้องมากขึ้น ผู้วิจัยได้ทำการพิจารณาข้อมูลท้องถิ่นประกอบด้วย คำในภาษาอังกฤษและภาษาไทยที่เป็นคู่กันต้องปรากฏในประโยคภาษาอังกฤษและภาษาไทยที่เป็นคู่กันเท่านั้น ซึ่งผู้วิจัยทำการจับคู่ประโยคโดยการใช้คะแนนสองชนิดคือ คะแนนจากคำที่รู้ความหมาย และคะแนนออฟเซตหรือความเฉของตำแหน่งคู่ประโยค โดยแต่ละคู่ประโยคภาษาอังกฤษและภาษาไทยจะได้คะแนนบวกเมื่อเปิดพจนานุกรมและพบคำที่เป็นความหมายของกันและกันในคู่ประโยคนั้น แต่จะถูกหักคะแนนตามออฟเซต ถ้าออฟเซตต่างกันมากก็ถูกหักคะแนนมากนั่นเอง ส่วนการจับคู่วลีก็ใช้หลักการเดียวกับการจับคู่คำที่กล่าวมา

บทที่ 3

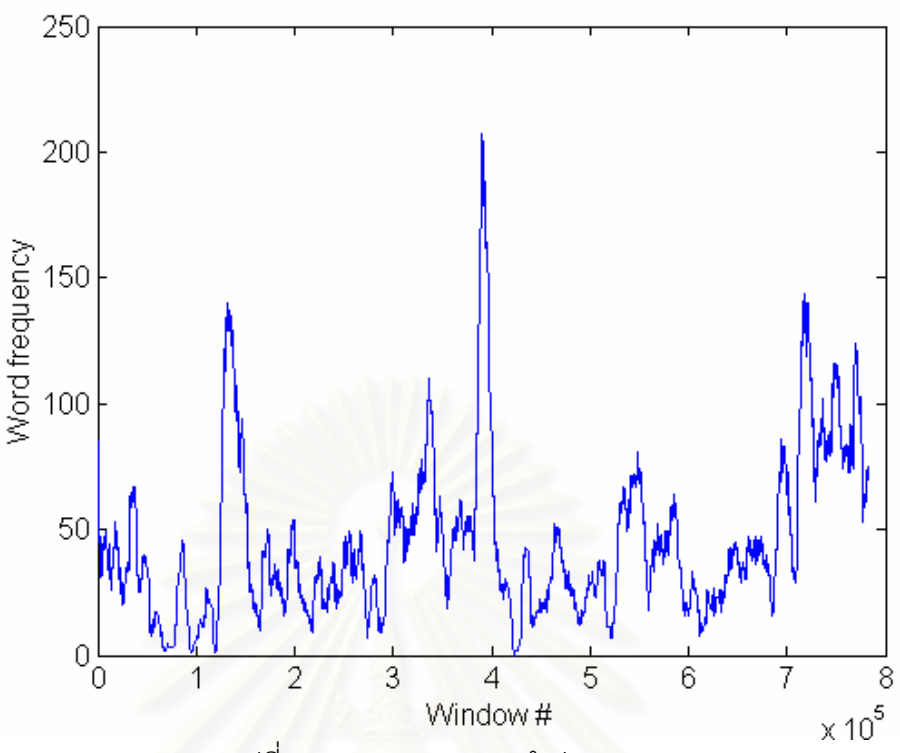
ขั้นตอนการดำเนินงานวิจัย

ในคลังข้อความขนาน แต่ละคำในภาษาหนึ่งควรจะมีความถี่และตำแหน่งการเกิดคำนั้น ๆ ใกล้เคียงกับคำที่เป็นคู่กันในอีกภาษาหนึ่ง ซึ่งคุณสมบัตินี้เป็นประโยชน์อย่างมากกับการจับคู่ส่วนที่ตรงกัน โดยการนับความถี่ของการปรากฏของแต่ละคำภายในหน้าต่างเลื่อน (Sliding Window) หน้าต่างเลื่อนจะเริ่มต้นที่คำแรกของข้อความและเลื่อนไปครั้งละหนึ่งคำจนจบข้อความ ผลลัพธ์ที่ได้จะเป็นลำดับของตัวเลขซึ่งคืออนุกรมเวลานั่นเอง โดยจะกล่าวถึงอีกครั้งในหัวข้อ 3.3 อย่างไรก็ตามในแต่ละภาษาที่ต่างกันความยาวของข้อความย่อมไม่เท่ากันทำให้ความยาวของอนุกรมเวลาไม่เท่ากันตามไปด้วย ดังนั้นจึงต้องปรับความยาวของอนุกรมเวลาให้เท่ากันก่อนเพื่อให้สะดวกในการเปรียบเทียบ หลังจากนั้นจึงวัดความเหมือนกันของอนุกรมเวลาโดยใช้ฟังก์ชันระยะห่าง คำคู่ใดที่คำนวณได้ระยะห่างน้อยที่สุดก็จะถูกจับคู่กัน

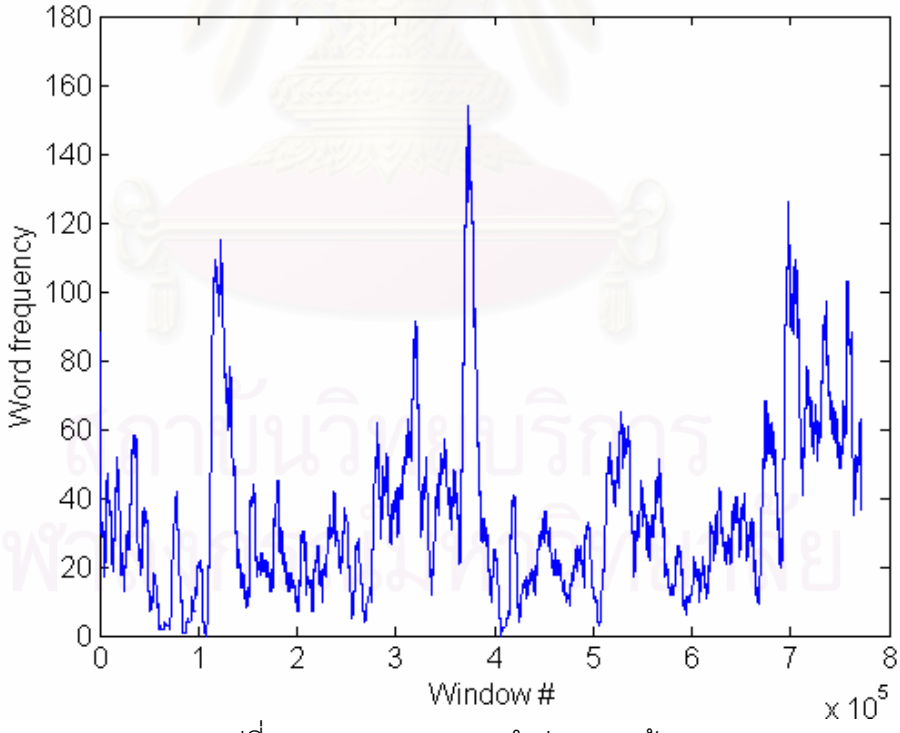
ถ้าเราสร้างอนุกรมเวลาของทุกคำในคลังข้อความขนาน แต่ละคำควรจะได้อนุกรมเวลาที่ไม่ซ้ำกัน และอนุกรมเวลาของคำที่เป็นคู่กันในสองภาษาที่แตกต่างกันควรจะมีรูปร่างที่คล้ายกัน เพราะมีตำแหน่งการเกิดและจำนวนความถี่ของคำนั้น ๆ ใกล้เคียงกัน จะเห็นตัวอย่างได้จากรูปที่ 3.1 และ 3.2 ซึ่งได้จากการสร้างอนุกรมเวลาของคำว่า “God” และ “พระเจ้า” จากคัมภีร์ไบเบิลฉบับภาษาอังกฤษและภาษาไทย ตามลำดับ

จากรูปที่ 3.1 และรูปที่ 3.2 พบว่าอนุกรมเวลาทั้งสองมีรูปร่างที่คล้ายกันมาก ส่วนรูปที่ 3.3 แสดงอนุกรมเวลาของคำว่า “bird” ที่ไม่ได้มีความหมายว่าพระเจ้า จะพบว่ารูปร่างแตกต่างจากรูปที่ 3.1 และ 3.2 เมื่อได้คำที่ระบุถึงสิ่งเดียวกันแล้วจึงนำไปใช้ในการจับคู่ประโยคที่ตรงกัน โดยถ้าประโยคคู่ใดมีคำที่ระบุถึงสิ่งเดียวกันมากที่สุด ประโยคนั้นจะถูกพิจารณาให้เป็นคู่กัน จากการให้คะแนนคู่ประโยค

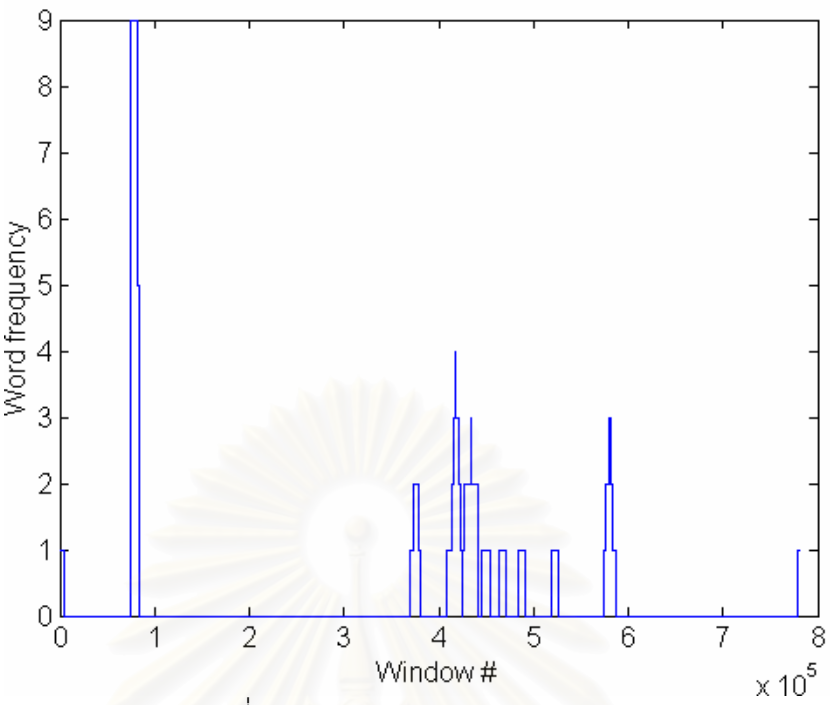
อย่างไรก็ตาม ก่อนที่จะทำการจับคู่คำและจับคู่ประโยคได้ จะต้องมีการเตรียมข้อมูลทั้งภาษาอังกฤษและภาษาไทยให้เหมาะสมกับการสกัดอนุกรมเวลาด้วย ซึ่งจะกล่าวถึงในหัวข้อ 3.2 และการจับคู่ให้ได้ผลดีนั้นจะต้องมีการปรับพารามิเตอร์ต่าง ๆ ให้เหมาะสมด้วย ซึ่งจะกล่าวถึงในหัวข้อ 4.3



รูปที่ 3.1 อนุกรมเวลาของคำว่า “God”



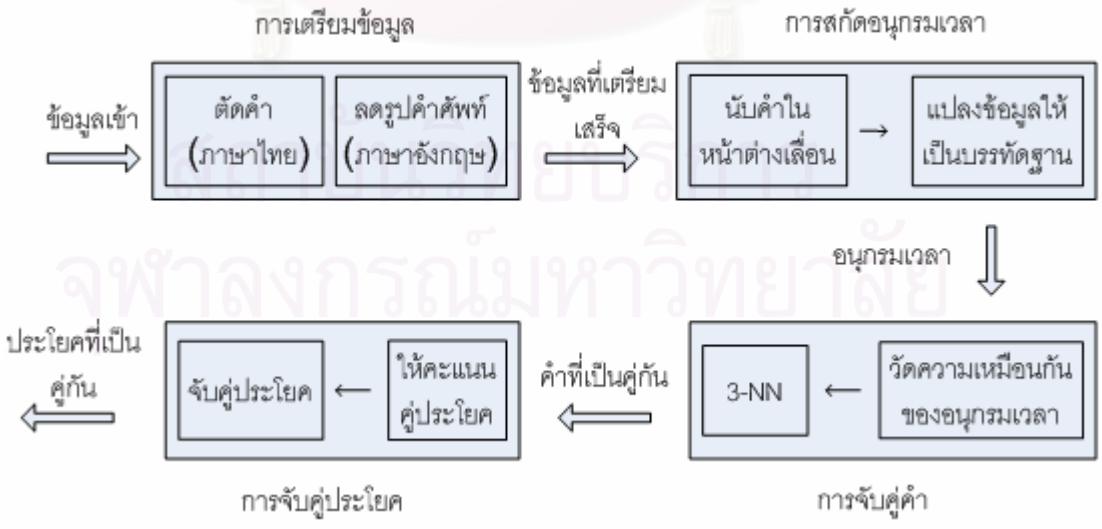
รูปที่ 3.2 อนุกรมเวลาของคำว่า “พระเจ้า”



รูปที่ 3.3 อนุกรมเวลาของคำว่า “bird”

3.1 แผนภาพการทำงาน

การดำเนินงานจะแบ่งเป็น 4 ส่วนได้แก่ การเตรียมข้อมูล (Data Preparation) การสกัดอนุกรมเวลา (Time Series Extraction) การจับคู่คำที่ตรงกัน และการจับคู่ประโยคที่ตรงกัน โดยแสดงได้ดังรูปที่ 3.4



รูปที่ 3.4 แผนภาพแสดงการดำเนินงาน

3.2 การเตรียมข้อมูล

3.2.1 การเตรียมข้อมูลภาษาอังกฤษ

3.2.1.1 การลดรูปคำศัพท์ เนื่องจากภาษาอังกฤษมีการผันคำตามบุรุษ จำนวน และเวลา ดังนั้นจึงต้องมีการแปลงคำที่ถูกผันให้กลับมาอยู่ในรูปรากศัพท์เดิม โดยการตัด -s -es -d -ed -ing ฯลฯ ออก การเปลี่ยนรูปของคำกริยาในภาษาอังกฤษจากช่อง 1 เป็นช่อง 2 และ 3 นอกจากจะเกิดจากการเติม -d หรือ -ed แล้วยังอาจเกิดจากการเปลี่ยนรูปหรือที่เรียกว่า irregular verb เช่น eat ate eaten เป็นต้น ซึ่งคำกริยาเหล่านี้ต้องแปลงกลับมาเป็นกริยาช่อง 1 ด้วยเช่นกัน ส่วนคำที่เติมหน่วยคำเติมหน้า (prefix) และหน่วยคำเติมหลัง (suffix) เพื่อสร้างให้ได้เป็นคำชนิดใหม่ เช่น การเติม -ous เพื่อสร้างคำคุณศัพท์ (adjective) ต้องมีการตัดหน่วยคำเติมเหล่านี้เพื่อแปลงกลับให้อยู่ในรูปรากศัพท์เดิมเช่นกัน เหตุที่ต้องมีการเปลี่ยนคำให้กลับมาอยู่ในรูปรากศัพท์เนื่องจากคำในภาษาไทยมิได้มีการผันเป็นหลายรูปเหมือนในภาษาอังกฤษ ดังนั้นไม่ว่าภาษาอังกฤษคำนั้นจะอยู่ในรูปใด เมื่อแปลเป็นภาษาไทยจะได้ความหมายเดียวกัน เช่น go went gone แปลเป็นคำว่า “ไป” เป็นต้น

โดยงานวิจัยนี้จะใช้โปรแกรม KTAGGER [3] ซึ่งเป็นโปรแกรมสำหรับการลดรูปคำศัพท์ในภาษาอังกฤษให้อยู่ในรูปทั่วไปหรือรากศัพท์ ซึ่งสามารถเลือกได้ว่าต้องการให้ผลลัพธ์ที่ได้แสดงในรูปแบบใด เช่น แสดงในรูปแบบ tab-delimited รูปแบบมาตรฐาน หรือ รูปแบบ SGML markup โดยขึ้นอยู่กับไฟล์ควบคุมที่เลือกใช้ คำสั่งที่ใช้ในโปรแกรม KTAGGER คือ ktagger.exe -x <sgml.ctl> -i <input.txt> -o <output.txt>

3.2.1.2 การจัดการกับเครื่องหมายต่าง ๆ ในประโยค บางตัวต้องมีการตัดทิ้ง และบางตัวต้องมีการเปลี่ยนไปเป็นเครื่องหมายอื่นแทน

- เครื่องหมายที่ต้องตัดทิ้ง ได้แก่ / | ‘ “ ! _ { } [ ] ( ) < > ; : - , การตัดเครื่องหมายเหล่านี้เพื่อป้องกันความสับสน เช่น “Hello!” และ “Hello” ถือเป็นคำเดียวกัน แต่การที่มีเครื่องหมายอัศเจรีย์อยู่ในคำด้วยทำให้โปรแกรมเข้าใจว่าเป็นคนละคำกันและสร้างอนุกรมเวลา 2 อนุกรม ซึ่งมีผลกระทบกับกระบวนการจับคู่คำ
- เครื่องหมายที่ต้องเปลี่ยนเป็นมหัพภาค (Full Stop) คือ เครื่องหมายคำถาม (Question Mark) เนื่องจากมหัพภาคแทนการจบประโยค ซึ่งคำถามถือเป็นประโยคด้วยเช่นกัน

3.2.1.3 การแบ่งย่อหน้า เนื่องจาก word processing ใช้การกดปุ่ม enter เป็นการขึ้นย่อหน้าใหม่ ดังนั้นในงานวิจัยนี้จะถือว่าการกดปุ่ม enter เป็นการขึ้นย่อหน้าใหม่ด้วยเช่นกัน เนื่องจากในภาษาอังกฤษมีการเว้นวรรคระหว่างคำ และแสดงการจบประโยคด้วยมหัพภาคหรือเครื่องหมายคำถามอยู่แล้ว ดังนั้นจึงไม่จำเป็นต้องตัดคำ และไม่ต้องเพิ่มเครื่องหมายวรรคตอนเพื่อบอกขอบเขตของวลีหรือประโยคเช่นในภาษาไทย ส่วนมหัพภาคที่เป็นส่วนหนึ่งของคำย่อเช่น U.S. จะไม่ถูกวิเคราะห์เป็น 2 ประโยคเพราะในโปรแกรมใช้วิธีการอ่านข้อความเข้ามาทีละส่วน ซึ่งแต่ละส่วนแบ่งแยกด้วยช่องว่าง (Space) แต่คำว่า U.S. ไม่มีช่องว่างระหว่างคำจึงถูกอ่านเข้ามาเป็นส่วนเดียวกันและถูกพิจารณาเป็นประโยคเดียวกัน

3.2.2 การเตรียมข้อมูลภาษาไทย

3.2.2.1 การตัดคำ เนื่องจากข้อความภาษาไทยเขียนติดกันโดยไม่จำเป็นต้องมีช่องว่างระหว่างคำเหมือนในภาษาอังกฤษ ดังนั้นจึงต้องนำข้อความเหล่านี้ไปตัดคำก่อน เพื่อความสะดวกในการจับคู่คำ โดยในงานวิจัยนี้ใช้โปรแกรมตัดคำ SWATH [2] ซึ่งพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค) โดยเหตุที่เลือกใช้โปรแกรมนี้เนื่องจากวิธีใช้ไม่ยุ่งยากและมีความแม่นยำในการตัดคำค่อนข้างสูง

3.2.2.2 การจัดการกับเครื่องหมายต่าง ๆ ในประโยค เครื่องหมายที่ต้องตัดทิ้งได้แก่ / | ‘ “ ! _ { } [ ] ( ) < > ; : - , โดยมีเหตุผลการตัดทิ้งเช่นเดียวกับที่กล่าวไปแล้วในภาษาอังกฤษ

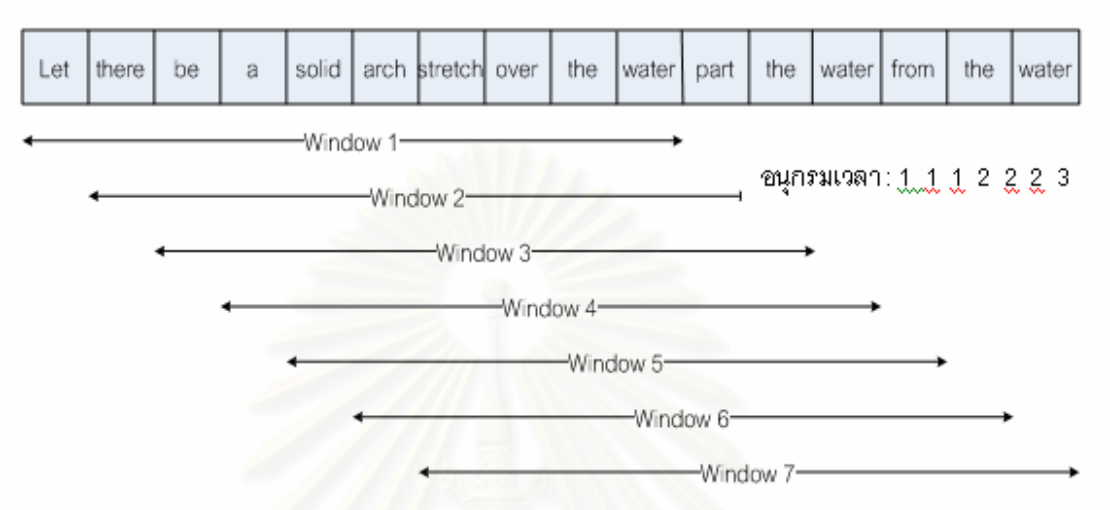
3.2.2.3 การแบ่งวรรคตอน ในภาษาไทยไม่มีการใช้เครื่องหมายวรรคตอนเมื่อจบประโยค แต่มีการใช้ช่องว่างเพื่อแบ่งวรรค โดยใน 1 ประโยคของภาษาไทยอาจจะประกอบด้วยวรรคเดียวหรือหลายวรรค ซึ่งในงานวิจัยนี้จะทดลองทั้งแบบจับคู่ 1 ประโยคภาษาอังกฤษกับ 1 ประโยคภาษาไทย และแบบจับคู่ 1 ประโยคภาษาอังกฤษกับ N วรรคในภาษาไทย

3.2.2.4 การแบ่งย่อหน้า ถือว่าการกดปุ่ม Enter ในภาษาไทยเป็นการขึ้นย่อหน้าใหม่เช่นเดียวกับในภาษาอังกฤษ

3.3 การสกัดอนุกรมเวลาจากคลังข้อความขนาน

หัวใจของวิธีที่นำเสนอนี้คือ การสกัดอนุกรมเวลาของแต่ละคำจากคลังข้อความขนาน การสกัดอนุกรมเวลาทำได้โดยการนับความถี่ของการปรากฏของแต่ละคำภายในหน้าต่างเลื่อน (Sliding Window) โดยหน้าต่างเลื่อนจะเริ่มต้นที่คำแรกของข้อความและเลื่อนไปครั้งละหนึ่งคำจนจบข้อความ (รูปที่ 3.5 แสดงตัวอย่างการสกัดอนุกรมเวลาของคำว่า “water”) ลำดับของตัวเลขที่

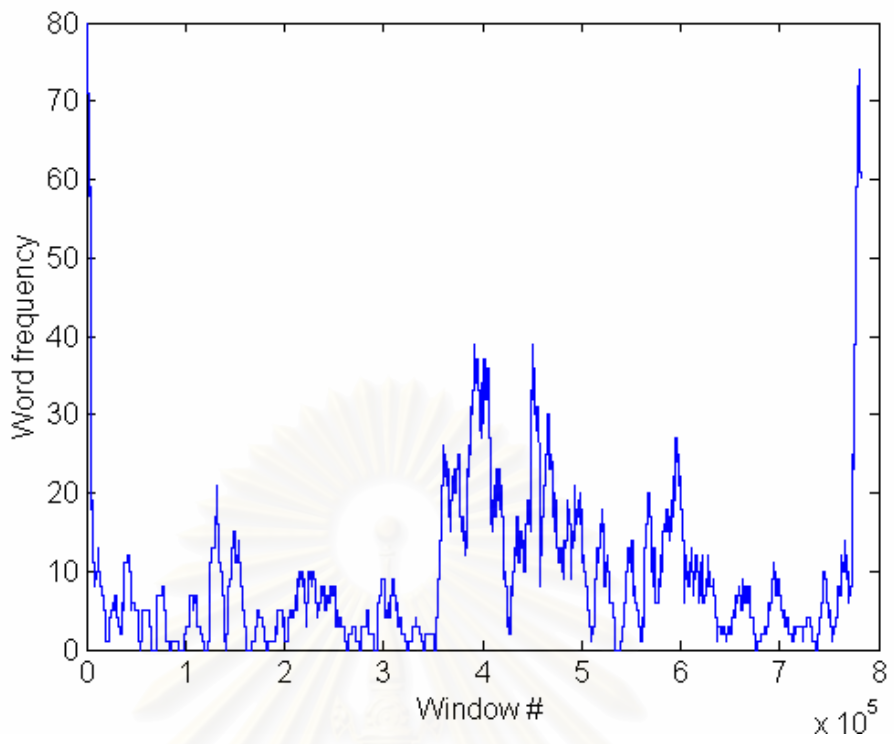
ได้คือข้อมูลอนุกรมเวลา หากนำข้อมูลชุดนี้มาวาดเป็นกราฟจะได้กราฟที่รูปร่างแตกต่างกัน ดังตัวอย่างในรูปที่ 3.1 ถึง 3.3 โดยแกน X จะแทนลำดับที่ของหน้าต่าง และแกน Y จะแทนจำนวนความถี่ที่ปรากฏคำนั้น ๆ ในแต่ละหน้าต่าง



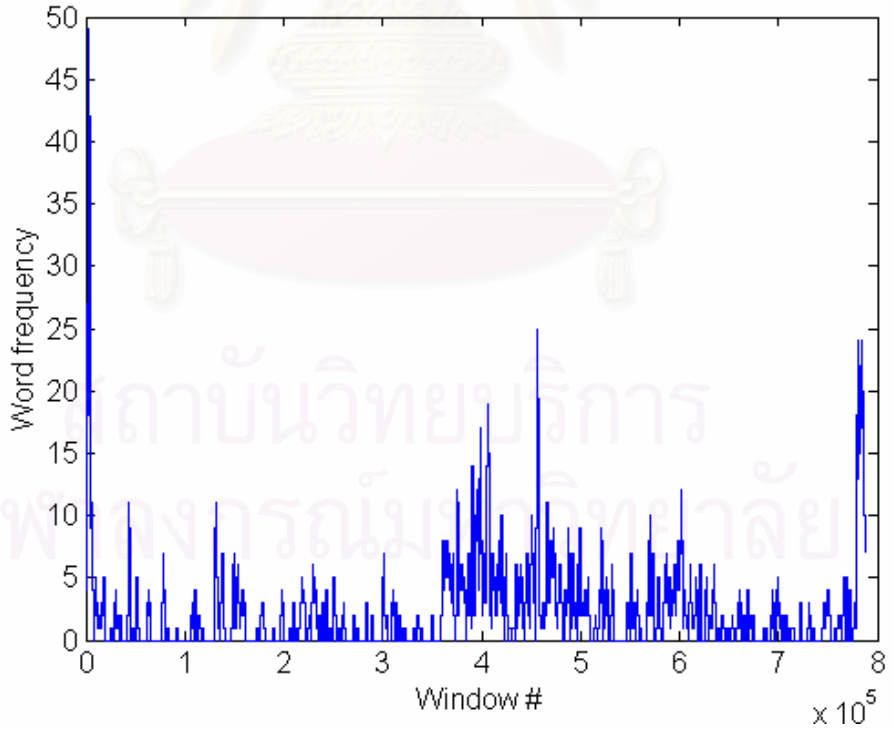
รูปที่ 3.5 ตัวอย่างการสกัดอนุกรมเวลาของคำว่า “water”

เนื่องจากความยาวของข้อความภาษาอังกฤษและภาษาไทยต่างกัน ทำให้ความยาวของอนุกรมเวลาต่างกัน ดังนั้นเพื่อให้สะดวกต่อการวัดความเหมือนกันของอนุกรมเวลา จึงทำการปรับลดความยาวของอนุกรมเวลาที่ยาวกว่าให้เท่ากับอนุกรมที่สั้นกว่า

นอกจากนี้ขนาดของหน้าต่าง (Window Size) เป็นอีกค่าที่มีความสำคัญ เพราะมีผลโดยตรงกับลักษณะของอนุกรมเวลาที่ได้รับ โดยถ้าใช้หน้าต่างขนาดเล็กจะทำให้อนุกรมเวลาที่ได้มีสัญญาณรบกวนมากกว่าการใช้หน้าต่างขนาดใหญ่ซึ่งให้อนุกรมเวลาที่มีความเรียบมากกว่า ดังแสดงในรูปที่ 3.6 และ 3.7 ซึ่งเป็นกราฟอนุกรมเวลาของคำว่า “earth” จากคัมภีร์ไบเบิลเหมือนกัน แต่ใช้ขนาดหน้าต่างไม่เท่ากัน อนุกรมเวลาที่มีสัญญาณรบกวนมากเกินไป จะทำให้ผลการวัดความเหมือนกันของอนุกรมเวลาคลาดเคลื่อนได้ แต่การใช้หน้าต่างขนาดใหญ่เกินไปก็มีผลต่อความถูกต้องในการจับคู่คำเช่นกัน ซึ่งรายละเอียดเกี่ยวกับการใช้หน้าต่างขนาดต่าง ๆ จะกล่าวถึงอีกครั้งในหัวข้อ 4.3.1



รูปที่ 3.6 อนุกรมเวลาของคำว่า “earth” โดยใช้หน้าต่างขนาด 8000 คำ



รูปที่ 3.7 อนุกรมเวลาของคำว่า “earth” โดยใช้หน้าต่างขนาด 2000 คำ

จากรูปที่ 3.6 จะเห็นว่าค่าความถี่ของคำที่แสดงในแนวแกน Y มีขอบเขตที่กว้างมาก การแปลงข้อมูลให้เป็นบรรทัดฐานจะช่วยให้ค่าข้อมูลในแนวแกน Y อยู่ในช่วงที่แคบลงได้ ในงานวิจัยนี้จะทำการแปลงค่าข้อมูลตามคะแนนมาตรฐานซี ดังที่กล่าวไปแล้วในหัวข้อ 2.1.5 เพราะการแปลงข้อมูลด้วยวิธีนี้ใช้ได้ดีในกรณีที่ไม่ทราบค่าสูงสุดและต่ำสุดของข้อมูล

3.4 การจับคู่คำที่ตรงกันในคลังข้อความขนาน

3.4.1 การลดขนาดของอนุกรมเวลา

การลดขนาดอนุกรมเป็นการทำให้อนุกรมเวลามีขนาดสั้นลง โดยยังคงลักษณะเด่นของข้อมูล ในงานวิจัยนี้วัดความเหมือนกันของอนุกรมเวลาจากฟังก์ชันระยะห่างซึ่งเป็นการคำนวณระยะห่างแบบจุดต่อจุด ดังนั้นจึงต้องปรับค่าในแนวแกน X ของอนุกรมของทั้งสองภาษาให้มีความยาวเท่ากัน ซึ่งในงานวิจัยนี้จะทำการลดขนาดอนุกรมด้วยวิธีง่าย ๆ เพียงหาผลต่างระหว่างความยาวของอนุกรมเวลา เพื่อให้ทราบว่าจะต้องกำจัดจุดจากอนุกรมที่ยาวกว่าเป็นจำนวนกี่จุด

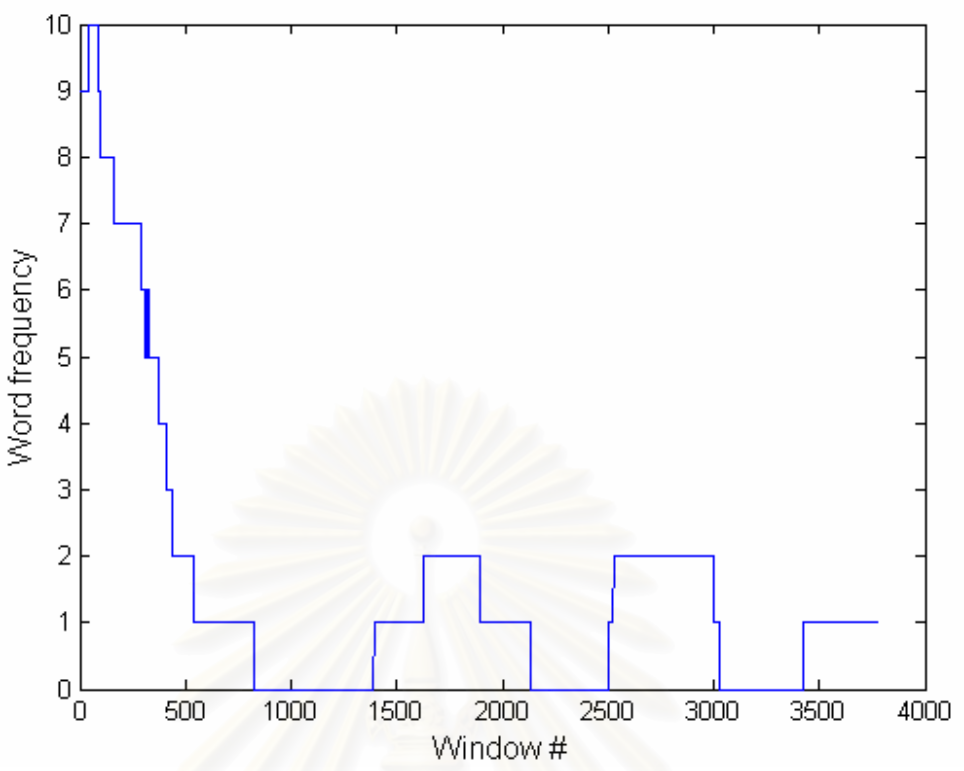
จำนวนจุดที่ต้องกำจัด = $\left|$ ความยาวอนุกรมเวลา$_1$ – ความยาวอนุกรมเวลา$_2$ $\right|$

โดยตำแหน่งที่จะถูกกำจัดออกได้แก่ ตำแหน่ง p ที่ (p mod a) = 0 กำหนดให้

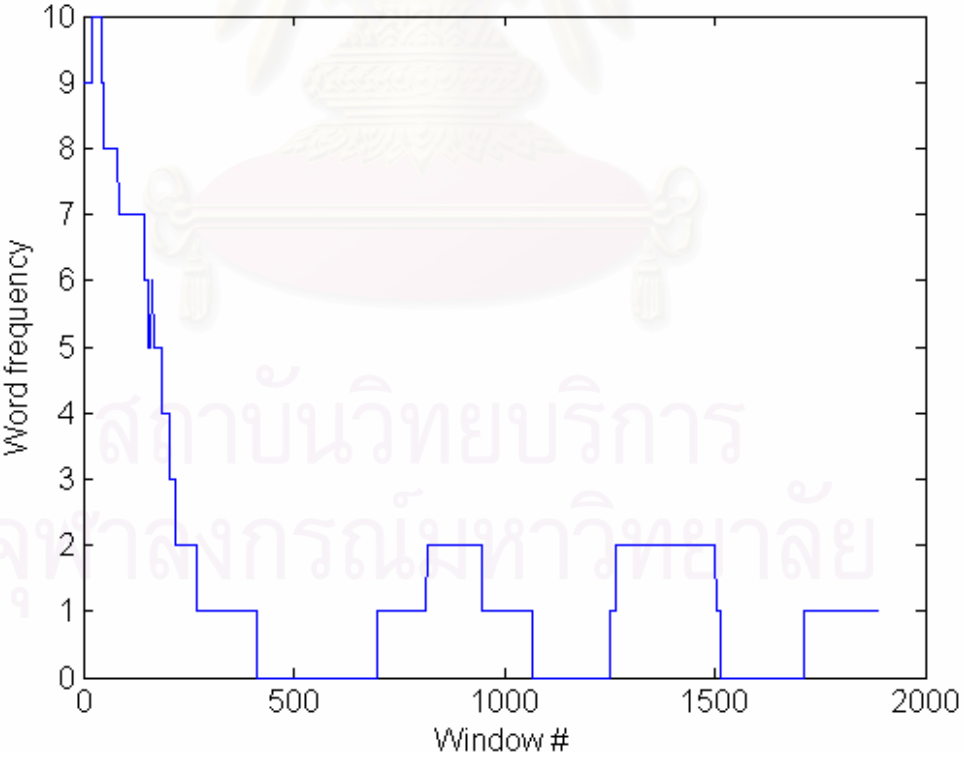
$$a = \left\lfloor \frac{\text{ความยาวของอนุกรมที่ยาวกว่า}}{\text{จำนวนจุดที่ต้องกำจัดออก}} \right\rfloor$$

เช่น อนุกรมเวลาของภาษาอังกฤษยาว 800 จุด และอนุกรมเวลาภาษาไทยยาว 900 จุด ดังนั้นอนุกรมเวลาของภาษาไทยจะต้องถูกกำจัดออกไปเป็นจำนวนเท่ากับ $\left|900 - 800\right| = 100$ จุด ค่า $a = \lfloor 900/100 \rfloor = 9$ ดังนั้นตำแหน่งที่ต้องถูกกำจัดได้แก่ p = 9, 18, 27, 36, 45, …

การลดความยาวอนุกรมด้วยวิธีนี้ นอกจากจะทำให้อนุกรมยาวเท่ากันแล้ว ยังใช้ลดความยาวอนุกรมให้สั้นลงตามสัดส่วนที่ต้องการได้ เช่น การลดความยาวลงเหลือ 50% ของความยาวทั้งหมด ยังคงให้กราฟรูปร่างเหมือนความยาว 100% ดังแสดงในรูปที่ 3.8 และ 3.9 เป็นต้น เนื่องจากในกรณีที่ข้อมูลเข้า (input) มีความยาวมาก จำนวนจุดที่ต้องใช้ในการคำนวณฟังก์ชันระยะห่างจะมากตามไปด้วย ทำให้การประมวลผลใช้เวลานาน อนุกรมที่สั้นลงจะช่วยลดระยะเวลาในการประมวลผลทำให้ได้ผลลัพธ์รวดเร็วขึ้น ในขณะที่ความถูกต้องยังคงมีนัยสำคัญไม่แตกต่างจากเดิม



รูปที่ 3.8 รูปร่างอนุกรมเวลาก่อนลดความยาว



รูปที่ 3.9 รูปร่างอนุกรมเวลาหลังลดความยาวลงเหลือ 50%

3.4.2 การวัดความเหมือน

งานวิจัยนี้จะคำนวณหาความเหมือนของอนุกรมเวลาด้วยฟังก์ชันระยะห่าง ระหว่างภาษาอังกฤษหนึ่งคำกับภาษาไทยทุกคำ โดยคำในภาษาไทยคำใดที่ให้ระยะห่างน้อยที่สุด 3 อันดับแรก (3-Nearest Neighbor) จะถูกเก็บไว้เป็นคู่ของภาษาอังกฤษคำนั้น ดำเนินการเช่นเดียวกันนี้กับทุกคำในภาษาอังกฤษยกเว้นคำหยุด เพราะเป็นคำที่พบในประโยคเกือบทุกประโยค ทำให้ไม่เหมาะที่จะใช้เป็นคำที่ระบุถึงสิ่งเดียวกัน สำหรับการจับคู่ประโยค การวัดความถูกต้องในการจับคู่คำนั้นจะใช้การหาค่าเฉลี่ยของส่วนกลับลำดับชั้น เนื่องจากคำที่มีระยะห่างน้อยที่สุดเป็นอันดับแรก เป็นคำที่มีความน่าจะเป็นว่าจะถูกต้องมากที่สุด ดังนั้นจึงควรได้คะแนนมากกว่าคำที่มีอันดับรองลงมาตามลำดับ

ตัวอย่าง กำหนดให้มีจำนวนคำภาษาอังกฤษที่ต้องการจับคู่ทั้งหมด 69 คำ แต่ละคำจับคู่กับคำภาษาไทยที่มีระยะห่างสั้นที่สุด 3 ลำดับแรก จำนวนคำที่ถูกต้องในแต่ละลำดับแสดงในตารางที่ 3.1 จะหาค่าเฉลี่ยของส่วนกลับลำดับชั้นได้ดังนี้ [(1*55)+(1/2*6)+(1/3*2)]/69 = 0.85

ตารางที่ 3.1 ส่วนกลับลำดับชั้น

ลำดับที่	จำนวนคำที่ถูกต้อง	ค่าส่วนกลับลำดับชั้น
1	55	1
2	6	1/2
3	2	1/3

เนื่องจากการจับคู่คำในงานวิจัยนี้มีวัตถุประสงค์เพื่อนำไปใช้ในการจับคู่ประโยค ดังนั้นการนับจำนวนคำที่ถูกต้องในการจับคู่คำ จะพิจารณาในระดับของประโยคคือ ถ้าคำที่จับคู่ได้อยู่ในประโยคเดียวกับคำที่ถูกต้อง ก็จะนับเป็นคำที่ถูกต้อง โดยนับเฉพาะคำในลำดับแรก เช่น ถ้าคำในลำดับที่ 1 อยู่ในประโยคที่ถูกต้องแล้วก็จะนับเป็นคำที่ถูกต้องในลำดับที่ 1 และจะไม่นับซ้ำซ้อนแม้ว่าคำในลำดับที่ 2 จะอยู่ในประโยคที่ถูกต้องก็ตาม

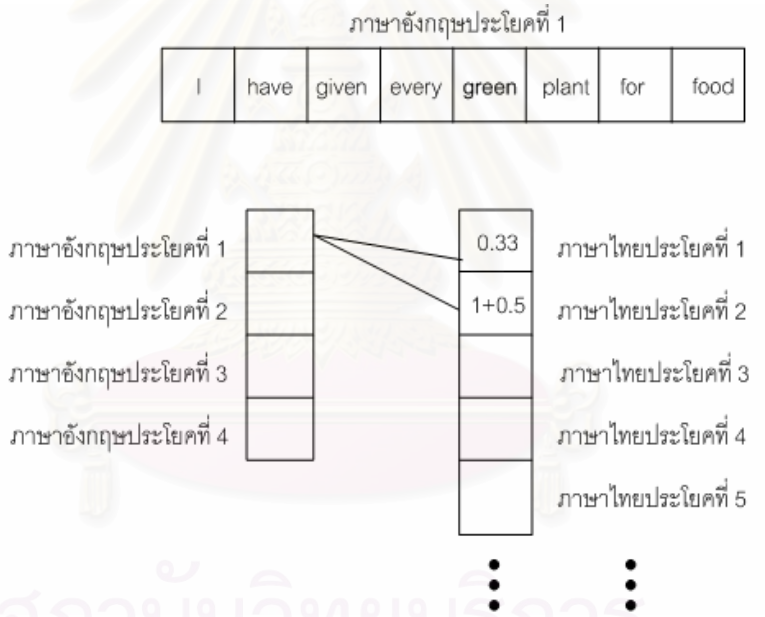
3.5 การจับคู่ประโยคในคลังข้อความขนาน

การจับคู่ประโยคอาศัยแนวคิดที่ว่า ถ้าประโยคคู่ใดมีคำที่ระบุถึงสิ่งเดียวกันมากที่สุดก็จะถูกพิจารณาให้เป็นคู่กัน โดยการทำงานจะแบ่งเป็น 2 ส่วน ได้แก่ การให้คะแนนคู่ประโยค และการจับคู่ประโยค ในงานวิจัยนี้คำที่ระบุถึงสิ่งเดียวกันเป็นผลที่ได้จากการจับคู่คำในหัวข้อ 3.4

3.5.1 การให้คะแนนคู่ประโยค

การให้คะแนน ทำได้โดยนำคำศัพท์ภาษาอังกฤษแต่ละคำในประโยค ไปค้นหาคู่คำศัพท์ภาษาไทย ซึ่งเป็นผลที่ได้มาจากการจับคู่คำที่ทำในหัวข้อ 3.4 เมื่อพบคู่คำศัพท์ภาษาไทยแล้วนำคำศัพท์ภาษาไทยนั้นไปตรวจสอบว่าอยู่ในข้อความภาษาไทยประโยคใด แล้วบวกคะแนนให้กับประโยคภาษาไทยนั้น โดยจะให้คะแนนเท่ากับค่าส่วนกลับลำดับชั้น เนื่องจากคำที่มีระยะห่างน้อยกว่าก็ควรจะเป็นคำที่น่าเชื่อถือมากกว่า ดังนั้นจึงให้คะแนนมากกว่า

คู่คำศัพท์ที่ได้ในภาษาไทยจะมี 3 คำ เรียงตามคำที่มีระยะห่างน้อยที่สุดสามอันดับแรก ดังนั้นคำที่มีระยะห่างน้อยที่สุดจะได้คะแนนเท่ากับ 1 ส่วนคำที่มีระยะห่างน้อยที่สุดเป็นอันดับสองจะได้คะแนนเท่ากับ 0.5 และคำที่มีระยะห่างน้อยที่สุดเป็นอันดับสามจะได้คะแนนเท่ากับ 0.33 ตัวอย่างเช่น การให้คะแนนคู่ประโยค “I have given every green plant for food.” ในรูปที่ 3.10



รูปที่ 3.10 แสดงการให้คะแนนประโยคภาษาไทย

สมมติคำว่า green อยู่ในประโยคที่ 1 ทำการจับคู่คำได้กับคำว่า เขียว พืชผัก และผล เรียงตามระยะห่างน้อยสุดสามอันดับแรก โดยสมมติให้

ภาษาไทยประโยคที่ 1 คือ เราให้บรรดาต้นไม้ซึ่งมีเมล็ดในผลเป็นอาหารแก่เจ้า

ภาษาไทยประโยคที่ 2 คือ เราให้บรรดาพืชผักเขียวสดเป็นอาหาร

คำว่า เขียว พบในประโยคที่ 2 ดังนั้นประโยคที่ 2 จะได้ 1 คะแนน

คำว่า พืชผัก พบในประโยคที่ 2 ดังนั้นประโยคที่ 2 จะได้เพิ่มอีก 0.5 คะแนน

คำว่า ผล พบในประโยคที่ 1 ดังนั้นประโยคที่ 1 จะได้ 0.33 คะแนน

เมื่อพิจารณาคำว่า green แล้วก็พิจารณาคำว่า plant ต่อไป ประโยคอื่น ๆ ก็ทำในลักษณะเดียวกันนี้

3.5.2 **การจับคู่ประโยค**

การจับคู่ประโยค ทำได้โดยหาว่าประโยคภาษาไทยใดที่ได้คะแนนรวมจากการให้คะแนนคู่ประโยคมากที่สุด ประโยคภาษาไทยนั้นจะถูกจับคู่กับประโยคภาษาอังกฤษที่กำลังพิจารณา ตัวอย่างเช่น จากรูปที่ 3.10 ภาษาอังกฤษประโยคที่ 1 เมื่อให้คะแนนคู่ประโยคจากทุกคำในประโยคภาษาอังกฤษ คือคำว่า ‘green’ ‘plant’ และ ‘food’ (ไม่พิจารณาคำว่า ‘I’ ‘have’ ‘give’ ‘every’ และ ‘for’ เนื่องจากเป็นคำหยุด) ประโยคภาษาไทยประโยคใดที่ได้คะแนนรวมสูงสุดก็จะถูกจับคู่กับภาษาอังกฤษประโยคที่ 1 นี้ เป็นต้น

สำหรับกรณีการจับคู่แบบ 1 ประโยคภาษาอังกฤษกับ 1 ประโยคภาษาไทย จะเลือกเอาประโยคภาษาไทยที่ได้คะแนนมากที่สุด ให้เป็นคู่ของประโยคภาษาอังกฤษที่กำลังพิจารณา ส่วนกรณีการจับคู่ประโยคแบบ 1 ประโยคในภาษาอังกฤษกับ N วรรคในภาษาไทยนั้น เนื่องจากภาษาไทยไม่มีเครื่องหมายแบ่งประโยค และ 1 ประโยคในภาษาไทยอาจจะประกอบด้วยหลายวรรค ดังนั้นจึงต้องมีวิธีการที่จะช่วยในการจับคู่ประโยคในภาษาไทย ซึ่งในที่นี้จะใช้การนับจำนวนคำเนื้อหา แต่อย่างไรก็ตาม จำนวนคำเนื้อหาในประโยคภาษาอังกฤษและภาษาไทยไม่ได้เท่ากันเสมอไป ดังนั้นจะมีการทำการทดลองหาอัตราส่วนของจำนวนคำเนื้อหาในภาษาอังกฤษและภาษาไทยในหัวข้อ 4.3.6

สำหรับกรณี 1:N มีการกำหนดเงื่อนไขเพื่อช่วยในการเลือกวรรคที่ถูกต้องดังนี้

1. เก็บวรรคที่มีคะแนนสูงสุด 5 อันดับแรกไว้ เพื่อเป็นตัวเลือกในการจับคู่ประโยคภาษาอังกฤษ เช่น ภาษาอังกฤษประโยคที่ 145 มีวรรคในภาษาไทยที่คะแนนมากที่สุด 5 อันดับแรกเป็นวรรคที่ {129, 148, 143, 142, 141} เป็นต้น การที่เก็บไว้ 5 อันดับ เนื่องด้วยจากการสังเกตส่วนใหญ่ประโยคภาษาอังกฤษ 1 ประโยคจะแปลเป็นภาษาไทยไม่เกิน 5 วรรค

2. นำวรรคที่เก็บไว้จากข้อ 1 มาเรียงตามหมายเลขลำดับที่ของวรรคจากน้อยไปมาก เพื่อจะกำจัดวรรคแรก หรือวรรคสุดท้ายที่ห่างจากวรรคอื่นมาก ๆ โดยถ้าวรรคแรกหรือวรรคสุดท้ายอยู่ห่างจากวรรคที่อยู่ติดกันเกินกว่าครึ่งหนึ่งของผลต่างระหว่างวรรคแรกและวรรคสุดท้าย วรรคนั้นก็จะถูกกำจัดทิ้งเช่น จากตัวอย่างในข้อ 1 เรียงตามลำดับที่ของวรรคจากน้อยไปมากได้เป็น {129, 141, 142, 143, 148} วรรคแรกกับวรรคสุดท้ายห่างกัน 148–129 = 19 และ วรรคแรกกับวรรคที่สองห่างกัน 141–129 = 12 ซึ่งมากกว่าครึ่งหนึ่งของ 19 ดังนั้น วรรคที่ 129 จะถูกตัดทิ้ง แต่อย่างไรก็ตาม ค่าครึ่งหนึ่งของ ผลต่างระหว่างวรรคแรกและวรรคสุดท้ายก็ไม่ควรน้อยกว่า 5 มิฉะนั้นแล้ว

อาจจะเป็นการกำจัดวรรคที่ควรเป็นคู่ประโยคทิ้งไปได้ ส่วนเหตุผลที่กำจัดวรรคที่อยู่ห่างมาก ๆ ออก เนื่องจากกลุ่มของวรรคที่จะเป็นคู่ของประโยคภาษาอังกฤษประโยคเดียวกันควรที่จะอยู่ใกล้กันมากกว่า โดยจะทำการกำจัดวรรคด้วยวิธีนี้ซ้ำ 2 รอบ เนื่องจากเราเก็บวรรคที่มีคะแนนสูงสุด 5 อันดับแรก ดังนั้นการวนซ้ำ 2 รอบนั้นจึงไม่มากไม่น้อยเกินไป

3. ถ้าวรรคใดที่มีจำนวนคำเนื้อหามาก แต่ได้คะแนนน้อยกว่าวรรคที่มีจำนวนคำเนื้อหาน้อยกว่า วรรคนั้นจะถูกกำจัดทิ้ง เนื่องจากวรรคที่มีคำเนื้อหามากมีโอกาสที่จะได้คะแนนมากกว่า แต่กลับได้คะแนนน้อยกว่าวรรคที่มีคำเนื้อหาน้อยกว่า แสดงว่าวรรคนั้นมีคำที่เป็นคู่กันน้อยจึงกำจัดทิ้งได้

4. รวมจำนวนคำเนื้อหาในวรรคต่าง ๆ ที่ไม่โดนกำจัด ถ้าเกินจากอัตราส่วนของจำนวนคำเนื้อหาในภาษาอังกฤษและภาษาไทย ส่วนที่เกินจะถูกตัดทิ้ง

5. วรรคที่อยู่ระหว่างวรรคอื่นที่จับคู่ได้ วรรคนั้นจะถูกจับคู่ไปด้วย เช่น ภาษาอังกฤษประโยคที่ 3 จับคู่กับวรรคที่ 3 และ 5 ดังนั้นวรรคที่ 4 จะถูกจับคู่ไปด้วยได้เป็น ภาษาอังกฤษประโยคที่ 3 จับคู่กับภาษาไทยวรรคที่ 3 4 และ 5

ในบทต่อ ๆ ไป จะเป็นรายละเอียดเกี่ยวกับการทดลองจับคู่คำและจับคู่ประโยคตามแนวคิดที่ได้นำเสนอไปแล้วในบทนี้

บทที่ 4

ข้อมูลเข้าและพารามิเตอร์ที่ใช้ในการทดลอง

ในบทนี้จะกล่าวถึงข้อมูลที่ใช้ในการทดลอง พารามิเตอร์ที่ใช้ในการทดลอง และการปรับพารามิเตอร์

4.1 ข้อมูลที่ใช้ในการทดลอง

4.1.1 ประเภทของข้อมูล

ในงานวิจัยนี้ใช้ข้อมูล 3 ประเภทได้แก่ คัมภีร์ไบเบิล ตัวอย่างคู่ประโยคจากดิกชันนารี และกฎหมายไทย เพื่อทดสอบว่าประเภทของข้อความจะมีผลต่อความถูกต้องในการจับคู่ประโยคหรือไม่มากน้อยเพียงใด

4.1.2 ขนาดของแฟ้มข้อมูล

ในการทดลองจะใช้แฟ้มข้อมูลขนาดต่าง ๆ กัน เพื่อทดสอบว่าความยาวของข้อความจะมีผลต่อความถูกต้องในการจับคู่หรือไม่ ขนาดของข้อมูลเข้าแสดงเปรียบเทียบไว้ในตารางที่ 4.1

ตารางที่ 4.1 ขนาดข้อมูลเข้า

ข้อมูลเข้า	ภาษาอังกฤษ			ภาษาไทย		
	จำนวนหน้า	จำนวนประโยค	จำนวนคำ	จำนวนหน้า	จำนวนวรรค	จำนวนคำ
ไบเบิลขนาดสั้น	1	31	816	1	108	812
ไบเบิลขนาดกลาง	5	175	4,253	5.5	547	4,500
ไบเบิลขนาดยาว	21.5	718	17,014	24	2,306	19,124
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น	1.5	70	1,029	2	-	1,365
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง	5	304	3,884	6.5	-	5,079
ข้อกฎหมายขนาดสั้น	1.5	50	1,264	1.5	153	1,187
ข้อกฎหมายขนาดกลาง	4	102	3,131	3.5	387	2,855

หมายเหตุ ตัวอย่างคู่ประโยคจากดิกชันนารีเป็นข้อมูลที่ได้รับมาจากหน่วยงานวิจัยอื่น ซึ่งมีการเตรียมข้อมูลโดยการตัดคำมาเรียบร้อยแล้ว จึงไม่ทราบจำนวนวรรค

รายละเอียดของแฟ้มข้อมูลมีดังต่อไปนี้

1. คัมภีร์ไบเบิลฉบับภาษาอังกฤษและภาษาไทย จากเว็บไซต์ที่มีข้อมูลไบเบิลหลายภาษาและหลายเวอร์ชัน โดยเลือกมาบางบท [29] ใช้ความยาวต่างกัน 3 ขนาด ได้แก่

   1) ปฐมกาล (Genesis) – บทที่ 1 ความยาวภาษาอังกฤษ 816 คำ และความยาวภาษาไทย 812 คำ (ประมาณ 1 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 124 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 160 คำ จำนวนประโยค 31 ประโยค (โดยต่อไปจะเรียกว่า ไบเบิลขนาดสั้น)

   2) ปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy) – บทที่ 1 ความยาวภาษาอังกฤษ 4,253 คำ (ประมาณ 5 หน้า) และความยาวภาษาไทย 4,500 คำ (ประมาณ 5.5 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 463 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 719 คำ จำนวนประโยค 175 ประโยค (โดยต่อไปจะเรียกว่า ไบเบิลขนาดกลาง)

   3) ปฐมกาล (Genesis) – บทที่ 1 ถึง 25 ความยาวภาษาอังกฤษ 17,014 คำ (ประมาณ 21.5 หน้า) และความยาวภาษาไทย 19,124 คำ (ประมาณ 24 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 927 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 1,529 คำ จำนวนประโยค 718 ประโยค (โดยต่อไปจะเรียกว่า ไบเบิลขนาดยาว)

2. ตัวอย่างประโยคจากดิกชันนารี [30] ใช้ความยาวต่างกัน 2 ขนาด ได้แก่

   1) ตัวอย่างคู่ประโยค 70 ตัวอย่าง ความยาวภาษาอังกฤษ 1,029 คำ (ประมาณ 1.5 หน้า) และความยาวภาษาไทย 1,365 คำ (ประมาณ 2 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 414 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 509 คำ จำนวนประโยค 70 ประโยค (โดยต่อไปจะเรียกว่า ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น)

   2) ตัวอย่างคู่ประโยค 304 ตัวอย่าง ความยาวภาษาอังกฤษ 3,884 คำ (ประมาณ 5 หน้า) และความยาวภาษาไทย 5,079 คำ (ประมาณ 6.5 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 1,096 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 1,248 คำ จำนวนประโยค 304 ประโยค (โดยต่อไปจะเรียกว่า ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง)

3. กฎหมายไทย จากเว็บไซต์รัฐสภา ซึ่งมีข้อมูลกฎหมายทั้งภาคภาษาไทยและภาษาอังกฤษ [31] ใช้ความยาวต่างกัน 2 ขนาด ได้แก่

1) ข้อกฎหมาย 50 ข้อ ความยาวภาษาอังกฤษ 1,264 คำ (ประมาณ 1.5 หน้า) และความยาวภาษาไทย 1,187 คำ (ประมาณ 1.5 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 315 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 357 คำ จำนวนประโยค 50 ประโยค (โดยต่อไปจะเรียกว่า ข้อกฎหมายขนาดสั้น)
2) ข้อกฎหมาย102 ข้อ ความยาวภาษาอังกฤษ 3,131 คำ (ประมาณ 4 หน้า) และความยาวภาษาไทย 2,855 คำ (ประมาณ 3.5 หน้า) คำศัพท์ที่แตกต่างกันในภาษาอังกฤษ 550 คำ คำศัพท์ที่แตกต่างกันในภาษาไทย 603 คำ จำนวนประโยค 102 ประโยค (โดยต่อไปจะเรียกว่า ข้อกฎหมายขนาดกลาง)

4.2 พารามิเตอร์ในการทดลอง

เนื่องจากผู้วิจัยต้องการหาวิธีที่จะทำให้การจับคู่คำและประโยคได้ผลดีมีประสิทธิภาพ จึงทำการวิเคราะห์หาว่ามีปัจจัยใดบ้าง ที่จะมีผลกระทบต่อความถูกต้องในการจับคู่คำและจับคู่ประโยค ซึ่งพารามิเตอร์ที่สำคัญ มีดังต่อไปนี้

1. ขนาดของหน้าต่าง
2. อัตราส่วนการลดความยาวของอนุกรมเวลา
3. ชนิดของฟังก์ชันระยะห่างที่ใช้ในการวัดความเหมือนของอนุกรมเวลา
4. จำนวนของคำหยุด
5. อันดับของคู่คำที่ใช้ในการให้คะแนนคู่ประโยค
6. อัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและคำเนื้อหาในภาษาไทย

4.3 การปรับพารามิเตอร์

เนื่องจากผู้วิจัยต้องการทราบว่าพารามิเตอร์แต่ละตัวมีผลต่อความถูกต้องในการจับคู่คำและประโยคอย่างไรบ้าง จึงทำการทดลองเบื้องต้นดังแสดงในหัวข้อ 4.3.1 ถึง 4.3.6 โดยในการทดลองเบื้องต้นสำหรับแต่ละพารามิเตอร์จะมีการทดลองในหลาย ๆ กรณีเช่น ใช้ข้อมูลเข้าหลาย ๆ ประเภท และขนาด พร้อมทั้งใช้ฟังก์ชันระยะห่างที่ต่างกัน

4.3.1 ขนาดของหน้าต่าง

ผู้วิจัยใช้ขนาดหน้าต่างที่แตกต่างกันได้แก่ 1% 5% 10% 20% 30% 40% และ 50% ของความยาวอนุกรมเวลา ตัวอย่างเช่น อนุกรมเวลายาว 1000 จุด ขนาดหน้าต่าง 5% ของความยาวอนุกรมเวลา จะเท่ากับ (5x1000)/100 = 50 จุด และทำการทดสอบดังการทดลองเบื้องต้นที่ 1 และ 2 เพื่อดูว่าขนาดของหน้าต่างที่แตกต่างกันมีผลต่อความถูกต้องในการจับคู่คำหรือไม่ เพียงใด

การทดลองเบื้องต้นที่ 1 การทดลองใช้หน้าต่างขนาดต่าง ๆ กับไบเบิลขนาดสั้น
ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)
ฟังก์ชันระยะห่าง : แมนฮัตตัน
จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 69 คำ
จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 126 คำ

ตารางที่ 4.2 ผลการทดลองจับคู่คำเมื่อใช้หน้าต่างขนาดต่าง ๆ กับไบเบิลขนาดสั้น

ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้อง
1%	43	10	3	0.71
5%	56	6	2	0.86
10%	49	10	2	0.79
20%	51	6	3	0.78
30%	46	9	5	0.76
40%	48	12	1	0.79
50%	49	7	3	0.78

การทดลองเบื้องต้นที่ 2 การทดลองใช้หน้าต่างขนาดต่าง ๆ กับข้อกฎหมายขนาดกลาง
ข้อมูลเข้า : ข้อกฎหมาย 102 ข้อ
ฟังก์ชันระยะห่าง : แมนฮัตตัน
จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 449 คำ
จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 552 คำ

ตารางที่ 4.3 ผลการทดลองจับคู่คำเมื่อใช้หน้าต่างขนาดต่าง ๆ กับข้อกฎหมายขนาดกลาง

ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้อง
1%	151	51	24	0.41
5%	175	60	44	**0.49**
10%	147	59	33	0.42
20%	171	54	28	0.46
30%	170	56	37	0.47
40%	165	44	39	0.45
50%	160	45	38	0.43

จากการผลการทดลองในตารางที่ 4.2 และ 4.3 พบว่าที่หน้าต่างขนาด 5% ของความยาวอนุกรมเวลาได้เปอร์เซ็นต์ความถูกต้องมากกว่าขนาดอื่น ๆ

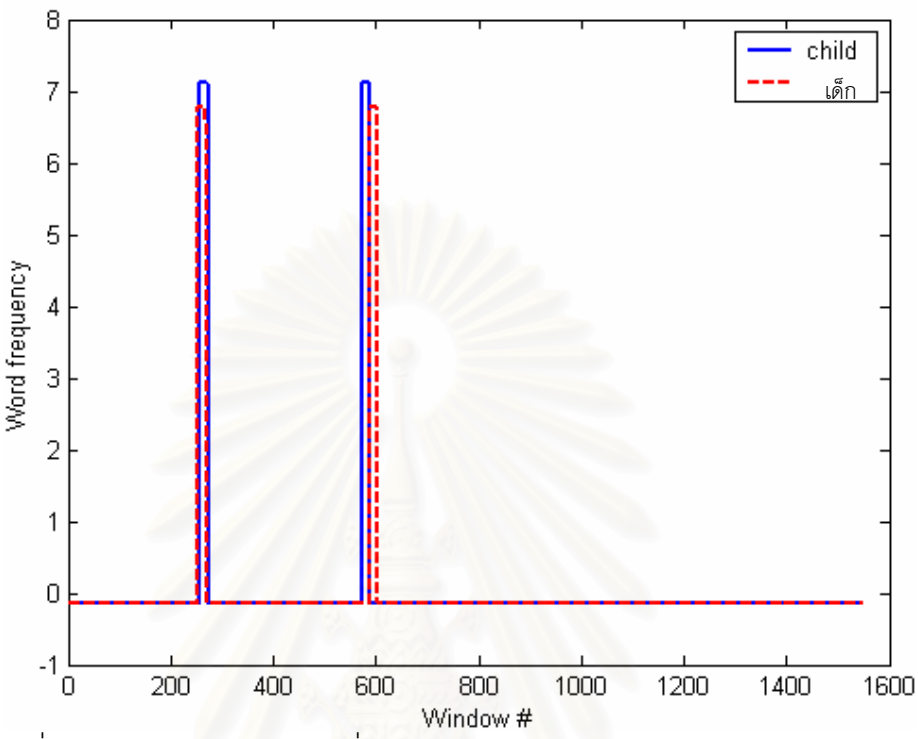
รูปที่ 4.1 รูปที่ 4.2 รูปที่ 4.3 และรูปที่ 4.4 เป็นอนุกรมเวลาของคำว่า “child” และคำว่า “เด็ก” ซึ่งจะเป็นตัวอย่างที่แสดงให้เห็นถึงสาเหตุหนึ่งที่ทำให้หน้าต่างขนาด 5% ของความยาวอนุกรมเวลา ให้ผลดีกว่าหน้าต่างขนาดอื่น ๆ

จากรูปที่ 4.1 พบว่าเมื่อใช้หน้าต่างขนาด 1% ของความยาวอนุกรมเวลา แท่งกราฟจะแคบ ดังนั้นโอกาสที่เส้นกราฟของคำในภาษาอังกฤษและภาษาไทยจะทับกันจะน้อยกว่าการใช้หน้าต่างขนาดใหญ่ขึ้น เนื่องจากส่วนใหญ่แล้วตำแหน่งของคำในภาษาอังกฤษและภาษาไทยไม่ได้อยู่ตรงกัน อาจจะมีการเยื้องกันเล็กน้อย เพราะบางครั้งภาษาไทยก็มีการใช้คำฟุ่มเฟือยกว่าภาษาอังกฤษ หรือบางครั้งภาษาอังกฤษก็มีการใช้คำฟุ่มเฟือยกว่าภาษาไทย ซึ่งการใช้ฟังก์ชันระยะห่างนั้นจะเป็นการคำนวณแบบจุดต่อจุด เมื่อแท่งกราฟไม่ทับกันก็ย่อมมีผลให้ผลรวมของระยะห่างมากขึ้นไปด้วย

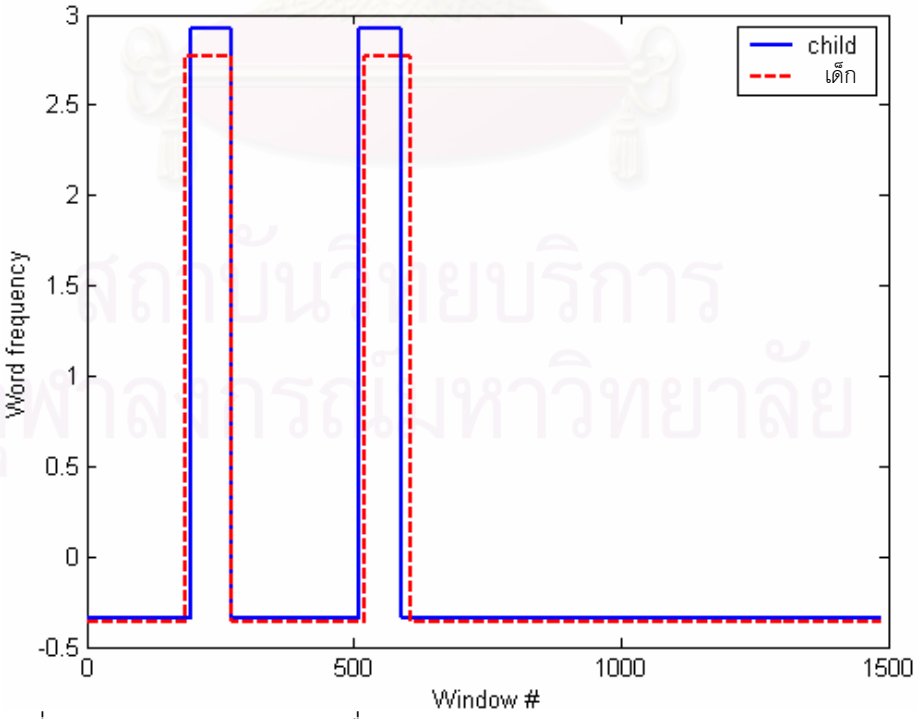
จากรูปที่ 4.2 และ 4.3 พบว่าเมื่อใช้หน้าต่างขนาด 5% และ 10% ของความยาวอนุกรมเวลา แท่งกราฟจะกว้างกว่ารูปที่ 4.1 ดังนั้นโอกาสที่กราฟของคำในภาษาอังกฤษและภาษาไทยจะทับกันจึงมีเพิ่มขึ้น และจะพบว่าหลังจากการแปลงตามค่าคะแนนมาตรฐานซีแล้วค่าในแนวแกน Y ในรูปที่ 4.2 จะน้อยกว่าค่าในแนวแกน Y ในรูปที่ 4.1 ดังนั้นเมื่อคำนวณระยะห่างโดยใช้ฟังก์ชันระยะห่างแล้ว ผลรวมจึงน้อยกว่า ทำให้มีโอกาสที่จะจับคู่คำได้ถูกต้องมากกว่า

จากรูปที่ 4.4 พบว่าเมื่อใช้หน้าต่างขนาด 20% ของความยาวอนุกรมเวลา แท่งกราฟมีความกว้างมากเกินไปจนทำให้แท่งกราฟทั้งสองแท่งมาชนกัน และทำให้ความชัดเจนในตำแหน่งของคำลดลงมาก เนื่องจากพบคำว่า “child” และ “เด็ก” สองครั้งควรจะได้กราฟสองแท่งอย่าง

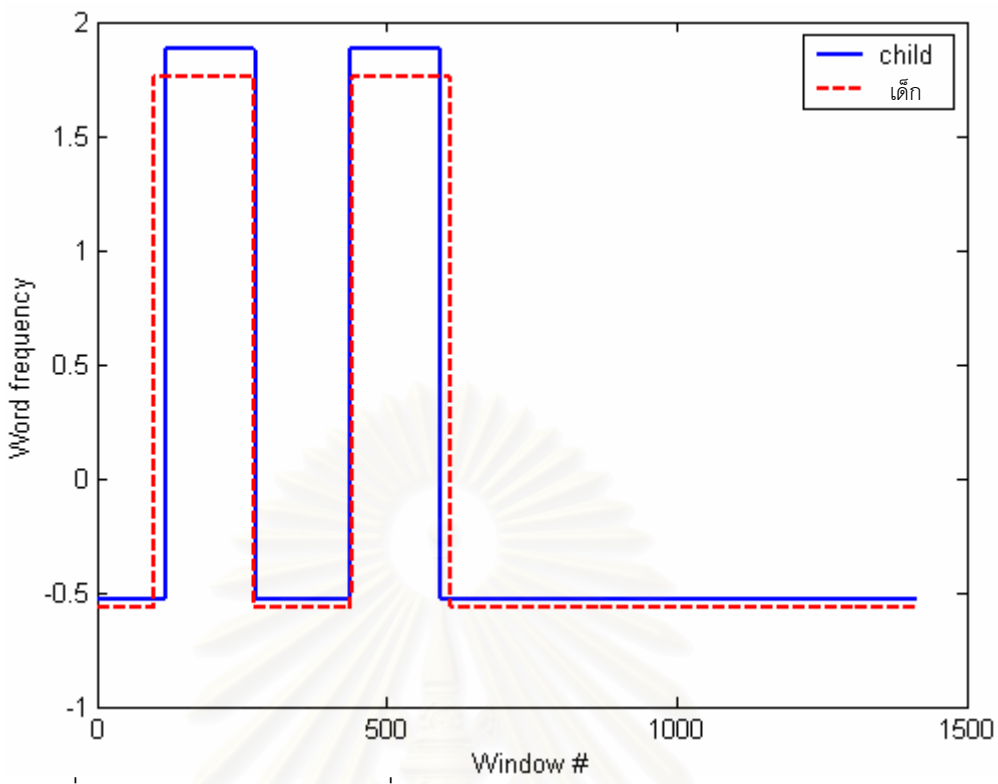
เด่นชัด แต่รูปกราฟที่ได้มีความคลาดเคลื่อนจากที่ควรจะเป็น จึงทำให้รูปกราฟที่ได้อาจจะไปคล้ายกับคำอื่น ที่มีตำแหน่งใกล้เคียงกว่าได้



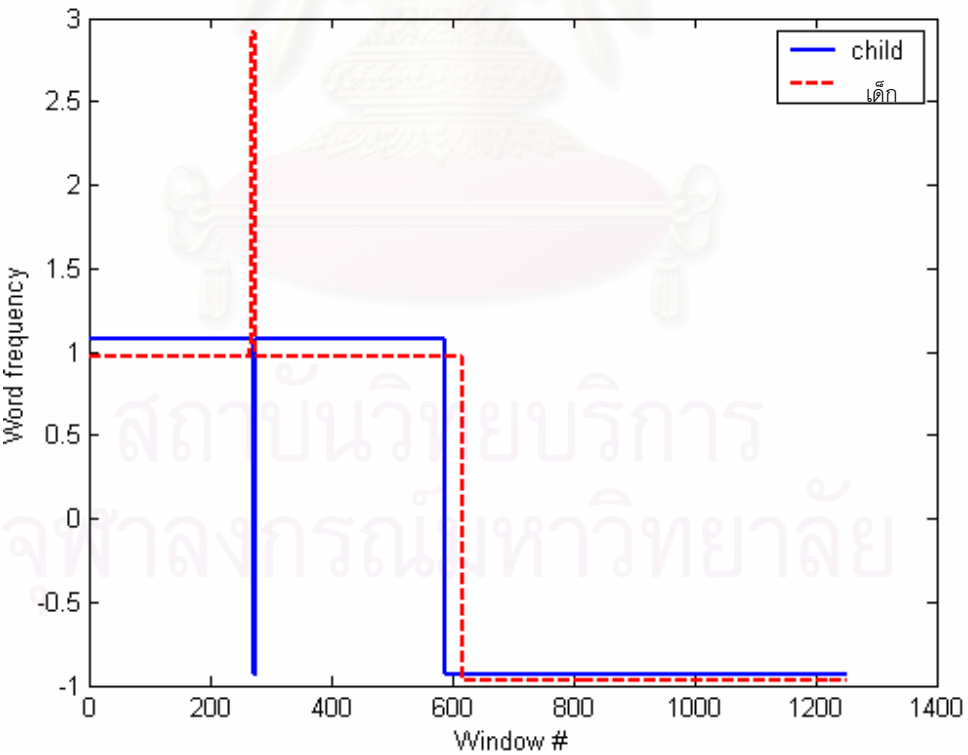
รูปที่ 4.1 ตัวอย่างอนุกรมเวลาเมื่อใช้หน้าต่างขนาด 1% ของความยาวอนุกรมเวลา



รูปที่ 4.2 ตัวอย่างอนุกรมเวลาเมื่อใช้หน้าต่างขนาด 5% ของความยาวอนุกรมเวลา



รูปที่ 4.3 ตัวอย่างอนุกรมเวลาเมื่อใช้หน้าต่างขนาด 10% ของความยาวอนุกรมเวลา



รูปที่ 4.4 ตัวอย่างอนุกรมเวลาเมื่อใช้หน้าต่างขนาด 20% ของความยาวอนุกรมเวลา

จากการทดลองสรุปได้ว่า ขนาดของหน้าต่างมีผลต่อความถูกต้องในการจับคู่คำเนื่องจากขนาดหน้าต่างที่ต่างกันรูปกราฟของอนุกรมเวลาที่ได้ก็จะต่างกันไปดังเช่นรูปที่ 4.1 รูปที่ 4.2 รูปที่ 4.3 และรูปที่ 4.4 การใช้หน้าต่างขนาดเล็กเกินไป มีผลให้โอกาสในการทับกันของกราฟของคำที่เป็นคู่กันในภาษาอังกฤษและภาษาไทยน้อยลง ซึ่งส่งผลให้ผลรวมระยะห่างมากขึ้น การใช้หน้าต่างขนาดใหญ่ขึ้นแท่งกราฟก็จะกว้างขึ้นทำให้โอกาสในการทับกันของกราฟของคำที่เป็นคู่กันในภาษาอังกฤษและภาษาไทยมีมากขึ้น แต่ถ้าหน้าต่างมีขนาดใหญ่เกินไปก็อาจทำให้แท่งกราฟสองแท่งมาชนกันได้ ทำให้รูปที่ได้คลาดเคลื่อนไปจากที่ควรจะเป็นและจับคู่ได้กับคำอื่นแทน ตัวอย่างที่กล่าวมาเป็นกรณีคำที่มีความถี่ในการปรากฏน้อย แต่สำหรับคำที่มีความถี่ในการปรากฏมาก ๆ ขนาดของหน้าต่างจะไม่มีผลต่อความถูกต้องในการจับคู่เท่าใดนักเพราะกราฟจะมีรูปร่างที่ค่อนข้างแตกต่างจากคำอื่น ๆ อยู่แล้ว แต่อย่างไรก็ตาม นอกจากขนาดของหน้าต่างแล้วยังมีตัวแปรอื่น ๆ ที่มีผลต่อความถูกต้องในการจับคู่คำ เช่น อัตราส่วนการลดความยาวของอนุกรมเวลา เป็นต้น ซึ่งจะกล่าวถึงในหัวข้อถัดไป

4.3.2 **อัตราส่วนการลดความยาวของอนุกรมเวลา**

จากที่กล่าวไปในหัวข้อ 2.1.6 ฟังก์ชันระยะห่างเป็นการคำนวณหาระยะห่างของทุกจุดในอนุกรมเวลา ทำให้ใช้เวลาในการคำนวณมาก ซึ่งการลดความยาวของอนุกรมเวลาจะช่วยให้โปรแกรมสามารถทำงานได้เร็วขึ้น เพราะจำนวนจุดที่นำไปคำนวณน้อยลง การทดลองต่อไปจะเป็นการทดสอบว่า การลดความยาวอนุกรมเวลามีผลต่อความถูกต้องในการจับคู่คำหรือไม่ โดยค่าเริ่มต้นคือ 1 หมายถึง ไม่มีการลดความยาวของอนุกรมเวลา ถ้าพารามิเตอร์มีค่า 0.5 หมายความว่า ความยาวของอนุกรมเวลาถูกลดลงครึ่งหนึ่งเช่น จากความยาว 2,000 จุด ลดลงเหลือ 1,000 จุด เป็นต้น

<u>การทดลองเบื้องต้นที่ 3</u> การทดลองลดความยาวของอนุกรมเวลาที่อัตราส่วนต่าง ๆ โดยทำการทดลองกับไบเบิลขนาดกลาง

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

ฟังก์ชันระยะห่าง : แมนฮัตตัน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 338 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 660 คำ

ขนาดหน้าต่าง : 20% ของความยาวอนุกรมเวลา

ตารางที่ 4.4 ผลการทดลองจับคู่คำเมื่อลดความยาวของอนุกรมเวลาที่อัตราส่วนต่าง ๆ กับไบเบิลขนาดกลาง

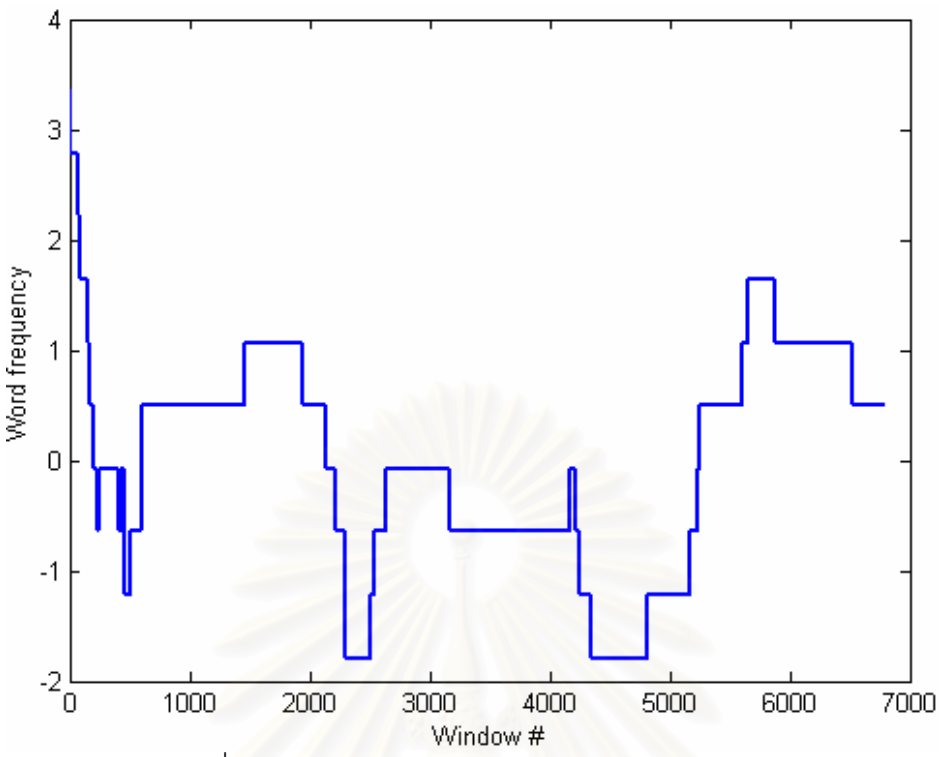
อัตราส่วนการลดความยาวอนุกรมเวลา	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้อง
1.0	85	32	33	0.33
0.9	88	32	33	0.34
0.8	80	28	38	0.32
0.7	89	33	23	0.33
0.6	88	35	20	0.33
0.5	75	39	26	0.31
0.4	75	27	23	0.28
0.3	77	25	30	0.29
0.2	70	34	36	0.29
0.1	70	29	30	0.28

จากผลการทดลองในตารางที่ 4.4 พบว่าการลดความยาวอนุกรมเวลา ยังคงให้ผลที่ใกล้เคียงกับการใช้ความยาวเริ่มต้นก่อนการลดความยาว และบางอัตราส่วนก็ให้ผลดีกว่าการใช้ความยาวเริ่มต้นก่อนการลดความยาว

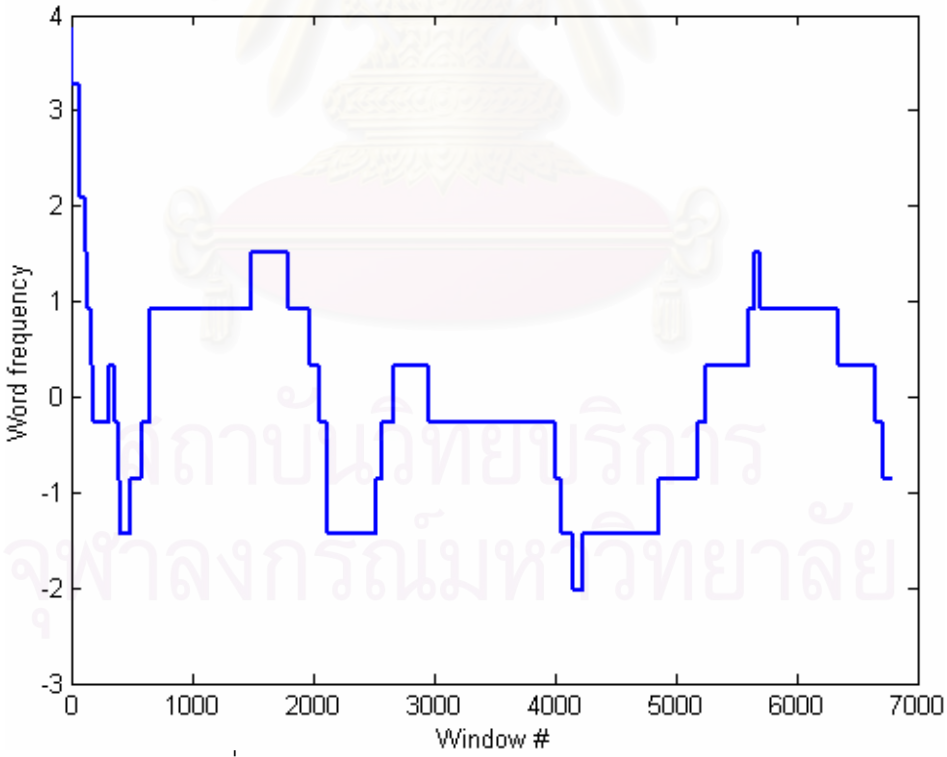
รูปที่ 4.5 ถึงรูปที่ 4.10 เป็นตัวอย่างอนุกรมเวลาของคำว่า heaven และคำว่า ฟ้า แบบลดและแบบไม่ลดความยาวอนุกรมเวลา

จากรูปที่ 4.5 กับ 4.6 และรูปที่ 4.8 กับ 4.9 หากดูจากรูปร่างโดยรวมแล้ว จะคล้ายคลึงกันและควรจะจับคู่กันได้ แต่จากรูปที่ 4.7 และ 4.10 หากพิจารณาในระดับจุดต่อจุดในรูปที่ 4.7 จำนวนจุดจะน้อยกว่า ดังนั้นผลรวมของระยะห่างที่ผิดพลาดก็จะน้อยกว่า ส่วนรูปที่ 4.10 จะมีจำนวนจุดมากกว่า ทำให้ผลรวมของระยะห่างที่ผิดพลาดมากกว่าและอาจจับคู่ได้กับคำอื่นที่มีผลรวมระยะห่างที่ผิดพลาดน้อยกว่า ดังนั้นการลดความยาวของอนุกรมเวลา ให้มีจำนวนจุดน้อยลง จึงมีผลต่อความถูกต้องในการจับคู่คำ

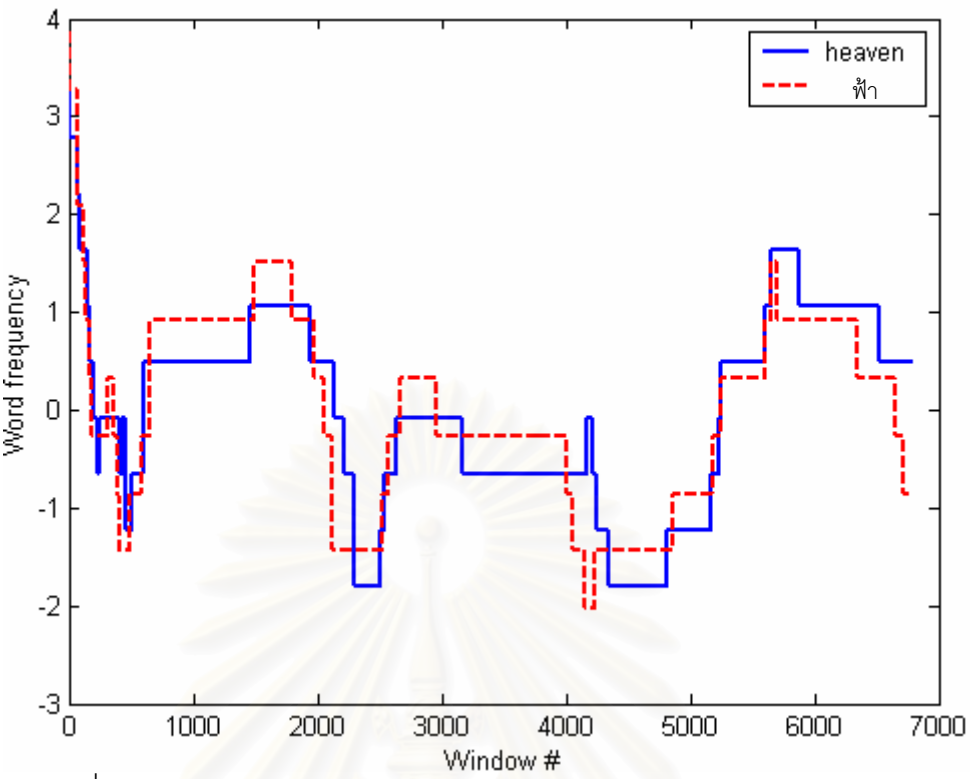
อย่างไรก็ตาม การลดความยาวของอนุกรมเวลามากเกินไปอาจจะทำให้รูปร่างอนุกรมเวลาผิดไปจากเดิมได้ ซึ่งมีผลต่อความถูกต้องในการจับคู่คำเช่นกัน ดังนั้นจึงไม่ควรลดความยาวอนุกรมเวลามากจนเกินไป เพราะบางครั้งการลดความยาวของอนุกรมเวลามากก็ให้ผลดีแต่บางครั้งก็อาจจะให้ผลไม่ดีได้เช่นกัน



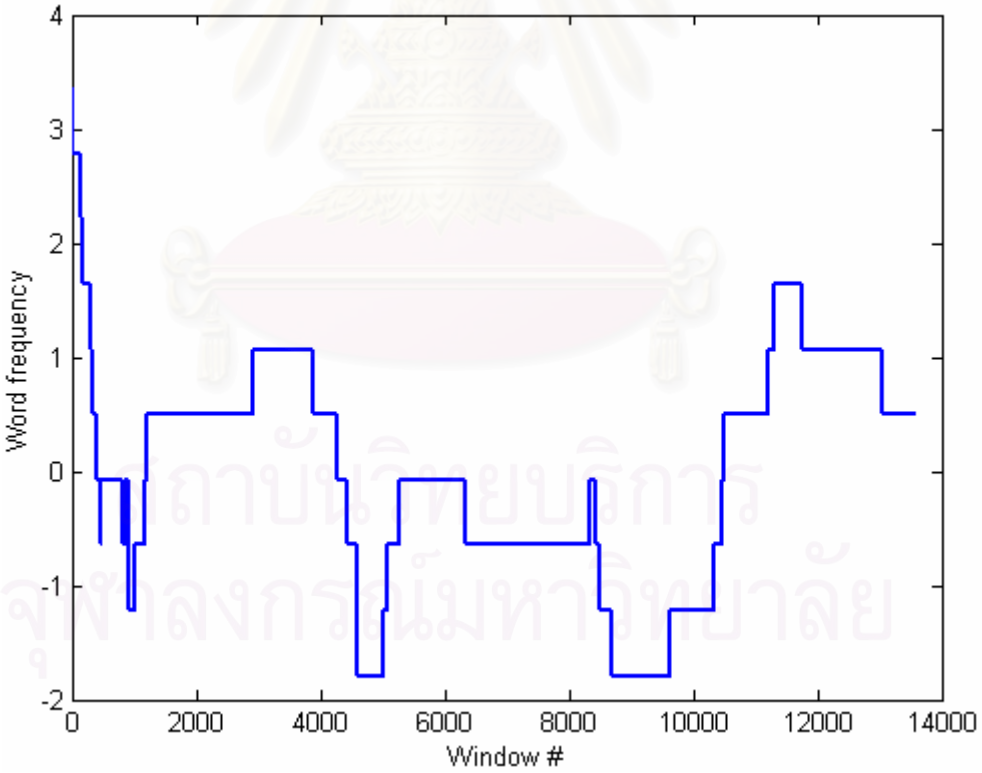
รูปที่ 4.5 อนุกรมเวลาของคำว่า “heaven” (ใช้อัตราส่วน 0.5)



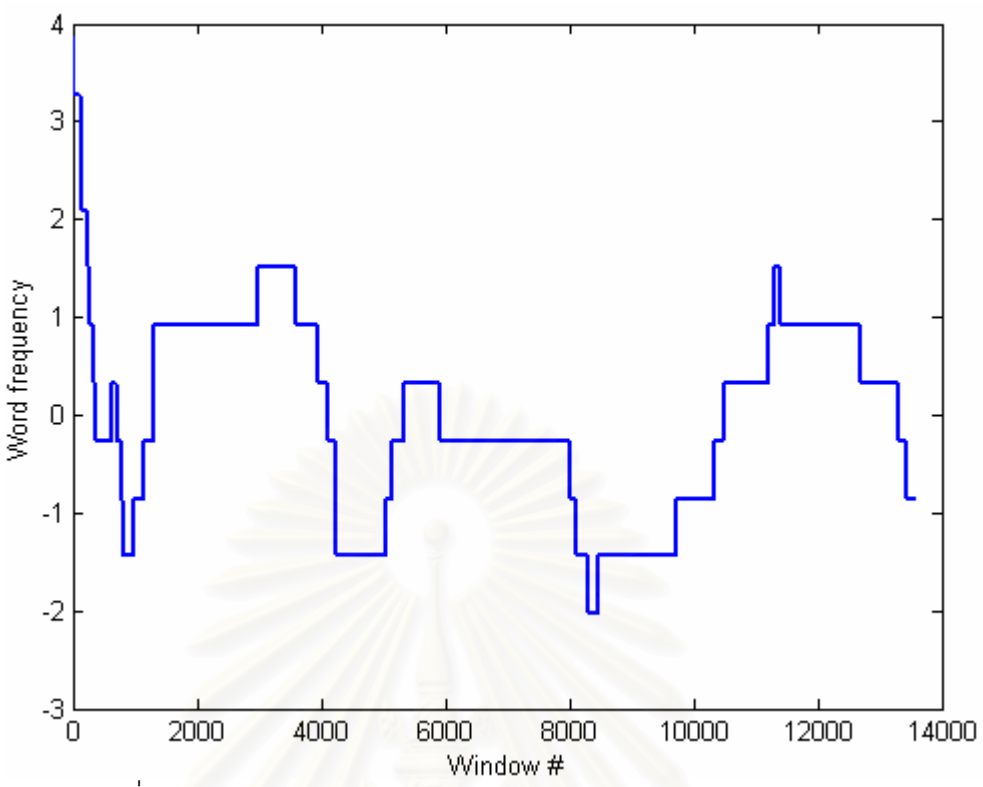
รูปที่ 4.6 อนุกรมเวลาของคำว่า “ฟ้า” (ใช้อัตราส่วน 0.5)



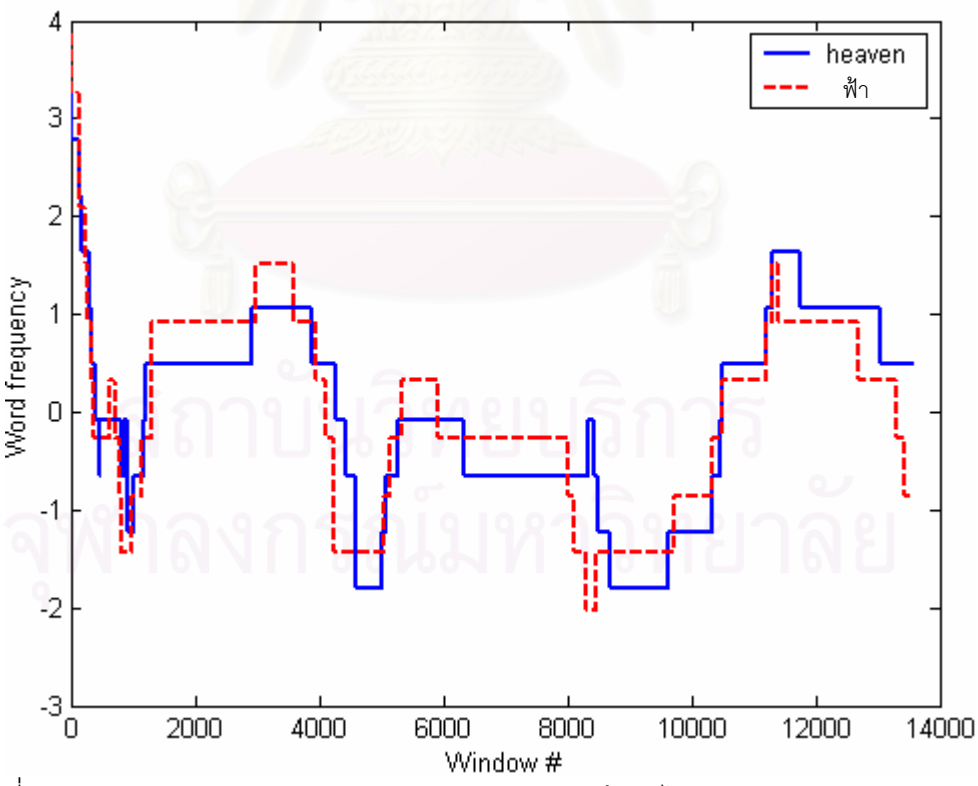
รูปที่ 4.7 อนุกรมเวลาของคำว่า “heaven” และ คำว่า “ฟ้า” (ใช้อัตราส่วน 0.5)



รูปที่ 4.8 อนุกรมเวลาของคำว่า “heaven” (ไม่ลดความยาวของอนุกรมเวลา)



รูปที่ 4.9 อนุกรมเวลาของคำว่า “ฟ้า” (ไม่ลดความยาวของอนุกรมเวลา)



รูปที่ 4.10 อนุกรมเวลาของคำว่า “heaven” และคำว่า “ฟ้า” (ไม่ลดความยาวของอนุกรมเวลา)

จากรูปที่ผ่านมา อาจตั้งสมมติฐานได้ว่ารูปร่างของอนุกรมเวลาของคำที่เป็นคู่กันมีความคล้ายคลึงกันแต่จับคู่กันได้ไม่ดีนัก เนื่องจากการใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันหรือยุคลิเดียนเป็นการคำนวณแบบจุดต่อจุด ซึ่งในหัวข้อ 4.3.3 จะทดลองใช้ฟังก์ชันระยะห่างแบบอื่น ที่มีความยืดหยุ่นกว่า ได้แก่ แบบไดนามิกไทม์วอร์ปปิง ซึ่งเป็นวิธีที่ไม่ได้พิจารณาเฉพาะจุดต่อจุด แต่จะพิจารณาจุดที่ใกล้เคียงด้วย เพื่อดูว่าผลการจับคู่คำจะดีขึ้นหรือไม่อย่างไร

<u>การทดลองเบื้องต้นที่ 4</u> การทดลองเพื่อเปรียบเทียบความถูกต้องในการจับคู่คำ ระหว่างไม่ลดความยาวอนุกรมและลดความยาวอนุกรมด้วยอัตราส่วน 0.5 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตัน

ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)

ฟังก์ชันระยะห่าง : แมนฮัตตัน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 69 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 126 คำ

ตารางที่ 4.5 ผลการทดลองจับคู่คำกรณีไม่ลดความยาวอนุกรมและลดความยาวอนุกรมด้วยอัตราส่วน 0.5 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตัน

	ไม่ลดความยาวอนุกรมเวลา				ลดความยาวด้วยอัตราส่วน 0.5			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้อง
1%	30	17	3	0.57	43	10	3	0.71
5%	48	11	4	0.79	56	6	2	**0.86**
10%	49	10	4	0.80	49	10	2	0.79
20%	50	6	5	0.79	51	6	3	0.78
30%	45	10	3	0.74	46	9	5	0.76

<u>การทดลองเบื้องต้นที่ 5</u> การทดลองเพื่อเปรียบเทียบความถูกต้องในการจับคู่คำ ระหว่างไม่ลดความยาวอนุกรมและลดความยาวอนุกรมด้วยอัตราส่วน 0.5 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียน

ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)

ฟังก์ชันระยะห่าง : ยุคลิเดียน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดกูเกิล) : 110 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 126 คำ

ตารางที่ 4.6 ผลการทดลองจับคู่คำกรณีไม่ลดความยาวอนุกรมและลดความยาวอนุกรมด้วยอัตราส่วน 0.5 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียน

	ไม่ลดความยาวอนุกรมเวลา				ลดความยาวด้วยอัตราส่วน 0.5			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้อง
1%	39	26	9	0.50	32	33	19	0.50
5%	64	20	8	0.70	71	15	12	**0.75**
10%	64	21	11	0.71	66	18	7	0.70
20%	64	21	7	0.70	70	10	6	0.70
30%	60	21	7	0.66	64	19	6	0.69

<u>การทดลองเบื้องต้นที่ 6</u> การทดลองเพื่อเปรียบเทียบความถูกต้องในการจับคู่ประโยค ระหว่างการไม่ลดความยาวอนุกรมและลดความยาวอนุกรมด้วยอัตราส่วน 0.5

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

ฟังก์ชันระยะห่าง : ยุคลิเดียน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 338 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 660 คำ

ตารางที่ 4.7 ผลการทดลองจับคู่ประโยคกรณีไม่ลดความยาวอนุกรมและลดความยาวอนุกรมด้วยอัตราส่วน 0.5 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับไบเบิลขนาดกลาง

	จำนวนประโยคที่จับคู่ได้ถูกต้อง	
ขนาดของหน้าต่าง	ไม่ลดความยาวอนุกรมเวลา	ลดความยาวด้วยอัตราส่วน 0.5
1%	25	29
5%	55	60
10%	52	59
20%	58	59
30%	58	58

ผู้วิจัยได้ทดลองลดความยาวอนุกรมเวลาด้วยอัตราส่วน 0.5 เปรียบเทียบกับการไม่ลดความยาวอนุกรม พร้อมกับกำหนดพารามิเตอร์หลาย ๆ รูปแบบดังการทดลองเบื้องต้นที่ 4 ถึง 6 จากผลการทดลองในตารางที่ 4.5 ถึง 4.7 พบว่าการลดอนุกรมเวลาด้วยอัตราส่วน 0.5 ส่วนใหญ่ผลที่ได้จะใกล้เคียงหรือดีกว่าการไม่ลดความยาวอนุกรม เพราะเป็นอัตราส่วนที่เป็นค่ากลาง ๆ ไม่มากไม่น้อยจนเกินไป ดังนั้นค่าที่เด่นชัดในอนุกรมเวลาก็จะยังอยู่ครบ ดังนั้นผู้วิจัยจึงจะใช้อัตราส่วน 0.5 ในการทดลองที่เหลือทั้งหมด

4.3.3 ชนิดของฟังก์ชันระยะห่างที่ใช้ในการวัดความเหมือนของอนุกรมเวลา

ในการทดลองนี้จะใช้ฟังก์ชันระยะห่าง 3 แบบได้แก่ ระยะห่างแมนฮัตตัน ระยะห่างยุคลิเดียน และไดนามิกไทม์วอร์ปปิง โดยทำการทดลองดังการทดลองเบื้องต้นที่ 7 8 และ 9

<u>การทดลองเบื้องต้นที่ 7</u> การทดลองเพื่อเปรียบเทียบผลการจับคู่คำระหว่างการใช้ฟังก์ชันระยะห่างแมนฮัตตันและยุคลิเดียน โดยทดลองกับไบเบิลขนาดยาว

ข้อมูลเข้า : บทที่ 1 ถึง 25 ของปฐมกาล (Genesis)

ฟังก์ชันระยะห่าง : แมนฮัตตัน และยุคลิเดียน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 771 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 1,465 คำ

ตารางที่ 4.8 ผลการทดลองจับคู่คำเมื่อใช้ฟังก์ชันระยะห่างแมนฮัตตันและยุคลิเดียน กับไบเบิลขนาดยาว

	แมนฮัตตัน				ยุคลิเดียน			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	154	55	41	0.25	116	47	34	0.20
5%	183	64	38	**0.30**	165	47	39	0.26
10%	193	48	40	**0.30**	179	50	37	0.28
20%	182	64	46	**0.30**	191	52	40	**0.30**
30%	136	45	35	0.22	125	47	28	0.20

<u>การทดลองเบื้องต้นที่ 8</u> การทดลองเพื่อเปรียบเทียบผลการจับคู่คำระหว่างการใช้ฟังก์ชันระยะห่างแมนฮัตตันและยุคลิเดียน โดยทดลองกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 304 ตัวอย่าง

ฟังก์ชันระยะห่าง : แมนฮัตตัน และ ยุคลิเดียน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 912 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 1194 คำ

ตารางที่ 4.9 ผลการทดลองจับคู่คำเมื่อใช้ฟังก์ชันระยะห่างแมนฮัตตันและยุคลิเดียน กับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	แมนฮัตตัน				ยุคลิเดียน			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	154	72	55	0.23	138	66	51	0.21
5%	167	103	73	**0.27**	147	126	87	0.26
10%	149	97	78	0.25	167	92	81	0.26
20%	77	35	49	0.12	69	43	38	0.11
30%	32	25	35	0.06	21	18	37	0.05

จากผลการทดลองในตารางที่ 4.8 และ 4.9 พบว่าผลที่ได้จากการใช้ฟังก์ชันระยะห่างแบบแมนฮัตตัน ส่วนใหญ่จะดีกว่าผลที่ได้จากการใช้ฟังก์ชันระยะห่างแบบยุคลิเดียน ซึ่งสาเหตุที่เป็นเช่นนี้สามารถอธิบายได้ด้วยตัวอย่างในตารางที่ 4.10 ซึ่งแสดงการคำนวณระยะห่างแบบแมนฮัตตันเปรียบเทียบกับแบบยุคลิเดียน จากตารางที่ 4.10 อนุกรมเวลาคู่ที่ 1 แตกต่างกัน 3 หน่วย (จากตำแหน่งที่ 3) อนุกรมเวลาคู่ที่ 2 แตกต่างกัน 3 หน่วย (จากตำแหน่งที่ 2 และ 3) อนุกรมเวลาทั้งสองคู่แตกต่างกัน 3 หน่วยเหมือนกัน ซึ่งการคำนวณด้วยสูตรแมนฮัตตันจะได้ระยะห่างเท่ากันคือ 3 แต่การคำนวณแบบยุคลิเดียนจะได้ระยะห่างต่างกันคือ 3 กับ 2.24 ซึ่งจะเห็นได้ว่ายุคลิเดียนมีความอ่อนไหว (Sensitive) ต่อค่าของข้อมูลต่าง ๆ มากกว่าแมนฮัตตัน ดังนั้นอาจกล่าวได้ว่า การหาความเหมือนกันของอนุกรมเวลา สำหรับโดเมนการจับคู่ข้อความในคลังข้อความขนานไม่ต้องการสูตรที่อ่อนไหวต่อค่าของข้อมูลมากนัก ดังนั้นการใช้สูตรแมนฮัตตันจึงสามารถจับคู่คำได้ถูกต้องมากกว่า

ตารางที่ 4.10 เปรียบเทียบระยะห่างที่คำนวณได้จากสูตรแมนฮัตตันและยุคลิเดียน

ตำแหน่งที่ 1 2 3 4 อนุกรมเวลาคู่ที่ 1 [3 7 23 4 3 7 20 4 3 6 18 4] อนุกรมเวลาคู่ที่ 2	
สูตรแมนฮัตตัน $D(X,Y) = \sum_{i=1}^{n} \lvert X_i - Y_i \rvert$	**สูตรยุคลิเดียน** $D(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$
คำนวณระยะห่างของอนุกรมเวลาคู่ที่ 1 $D(X,Y) = \lvert 3-3 \rvert + \lvert 7-7 \rvert + \lvert 23-20 \rvert + \lvert 4-4 \rvert$ $= 0 + 0 + 3 + 0$ $= 3$	คำนวณระยะห่างของอนุกรมเวลาคู่ที่ 1 $D(X,Y) = \sqrt{(3-3)^2 + (7-7)^2 (23-20)^2 + (4-4)^2}$ $= \sqrt{0+0+9+0}$ $= 3$
คำนวณระยะห่างของอนุกรมเวลาคู่ที่ 2 $D(X,Y) = \lvert 3-3 \rvert + \lvert 7-6 \rvert + \lvert 20-18 \rvert + \lvert 4-4 \rvert$ $= 0 + 1 + 2 + 0$ $= 3$	คำนวณระยะห่างของอนุกรมเวลาคู่ที่ 2 $D(X,Y) = \sqrt{(3-3)^2 + (7-6)^2 (20-18)^2 + (4-4)^2}$ $= \sqrt{0+1+4+0}$ $= 2.24$

นอกจากนี้จากที่เคยกล่าวไว้แล้วในหัวข้อ 4.3.2 ว่ารูปร่างของอนุกรมเวลาของคำที่เป็นคู่กันมีความคล้ายคลึงกันแต่จับคู่กันไม่ได้เพราะเส้นกราฟเยื้องกัน ดังนั้นเราจึงจะทำการทดลองใช้ฟังก์ชันระยะห่างแบบอื่น ที่มีความยืดหยุ่นกว่า ไม่ได้พิจารณาเฉพาะจุดต่อจุด แต่จะพิจารณาจุดที่ใกล้เคียงด้วยคือ แบบไดนามิกไทม์วอร์ปปิง ดังการทดลองเบื้องต้นที่ 9

<u>การทดลองเบื้องต้นที่ 9</u> การทดลองเพื่อเปรียบเทียบผลการจับคู่คำระหว่างการใช้ฟังก์ชันระยะห่างแมนฮัตตัน ยุคลิเดียน และไดนามิกไทม์วอร์ปปิง โดยทดลองกับไบเบิลขนาดกลาง

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

ฟังก์ชันระยะห่าง : แมนฮัตตัน ยุคลิเดียน และไดนามิกไทม์วอร์ปปิง

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 338 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 660 คำ

ตารางที่ 4.11 ผลการทดลองจับคู่คำเมื่อใช้ฟังก์ชันระยะห่างแมนฮัตตัน ยุคลิเดียน และไดนามิกไทม์วอร์ปปิง

	แมนฮัตตัน				ยุคลิเดียน			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	59	30	29	0.25	48	26	18	0.20
5%	93	36	31	0.36	81	36	23	0.32
10%	83	33	24	0.32	68	38	26	0.28
20%	75	39	26	0.31	71	41	18	0.29
30%	87	29	24	0.32	81	39	23	0.32

	ไดนามิกไทม์วอร์ปปิง			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	67	22	24	0.25
5%	64	32	20	0.26
10%	58	36	23	0.25
20%	56	31	21	0.23
30%	55	32	17	0.23

จากการผลการทดลองในตารางที่ 4.11 พบว่าแมนฮัตตันและยุคลิเดียนยังคงให้ผลดีกว่าไดนามิกไทม์วอร์ปปิง โดยไดนามิกไทม์วอร์ปปิงจะให้ผลดีกว่าในกรณีขนาดของหน้าต่างเป็น 1% ของความยาวอนุกรมเวลา ซึ่งถ้าสังเกตจากรูปที่ 4.1 ซึ่งเป็นตัวอย่างอนุกรมที่ใช้ขนาดหน้าต่างเป็น 1% ของความยาวอนุกรมเวลา อาจกล่าวได้ว่าไดนามิกไทม์วอร์ปปิงช่วยแก้ปัญหาในกรณีการเยื้องได้ แต่ก็ไม่สามารถช่วยให้ความถูกต้องของการจับคู่เพิ่มขึ้นมากนักเพราะความถูกต้องในการจับคู่คำยังขึ้นกับพารามิเตอร์อื่น ๆ อีกมาก และยังขึ้นอยู่ลักษณะโครงสร้างของคำและภาษาด้วย ซึ่งจะกล่าวถึงรายละเอียดในภายหลัง

นอกจากนี้การใช้ไดนามิกไทม์วอร์ปปิงยังใช้เวลาในการคำนวณมากกว่าแบบแมนฮัตตันและยุคลิเดียนหลายเท่า เพราะเป็นกำหนดการพลวัตใช้เวลาในการคำนวณ $O(n^2)$ แต่แมนฮัตตันและยุคลิเดียนใช้เวลา $O(n)$ ดังนั้นในงานวิจัยนี้จึงเลือกใช้แบบแมนฮัตตันและยุคลิเดียนในการทดลองในบทที่ 5 เพราะทั้งคู่ให้ผลความถูกต้องมากกว่าและใช้เวลาในการคำนวณน้อยกว่า

4.3.4 จำนวนของคำหยุด

เนื่องจากคำหยุดในภาษาอังกฤษมีเป็นจำนวนมาก ดังนั้นผู้วิจัยจึงทำการทดลองสองแบบ คือ แบบกำจัดกลุ่มคำหยุดทั่วไป ตามที่แสดงไว้ในภาคผนวก ก และแบบกำจัดคำหยุดบางคำดังแสดงในตารางที่ 4.12 ซึ่งเป็นคำหยุดที่ใช้ในกูเกิล (Google) ดังนั้นในงานวิจัยนี้จะขอเรียกกลุ่มคำหยุดนี้ว่า กลุ่มคำหยุดกูเกิล ส่วนภาษาไทยมีคำหยุดเพียง 77 คำจึงกำจัดทั้งหมด

ตารางที่ 4.12 กลุ่มคำหยุดกูเกิล

กลุ่มคำหยุดกูเกิล
I a about an are as at be by for from how in is it of on or that the this to was what when will where who with and

<u>การทดลองเบื้องต้นที่ 10</u> การทดลองกำจัดกลุ่มคำหยุดกูเกิลและกำจัดกลุ่มคำหยุดทั่วไปโดยทดลองกับไบเบิลขนาดกลาง

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

ฟังก์ชันระยะห่าง : ยุคลิเดียน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 338 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดกูเกิล) : 439 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 660 คำ

จำนวนประโยค : 175 ประโยค

ตารางที่ 4.13 ผลการทดลองจับคู่ประโยคกรณีกำจัดกลุ่มคำหยุดกูเกิลและกำจัดกลุ่มคำหยุดทั่วไปกับไบเบิลขนาดกลาง

	จำนวนประโยคที่ถูกต้อง	
ขนาดของหน้าต่าง	กำจัดกลุ่มคำหยุดกูเกิล	กำจัดกลุ่มคำหยุดทั่วไป
1%	26	29
5%	56	60
10%	52	59
20%	46	59
30%	55	58

<u>การทดลองเบื้องต้นที่ 11</u> การทดลองกำจัดกลุ่มคำหยุดกูเกิลและกำจัดกลุ่มคำหยุดทั่วไปโดยทดลองกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี

ฟังก์ชันระยะห่าง : ยุคลิเดียน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 912 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดกูเกิล) : 1071 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 1194 คำ

จำนวนประโยค : 304 ประโยค

ตารางที่ 4.14 ผลการทดลองจับคู่ประโยคกรณีกำจัดกลุ่มคำหยุดกูเกิลและกำจัดกลุ่มคำหยุดทั่วไปกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	จำนวนประโยคที่ถูกต้อง	
ขนาดของหน้าต่าง	กำจัดกลุ่มคำหยุดกูเกิล	กำจัดกลุ่มคำหยุดทั่วไป
1%	81	70
5%	111	88
10%	120	106
20%	59	60
30%	36	34

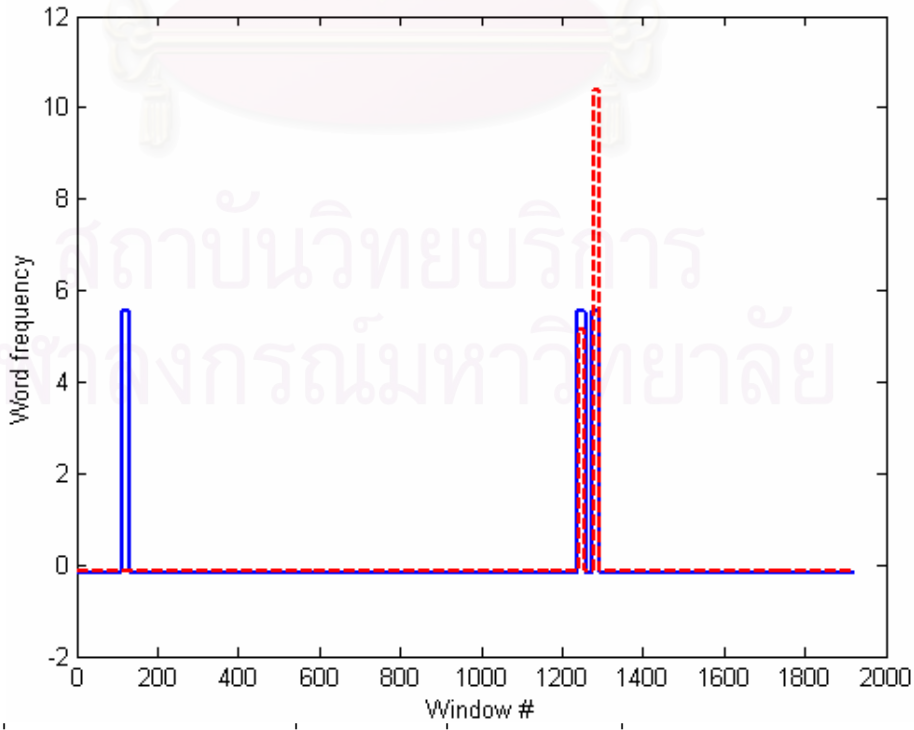
จากผลการทดลองในตารางที่ 4.13 และ 4.14 พบว่าการใช้ไบเบิลขนาดกลางเป็นข้อมูลเข้า จะให้ผลดีเมื่อกำจัดกลุ่มคำหยุดทั่วไป ส่วนการใช้ตัวอย่างประโยคจากดิกชันนารีขนาดกลางเป็นข้อมูลเข้าจะให้ผลดีเมื่อกำจัดกลุ่มคำหยุดกูเกิล

ตัวอย่างประโยคจากดิกชันนารีขนาดกลางมีความยาว 3,884 คำ มีคำภาษาอังกฤษที่แตกต่างกันทั้งหมด 1,096 คำ จะเห็นว่ามีการใช้คำที่หลากหลายมาก ทำให้ความถี่ในการปรากฏของแต่ละคำมีค่าน้อย คำหยุดก็เช่นเดียวกันมีการปรากฏไม่มากนัก ทั้งนี้การกำจัดคำหยุดมีสาเหตุมาจากคำหยุดมักจะเป็นคำที่พบบ่อยในเอกสารและไม่มีความหมายเฉพาะเจาะจง เช่น the on and of for เป็นต้น ซึ่งจะพบบ่อยในเกือบทุกประโยค แต่ก็ยังมีคำหยุดบางส่วนที่ไม่ได้พบบ่อยในเกือบทุกประโยค เช่น every put make give เป็นต้น เมื่อคำหยุดเหล่านี้มาอยู่ในข้อมูลเข้าที่มีการใช้คำหลากหลาย มีความถี่ในการปรากฏของคำทั่ว ๆ ไปน้อย คำหยุดเหล่านี้ก็มีความถี่ในการปรากฏน้อยเช่นกัน จึงมีลักษณะเหมือนคำทั่ว ๆ ไป และมีประโยชน์ในการจับคู่ประโยคเช่นเดียวกัน

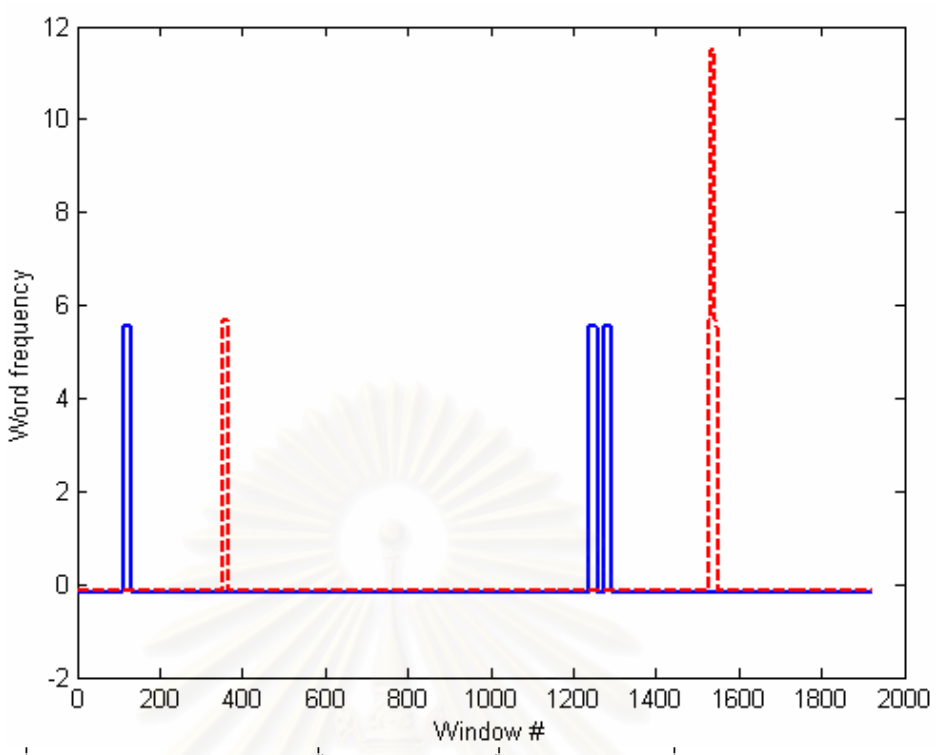
ดังที่กล่าวไปแล้วในข้างต้นว่าในการพิจารณาความถูกต้องของคำที่จับคู่ได้ เราไม่ได้สนใจเฉพาะความหมายของคำเท่านั้น แต่เราสนใจในระดับประโยคด้วยว่าถ้าคำนั้นอยู่ในประโยคเดียวกับคำที่ถูกต้อง เราก็จะถือว่าถูกต้อง เพราะสามารถช่วยเพิ่มคะแนนให้กับประโยคที่ถูกต้องได้ในการจับคู่ประโยค คำหยุดบางคำที่ไม่ได้อยู่ในกลุ่มคำหยุดกูเกิลจึงมีประโยชน์สามารถช่วยเพิ่มคะแนนให้กับประโยคที่ถูกต้องได้ ดังนั้นกรณีใช้ข้อมูลเข้าเป็นตัวอย่างประโยคจากดิกชันนารีจึงได้ผลดีเมื่อกำจัดเพียงกลุ่มคำหยุดกูเกิล

การที่คำทั่วไป หรือคำหยุดที่มีความถี่น้อยสามารถจับคู่ได้กับคำที่อยู่ในประโยคเดียวกับคำที่ถูกต้อง เนื่องจากผลรวมของระยะห่างที่สามารถหาได้จากรูปที่มีเส้นกราฟซ้อนทับกันจะมีค่าที่น้อยกว่าผลรวมของระยะห่างที่หาได้จากรูปที่เส้นกราฟไม่ได้ซ้อนทับกัน ดังแสดงในรูปที่ 4.11 และ 4.12 ดังนั้นคำมีความถี่น้อยจึงจับคู่ได้กับคำที่มีตำแหน่งใกล้เคียงกันในอีกภาษาหนึ่ง ซึ่งก็มีความเป็นไปได้มากว่าจะอยู่ในประโยคเดียวกับคำที่ถูกต้อง

ข้อมูลไบเบิลขนาดกลางที่ใช้ในการทดลองนี้ยาว 4,253 คำ มีการใช้คำภาษาอังกฤษที่แตกต่างกัน 463 คำ จะเห็นว่ามีการใช้คำที่หลากหลายน้อยกว่าข้อมูลตัวอย่างประโยคจากดิกชันนารีขนาดกลาง ดังนั้นความถี่ของแต่ละคำที่ปรากฏก็จะมากกว่า คำหยุดก็เช่นเดียวกัน เช่น คำว่า every มีการใช้ 34 ครั้ง คำว่า put มีการใช้ 23 ครั้ง คำว่า make มีการใช้ 34 ครั้ง และคำว่า give มีการใช้ 47 ครั้ง เป็นต้น ซึ่งเมื่อคำหยุดเหล่านี้ไม่ได้ถูกกำจัดออก และคำหยุดเหล่านี้จับคู่ผิดก็จะมีผลให้การให้คะแนนคู่ประโยคผิดพลาดไปด้วย ดังนั้นสำหรับการใช้ไบเบิลขนาดกลางเป็นข้อมูลเข้า จึงให้ผลดีเมื่อกำจัดกลุ่มคำหยุดทั่วไป



รูปที่ 4.11 แสดงอนุกรมเวลาที่ซ้อนทับกันเนื่องจากเป็นคำที่อยู่ในตำแหน่งใกล้เคียงกัน



รูปที่ 4.12 แสดงอนุกรมเวลาที่ไม่ซ้อนทับกันเนื่องจากเป็นคำที่อยู่ในตำแหน่งห่างกัน

ดังนั้นจึงกล่าวได้ว่า การเลือกกำจัดกลุ่มคำหยุดกูเกิลหรือกลุ่มคำหยุดทั่วไป มีผลกับความถูกต้องในการจับคู่ประโยค โดยแบบใดจะให้ผลดีกว่านั้นก็ขึ้นอยู่กับความถี่ในการปรากฏของคำหยุด ถ้าความถี่น้อยคำหยุดนั้นอาจจะเป็นประโยชน์ในการจับคู่ประโยคดังการทดลองเบื้องต้นที่ 11

4.3.5 **อันดับของคู่คำที่ใช้ในการให้คะแนนคู่ประโยค**

เนื่องจากในการจับคู่คำได้เก็บคู่คำไว้ 3 อันดับ ดังนั้นจึงทำการทดลองให้คะแนนคู่ประโยคในแบบต่าง ๆ ได้แก่ แบบใช้คู่คำอันดับหนึ่งเท่านั้น แบบใช้คู่คำอันดับหนึ่งกับสอง และแบบใช้คู่คำอันดับหนึ่ง สองและสาม ดังการทดลองเบื้องต้นที่ 12 13 และ 14 โดยค่าที่แสดงในตารางคือจำนวนประโยคที่จับคู่ได้ถูกต้อง

<u>การทดลองเบื้องต้นที่ 12</u> การทดลองให้คะแนนคู่ประโยคในแบบต่าง ๆ กับไบเบิลขนาดสั้น

ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)

ฟังก์ชันระยะห่าง : แมนฮัตตัน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 69 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 126 คำ

จำนวนประโยค : 31 ประโยค

ตารางที่ 4.15 ผลการทดลองจับคู่ประโยคเมื่อให้คะแนนคู่ประโยคในแบบต่าง ๆ กับไบเบิลขนาดสั้น

ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	26	27	24
5%	28	28	28
10%	21	24	26
20%	26	25	25
30%	20	22	21

การทดลองเบื้องต้นที่ 13 การทดลองให้คะแนนคู่ประโยคในแบบต่าง ๆ กับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 304 ตัวอย่าง

ฟังก์ชันระยะห่าง : แมนฮัตตัน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 912 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 1194 คำ

จำนวนประโยค : 304 ประโยค

ตารางที่ 4.16 ผลการทดลองจับคู่ประโยคเมื่อให้คะแนนคู่ประโยคในแบบต่าง ๆ กับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	82	80	84
5%	108	107	109
10%	92	95	100
20%	63	67	65
30%	25	32	40

การทดลองเบื้องต้นที่ 14 การทดลองให้คะแนนคู่ประโยคในแบบต่าง ๆ กับข้อมูลกฎหมายขนาดกลาง

ข้อมูลเข้า : ข้อกฎหมาย 102 ข้อ

ฟังก์ชันระยะห่าง : แมนฮัตตัน

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (ไม่รวมกลุ่มคำหยุดทั่วไป) : 449 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (ไม่รวมกลุ่มคำหยุดภาษาไทย) : 552 คำ
จำนวนประโยค : 102 ประโยค

ตารางที่ 4.17 ผลการทดลองจับคู่ประโยคเมื่อให้คะแนนคู่ประโยคในแบบต่าง ๆ กับข้อมูลกฎหมายขนาดกลาง

ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	50	49	48
5%	55	53	52
10%	45	49	48
20%	55	48	47
30%	43	45	48

จากผลการทดลองตารางที่ 4.15 พบว่าทั้งสามแบบให้ผลดีพอ ๆ กันขึ้นอยู่กับขนาดของหน้าต่าง ส่วนผลการทดลองตารางที่ 4.16 พบว่าการใช้คู่คำอันดับที่ 1 2 และ 3 มีแนวโน้มให้ผลดีกว่าแบบอื่น และผลการทดลองตารางที่ 4.17 การใช้คู่คำอันดับที่ 1 มีแนวโน้มให้ผลดีกว่าแบบอื่น จากทั้ง 3 การทดลองอาจกล่าวได้ว่าไม่มีแบบใดที่ให้ผลดีที่สุด แต่ละแบบสามารถให้ผลดีกว่าแบบอื่นได้ ทั้งนี้ขึ้นอยู่กับปัจจัยอื่น ๆ ด้วย เช่น ขนาดของหน้าต่าง ชนิดและลักษณะของข้อมูลที่นำมาใช้ในการทดลอง เป็นต้น

จากการวิเคราะห์ลักษณะของข้อมูลที่นำมาใช้ในการทดลองพบว่า ข้อมูลที่เป็นข้อกฎหมายจะเป็นภาษาทางการ ใช้คำศัพท์ที่เป็นทางการ ไม่ใช้คำฟุ่มเฟือยเหมือนข้อความทั่วไป จำนวนคำที่แตกต่างกันจะมากหรือน้อยขึ้นอยู่กับจำนวนหมวดกฎหมายที่เลือก ถ้าเลือกหลายหมวดก็จะมีคำที่แตกต่างกันมาก แต่กฎหมายหมวดเดียวกันจะใช้คำซ้ำ ๆ กัน เช่น หมวดรัฐสภาจะพบคำที่เกี่ยวกับ รัฐสภา วุฒิสภา รัฐธรรมนูญ เป็นต้น ถ้าเป็นหมวดศาล จะพบคำที่เกี่ยวกับ ศาล ตุลาการ ผู้พิพากษา เป็นต้น คำไหนที่ใช้ซ้ำบ่อย ๆ จะมีความถี่มากและมีโอกาสจับคู่คำได้มาก เพราะคำเหล่านี้เมื่ออยู่ในภาษาอังกฤษก็จะใช้เป็นคำเดียวกันตลอดทั้งเอกสาร ไม่มีการเล่นคำ เพราะเป็นเอกสารที่เป็นทางการ ดังนั้นคู่คำในอันดับที่ 1 จึงมีแนวโน้มที่จะจับคู่ได้ถูกต้องมาก เมื่อจับคู่ประโยคโดยคิดคะแนนจากคู่คำในอันดับที่ 1 จึงจับคู่ประโยคได้ถูกต้องมาก ส่วนข้อมูลที่เป็นตัวอย่างประโยคจากดิกชันนารี เนื่องจากในดิกชันนารีมีคำศัพท์เป็นจำนวนมาก และตัวอย่างประโยคที่แต่งขึ้นมาสำหรับคำศัพท์แต่ละคำก็ไม่จำเป็นต้องเป็นเรื่องเดียวกัน ดังนั้นจึงมีการใช้คำศัพท์ที่หลากหลาย ความถี่ในการปรากฏของคำจะน้อย คำหนึ่งคำอาจจะแปลได้หลายอย่าง มีการใช้คำฟุ่มเฟือย ซึ่งเหล่านี้ล้วนมีผลต่อการจับคู่คำทำให้คู่คำที่ถูกต้องอาจจะเป็นคู่คำในอันดับ

ใดก็ได้ ดังนั้นการจับคู่ประโยคโดยคิดคะแนนจากคู่คำทั้ง 3 อันดับจึงให้ผลดีกว่าการใช้คู่คำอันดับที่ 1 เพียงอย่างเดียว เพราะถ้าคู่คำอันดับที่ 1 ไม่ถูกก็ยังมีคะแนนจากคู่คำอันดับที่ 2 และ 3 มาช่วยได้ ดังนั้นอาจกล่าวได้ว่า การคิดคะแนนคู่ประโยคจากคู่คำอันดับต่างกัน มีผลต่อความถูกต้องในการจับคู่ประโยค โดยถ้าข้อความเข้ามีการใช้คำที่หลากหลาย ความหมายไม่แน่นอน ควรคิดคะแนนจากคู่คำทั้งสามอันดับ ส่วนข้อมูลเข้าที่มีเนื้อหาเป็นทางการ ไม่มีการเล่นคำ แต่ละคำมีความหมายชัดเจน สามารถคิดคะแนนจากคู่คำอันดับหนึ่งเพียงอันดับเดียวได้

4.3.6 อัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและคำเนื้อหาในภาษาไทย

เนื่องจากการจับคู่แบบ 1:N นั้น เราไม่ทราบว่าหนึ่งประโยคภาษาไทยจะประกอบด้วยกี่วรรค ดังนั้นจึงต้องหาวิธีที่จะช่วยในการหาจำนวนวรรค โดยในที่นี้จะใช้การนับจำนวนคำเนื้อหา โดยตั้งสมมติฐานว่า จำนวนคำเนื้อหาในแต่ละประโยคภาษาอังกฤษและภาษาไทยที่เป็นคู่กันควรจะใกล้เคียงกัน แต่อย่างไรก็ตาม ไม่สามารถระบุได้ว่าจำนวนคำเนื้อหาในประโยคภาษาอังกฤษและภาษาไทยจะเท่ากันหรือมากน้อยกว่ากันเพียงใด ผู้วิจัยจึงได้ทำการทดลองเบื้องต้นที่ 15 เพื่อดูว่าการกำหนดอัตราส่วนระหว่างจำนวนคำเนื้อหาในประโยคภาษาอังกฤษและภาษาไทยเป็นเท่าใด จึงจะให้ผลดีและเหมาะสำหรับใช้ในการหาจำนวนวรรค โดยอัตราส่วนที่ใช้จะเป็นจำนวนคำเนื้อหาในประโยคภาษาอังกฤษต่อจำนวนคำเนื้อหาในประโยคภาษาไทย เช่น การใช้อัตราส่วน 1.2 จะหมายถึง ถ้าประโยคภาษาอังกฤษที่กำลังพิจารณามีคำเนื้อหา 12 คำ จะตัดประโยคภาษาไทยเมื่อนับคำเนื้อหาในภาษาไทยได้ 10 คำ เป็นต้น

โดยในการทดลองจะให้ 1 คะแนนกับประโยคที่สามารถจับคู่ได้ถูกต้อง และประโยคที่จับคู่ได้เนื้อหาครบถ้วน แม้จะมีวรรคเกินมาบ้าง ซึ่งในกรณีนี้จะเรียกว่า ครบ ส่วนประโยคที่จับคู่ได้เพียงบางส่วน หรือมีบางวรรคขาดหายไปจะให้คะแนน 0.5 ในกรณีนี้จะเรียกว่า ขาด และการหาค่าเฉลี่ยนั้นจะได้จากการนำคะแนนทั้งกรณีครบและขาดมาบวกกันและหารด้วยจำนวนประโยคทั้งหมด

<u>การทดลองเบื้องต้นที่ 15</u> การทดลองกำหนดอัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและภาษาไทยให้เป็นค่าต่าง ๆ โดยทดลองกับไบเบิลขนาดสั้น ไบเบิลขนาดกลาง และข้อกฎหมายขนาดสั้น

ตารางที่ 4.18 ผลการทดลองจับคู่ประโยคโดยใช้อัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและภาษาไทยค่าต่าง ๆ

	ไบเบิลขนาดสั้น			ไบเบิลขนาดกลาง			ข้อกฎหมายขนาดสั้น			
อัตราส่วน	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	เฉลี่ยรวม
0.7	5	4.5	**0.31**	10	20.5	**0.17**	7	4	0.22	**0.23**
0.8	5	4	0.29	10	17.5	0.16	9	3.5	0.25	**0.23**
0.9	6	3.5	**0.31**	10	10.5	0.12	9	4	0.26	**0.23**
1.0	6	3.5	**0.31**	10	8	0.10	8	3.5	0.23	0.21
1.1	5	3.5	0.27	10	7.5	0.10	8	3	0.22	0.20
1.2	6	2.5	0.27	11	6	0.10	9	3.5	0.25	0.21
1.3	6	2	0.26	11	5	0.09	10	3.5	**0.27**	0.21

จากผลการทดลองในตารางที่ 4.18 พบว่าค่าเฉลี่ยรวมของความถูกต้องมากที่สุด คือ 0.23 แต่เนื่องจากมีอัตราส่วน 0.7 0.8 และ 0.9 ที่ได้ค่าเฉลี่ยรวมเท่ากัน ดังนั้นในงานวิจัยนี้ผู้วิจัยจะเลือกใช้ค่า 0.8 เนื่องจากเป็นค่ากึ่งกลางระหว่าง 0.7 และ 0.9

จากการทดลองเบื้องต้นทั้งหมด สามารถสรุปได้ว่าพารามิเตอร์ทั้ง 6 ตัวที่กล่าวถึงในบทนี้ล้วนมีผลต่อความถูกต้องในการจับคู่คำและจับคู่ประโยค ดังนั้นจึงนำพารามิเตอร์เหล่านี้ไปใช้กับการทดลองในบทที่ 5 ต่อไป

บทที่ 5

การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงการทดลองจับคู่คำและจับคู่ประโยคตามแนวคิดที่ได้นำเสนอไปแล้ว เพื่อดูว่าวิธีที่นำเสนอจะมีความสามารถในการจับคู่คำและจับคู่ประโยคได้มากน้อยเพียงไร โดยแบ่งการทดลองออกเป็น 4 ส่วนได้ แก่ การจับคู่คำ การจับคู่ประโยคแบบ 1 : 1 การเปรียบเทียบกับวิธีอื่นที่ไม่ใช้อนุกรมเวลา และการจับคู่ประโยคแบบ 1 : N โดยเปอร์เซ็นต์ความถูกต้องที่มากที่สุดในแต่ละการทดลองจะแสดงเป็นตัวหนา

5.1 การทดลองจับคู่คำ

ในการจับคู่คำจะทำการทดลองกับข้อมูลเข้าดังนี้ ไบเบิลขนาดสั้น ไบเบิลขนาดกลาง ไบเบิลขนาดยาว ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง ข้อกฎหมายขนาดสั้น และข้อกฎหมายขนาดกลาง โดยทดลองกับหน้าต่างขนาด 1% 5% 10% 20% และ30% ของความยาวอนุกรมเวลา โดยลดความยาวอนุกรมเวลาด้วยอัตราส่วน 0.5 ทดลองใช้ฟังก์ชันระยะห่าง 2 แบบ คือ แมนฮัตตันและยุคลิเดียน โดยวัดความถูกต้องด้วยการหาค่าเฉลี่ยของส่วนกลับลำดับชั้น แล้วจึงจะทำการวิเคราะห์ผลโดยรวม

<u>การทดลองที่ 1</u> การทดลองจับคู่คำกับไบเบิลขนาดสั้น

ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 69 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดกูเกิล) : 110 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 126 คำ

ตารางที่ 5.1 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับไบเบิลขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	63	14	6	0.65	43	10	3	0.71
5%	75	16	3	**0.76**	56	6	2	**0.86**
10%	70	15	6	0.72	49	10	2	0.79
20%	72	12	5	0.72	51	6	3	0.78
30%	66	14	7	0.68	46	9	5	0.76
	ค่าเฉลี่ยความถูกต้อง			0.70	ค่าเฉลี่ยความถูกต้อง			0.78

ตารางที่ 5.2 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับไบเบิลขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	32	33	19	0.50	21	20	11	0.50
5%	71	15	12	**0.75**	50	7	6	**0.80**
10%	66	18	7	0.70	47	10	2	0.76
20%	70	10	6	0.70	50	5	2	0.77
30%	64	19	6	0.69	47	10	2	0.76
	ค่าเฉลี่ยความถูกต้อง			0.67	ค่าเฉลี่ยความถูกต้อง			0.72

<u>การทดลองที่ 2</u> การทดลองจับคู่คำกับไบเบิลขนาดกลาง

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 338 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกำจัดกลุ่มคำหยุดกูเกิล) : 439 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 660 คำ

ตารางที่ 5.3 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับไบเบิลขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	74	46	36	0.25	59	30	29	0.25
5%	111	51	47	**0.35**	93	36	31	**0.36**
10%	102	46	35	0.31	83	33	24	0.32
20%	94	53	39	0.30	75	39	26	0.31
30%	106	38	32	0.31	87	29	24	0.32
	ค่าเฉลี่ยความถูกต้อง			0.30	ค่าเฉลี่ยความถูกต้อง			0.31

ตารางที่ 5.4 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับไบเบิลขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	67	25	28	0.20	48	26	18	0.20
5%	96	48	38	**0.30**	81	36	23	**0.32**
10%	84	54	37	0.28	68	38	26	0.28
20%	92	55	27	0.29	71	41	18	0.29
30%	94	54	38	**0.30**	81	39	23	**0.32**
	ค่าเฉลี่ยความถูกต้อง			0.27	ค่าเฉลี่ยความถูกต้อง			0.28

<u>การทดลองที่ 3</u> การทดลองจับคู่คำกับไบเบิลขนาดยาว

ข้อมูลเข้า : บทที่ 1 ถึง 25 ของปฐมกาล (Genesis)

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 771 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดกูเกิล) : 902 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 1,465 คำ

ตารางที่ 5.5 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับไบเบิลขนาดยาว

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	174	68	46	0.25	154	55	41	0.25
5%	210	69	44	**0.29**	183	64	38	**0.30**
10%	208	70	45	**0.29**	193	48	40	**0.30**
20%	200	72	56	0.28	182	64	46	**0.30**
30%	143	56	47	0.21	136	45	35	0.22
	ค่าเฉลี่ยความถูกต้อง			0.26	ค่าเฉลี่ยความถูกต้อง			0.27

ตารางที่ 5.6 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับไบเบิลขนาดยาว

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	131	56	35	0.19	116	47	34	0.20
5%	183	61	44	0.25	165	47	39	0.26
10%	194	66	46	0.27	179	50	37	0.28
20%	209	60	52	**0.28**	191	52	40	**0.30**
30%	137	55	38	0.20	125	47	28	0.20
	ค่าเฉลี่ยความถูกต้อง			0.24	ค่าเฉลี่ยความถูกต้อง			0.25

<u>การทดลองที่ 4</u> การทดลองจับคู่คำกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 70 ตัวอย่าง

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 302 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดกูเกิล) : 391 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 452 คำ

ตารางที่ 5.7 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	93	32	10	0.29	70	23	6	0.28
5%	143	42	19	0.44	108	32	13	0.42
10%	144	68	22	**0.47**	115	50	14	**0.48**
20%	70	36	37	0.26	48	28	32	0.24
30%	71	31	30	0.25	57	20	24	0.25
	ค่าเฉลี่ยความถูกต้อง			0.34	ค่าเฉลี่ยความถูกต้อง			0.33

ตารางที่ 5.8 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	97	34	14	0.30	79	21	12	0.31
5%	102	57	16	0.35	78	45	9	0.34
10%	126	65	46	**0.44**	99	49	34	**0.45**
20%	72	42	38	0.27	52	32	33	0.26
30%	73	34	33	0.25	54	24	28	0.25
	ค่าเฉลี่ยความถูกต้อง			0.32	ค่าเฉลี่ยความถูกต้อง			0.32

<u>การทดลองที่ 5</u> การทดลองจับคู่คำกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 304 ตัวอย่าง

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 912 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดกูเกิล) : 1071 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 1,181 คำ

ตารางที่ 5.9 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	194	87	60	0.24	154	72	55	0.23
5%	199	117	87	**0.27**	167	103	73	**0.27**
10%	170	112	90	0.24	149	97	78	0.25
20%	91	42	47	0.12	77	35	49	0.12
30%	37	30	38	0.06	32	25	35	0.06
	ค่าเฉลี่ยความถูกต้อง			0.19	ค่าเฉลี่ยความถูกต้อง			0.19

ตารางที่ 5.10 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	157	82	55	0.20	138	66	51	0.21
5%	166	142	102	0.25	147	126	87	**0.26**
10%	194	104	94	**0.26**	167	92	81	**0.26**
20%	84	46	39	0.11	69	43	38	0.11
30%	26	25	42	0.05	21	18	37	0.05
	ค่าเฉลี่ยความถูกต้อง			0.17	ค่าเฉลี่ยความถูกต้อง			0.18

<u>การทดลองที่ 6</u> การทดลองจับคู่คำกับข้อกฎหมายขนาดสั้น

ข้อมูลเข้า : ข้อกฎหมาย 50 ข้อ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 263 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดกูเกิล) : 298 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 317 คำ

ตารางที่ 5.11 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับข้อกฎหมายขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	106	47	22	0.46	92	44	16	0.45
5%	143	52	30	**0.60**	124	46	27	**0.59**
10%	140	53	20	0.58	121	44	20	0.57
20%	137	34	27	0.55	120	26	25	0.54
30%	138	51	32	0.58	120	45	29	0.58
	ค่าเฉลี่ยความถูกต้อง			0.55	ค่าเฉลี่ยความถูกต้อง			0.55

ตารางที่ 5.12 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับข้อกฎหมายขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	100	31	18	0.41	87	28	10	0.40
5%	132	41	28	0.52	115	35	27	0.54
10%	137	49	23	**0.57**	118	43	22	**0.56**
20%	143	32	29	**0.57**	124	28	29	**0.56**
30%	130	51	34	0.56	113	45	31	0.55
	ค่าเฉลี่ยความถูกต้อง			0.53	ค่าเฉลี่ยความถูกต้อง			0.52

<u>การทดลองที่ 7</u> การทดลองจับคู่คำกับข้อกฎหมายขนาดกลาง

ข้อมูลเข้า : ข้อกฎหมาย 102 ข้อ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดทั่วไป) : 449 คำ

จำนวนคำที่แตกต่างกันในภาษาอังกฤษ (กรณีกำจัดกลุ่มคำหยุดกูเกิล) : 528 คำ

จำนวนคำที่แตกต่างกันในภาษาไทย (กำจัดกลุ่มคำหยุดภาษาไทย) : 545 คำ

ตารางที่ 5.13 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับข้อกฎหมายขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	173	77	36	0.42	151	51	24	0.41
5%	188	77	49	**0.46**	175	60	44	**0.49**
10%	192	73	43	**0.46**	147	59	33	0.42
20%	193	67	41	0.45	171	54	28	0.46
30%	186	79	54	**0.46**	170	56	37	0.47
	ค่าเฉลี่ยความถูกต้อง			0.45	ค่าเฉลี่ยความถูกต้อง			0.45

ตารางที่ 5.14 ผลการทดลองจับคู่คำโดยใช้ฟังก์ชันระยะห่างแบบยุคลิเดียนกับข้อกฎหมายขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล				กำจัดกลุ่มคำหยุดทั่วไป			
ขนาดของหน้าต่าง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	%ความถูกต้อง
1%	156	72	32	0.38	134	51	30	0.38
5%	183	74	44	0.44	177	42	43	0.47
10%	192	72	43	0.46	149	63	34	0.43
20%	197	83	34	**0.47**	169	69	32	**0.48**
30%	189	72	53	0.46	176	48	45	**0.48**
	ค่าเฉลี่ยความถูกต้อง			0.44	ค่าเฉลี่ยความถูกต้อง			0.45

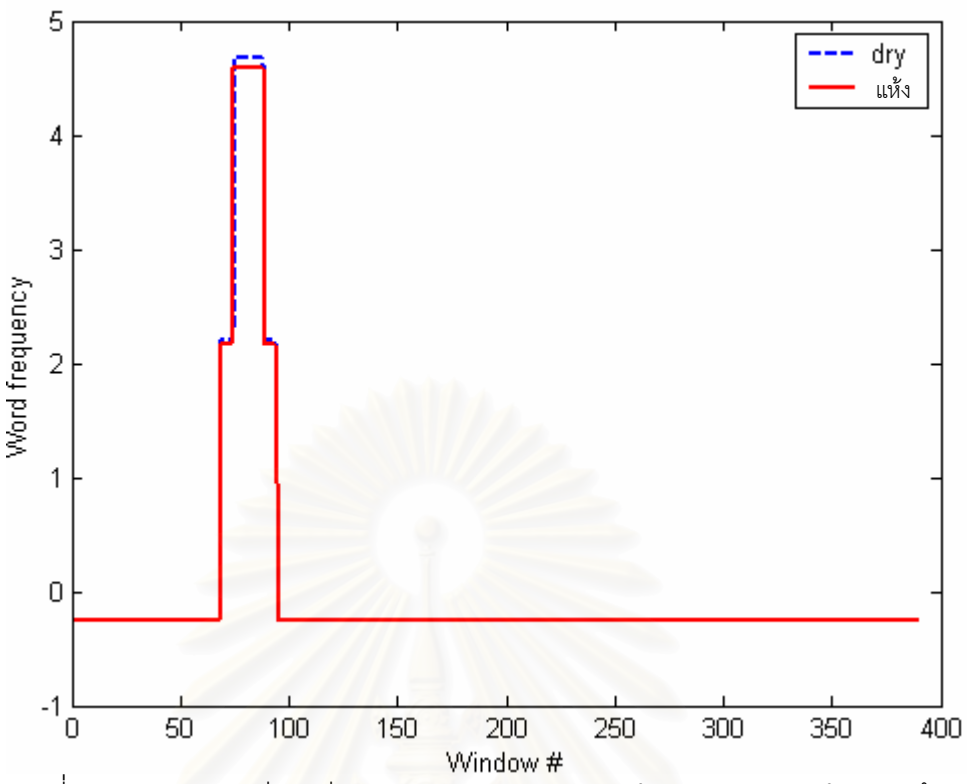
5.2 วิเคราะห์ผลการทดลองจับคู่คำ

จากผลการทดลองพบว่าการใช้ไบเบิลขนาดสั้นเป็นข้อมูลเข้าได้เปอร์เซ็นต์ความถูกต้องในการจับคู่คำมากที่สุดคือ 86% รองลงมา ได้แก่ การใช้ข้อกฎหมายขนาดสั้นเป็นข้อมูลเข้าได้ความถูกต้อง 60% จากการสังเกตพบว่าข้อมูลเข้าที่มีขนาดสั้นจะให้ผลดีกว่าข้อมูลขนาดยาวเนื่องจากข้อมูลสั้นจะมีความถี่ในการปรากฏของแต่ละคำน้อยกว่าข้อมูลขนาดยาว ซึ่งความถี่น้อยมีผลดีกับการจับคู่คำดังที่กล่าวไปแล้วในหัวข้อ 4.3.4 (รูปที่ 4.11 และ 4.12) นอกจากนี้ ข้อมูลขนาดสั้นมีแนวโน้มการใช้คำที่แตกต่างกันน้อยกว่าข้อมูลขนาดยาวเช่น ไบเบิลขนาดสั้นใช้คำศัพท์ภาษาไทยที่แตกต่างกัน 160 คำ ในขณะที่ไบเบิลขนาดยาวใช้คำศัพท์ภาษาไทยที่แตกต่างกัน 1,529 คำ ซึ่งการใช้คำศัพท์ภาษาไทยที่หลากหลาย จะทำให้คำศัพท์ภาษาอังกฤษแต่ละคำมีตัวเลือกในการจับคู่มาก และมีโอกาสจับคู่ผิดได้มาก แต่บางครั้งข้อมูลที่ยาวก็มีการใช้คำศัพท์ที่แตกต่างกันน้อยกว่าข้อมูลที่สั้นกว่า เช่น ไบเบิลขนาดยาวมีภาษาอังกฤษยาว 17,014 คำ แต่ใช้คำศัพท์ที่แตกต่างกันเพียง 927 คำ ในขณะที่ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง มีภาษาอังกฤษยาว 3,884 คำ แต่มีการใช้คำศัพท์ภาษาอังกฤษที่แตกต่างกันถึง 1,096 คำ ซึ่งจากสาเหตุนี้ ทำให้การใช้ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลางเป็นข้อมูลเข้า จับคู่คำได้น้อยกว่าข้อมูลเข้าอื่น โดยมีความถูกต้องเพียง 27%

จำนวนการใช้คำศัพท์ที่แตกต่างกัน ไม่ได้แปรผันตามขนาดความยาวของข้อความเท่านั้น แต่ยังขึ้นอยู่กับประเภทของข้อความด้วยเช่น ข้อความประเภทตัวอย่างประโยคจากดิกชันนารีจะมีการใช้คำที่แตกต่างกันมากกว่าข้อมูลประเภทคัมภีร์ทางศาสนา และข้อมูลประเภทกฎหมาย

ดังนั้นความถูกต้องในการจับคู่คำ นอกจากจะขึ้นอยู่กับพารามิเตอร์ต่าง ๆ ดังที่กล่าวไปแล้วในหัวข้อ 4.3 ความถูกต้องยังขึ้นอยู่กับความยาวของข้อมูลเข้า ประเภทของข้อมูลเข้า และจำนวนคำศัพท์ที่แตกต่างกันในข้อความอีกด้วย

เริ่มแรกได้ตั้งสมมติฐานไว้ว่า คำในภาษาอังกฤษและภาษาไทยที่เป็นคู่กัน ควรจะมีรูปร่างของกราฟที่ใกล้เคียงกัน มีความถี่และตำแหน่งในการปรากฏคำไม่ต่างกันมากนัก เช่น คำว่า dry จับคู่ได้กับคำว่าแห้ง ซึ่งกราฟการปรากฏของทั้งสองคำใกล้เคียงกันมากดังรูปที่ 5.1



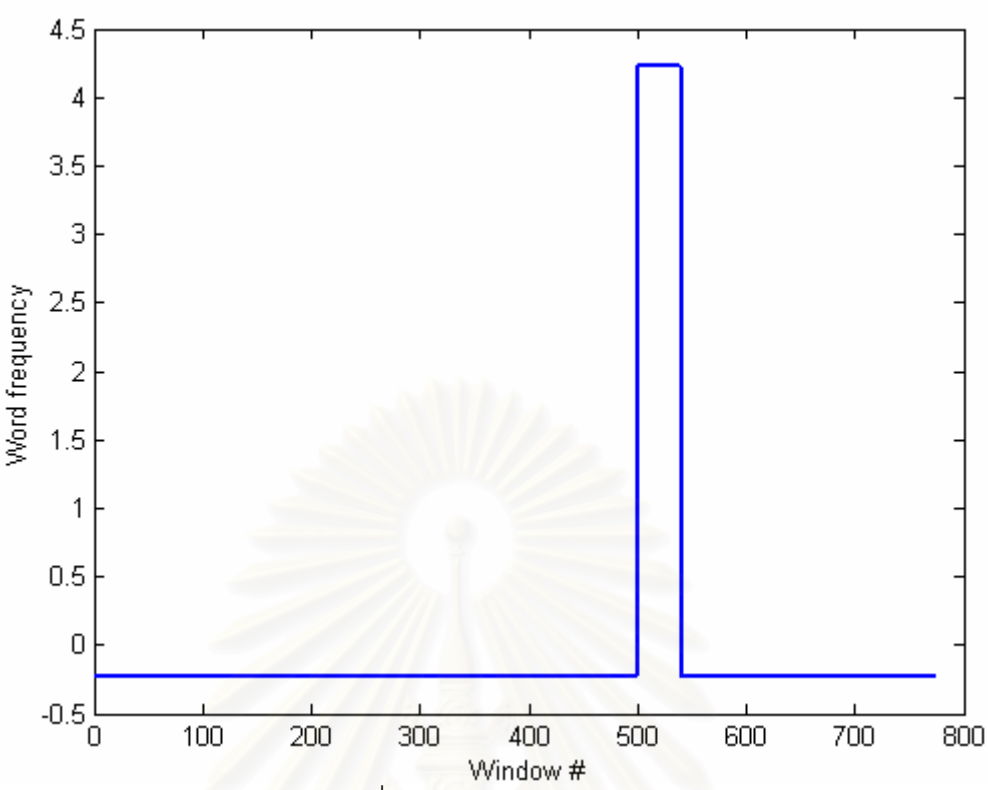
รูปที่ 5.1 แสดงการเปรียบเทียบรูปร่างอนุกรมเวลาของคำว่า “dry” และคำว่า “แห้ง”

แต่จากการทดลองพบว่า คำที่เป็นคู่กันและได้กราฟลักษณะคล้ายกันดังตัวอย่างข้างต้นนั้นมีไม่มากนักเนื่องจากหลาย ๆ ปัจจัยดังนี้

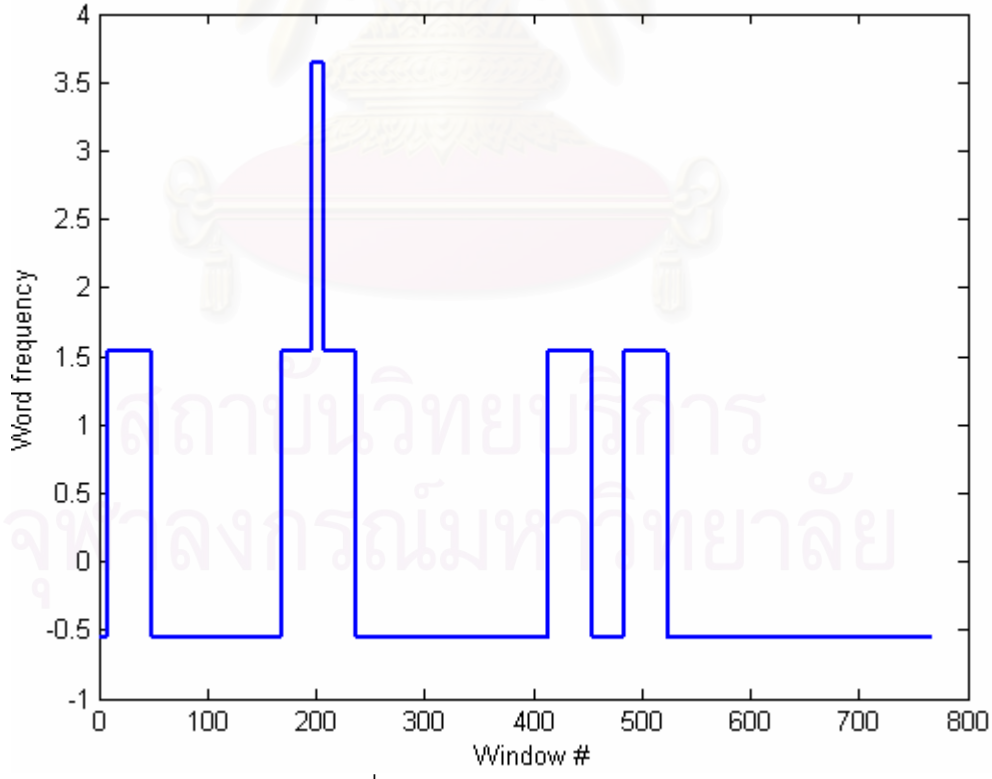
1. การแปลข้อความไม่ได้แปลแบบคำต่อคำ บางครั้งเป็นการแปลโดยรวมทั้งประโยค โดยเลือกใช้คำศัพท์ตามความเหมาะสมกับแต่ละประโยค เช่น คำว่า birth แปลเป็นภาษาไทยได้คำว่า เกิด แต่ยังมีภาษาอังกฤษอีกหลายคำที่สามารถแปลเป็นคำว่าเกิดได้ ดังตารางที่ 5.15 ทำให้กราฟของคำว่า birth และกราฟของคำว่า เกิด มีรูปร่างต่างกันดังแสดงในรูปที่ 5.2 และ 5.3

ตารางที่ 5.15 ตัวอย่างประโยคที่ไม่ได้แปลแบบคำต่อคำ

ภาษาอังกฤษ	ภาษาไทย
And God said, Let there be light: and **there was** light	พระเจ้าตรัสว่า "จงให้มีความสว่าง" แล้วความสว่างก็**เกิด**ขึ้น
And God said, Let grass **come up** on the earth	พระเจ้าตรัสว่า "จงให้แผ่นดิน**เกิด**ต้นหญ้า”
And every sort of living and moving thing **with** which the waters were full	บรรดาสัตว์ที่มีชีวิตแหวกว่ายไปมาตามชนิดของมัน**เกิด**ขึ้นบริบูรณ์ในน้ำนั้น



รูปที่ 5.2 กราฟของคำว่า “birth”

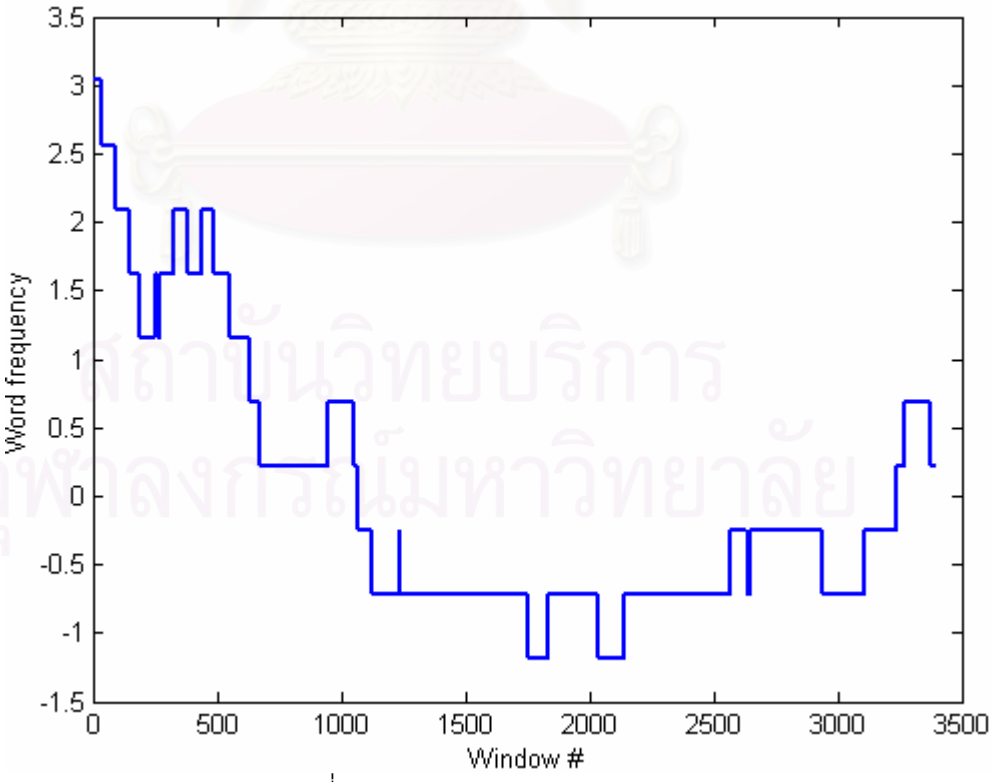


รูปที่ 5.3 กราฟของคำว่า “เกิด”

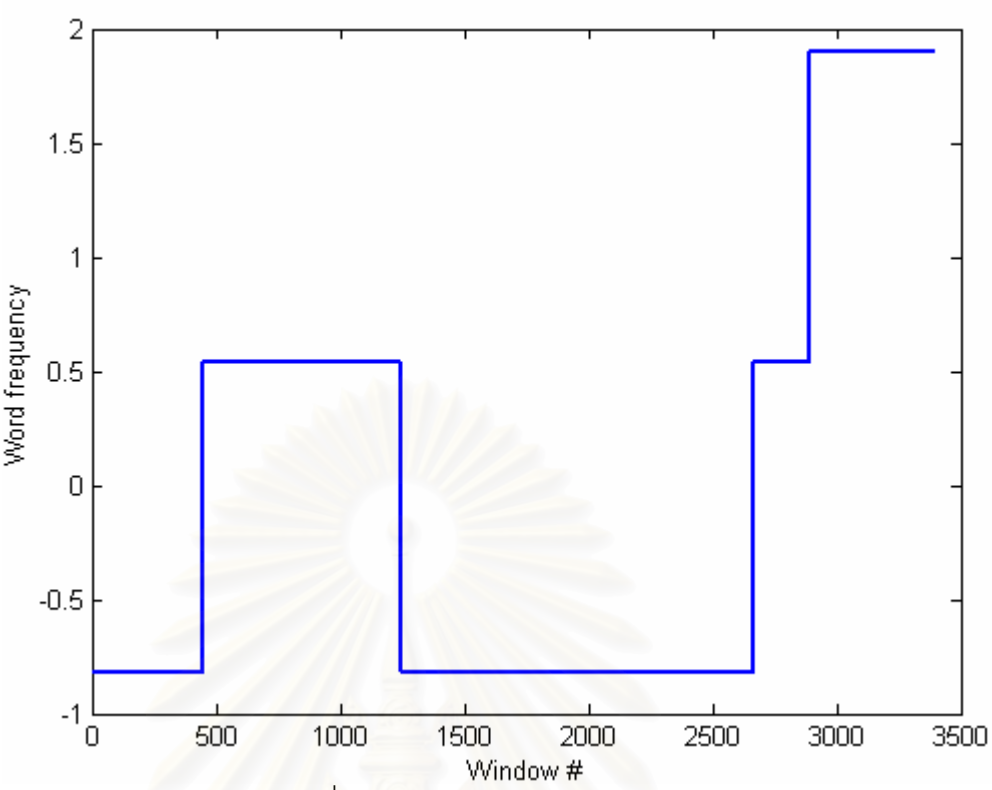
2. กลุ่มคำในภาษาอังกฤษสามารถแปลความหมายเป็นหนึ่งคำในภาษาไทยได้ เช่น “And God said, Let the earth give birth to all sorts of living things, cattle and all things moving on the earth, and beasts of the earth after their sort and it was so.” แปลได้เป็น “พระเจ้าตรัสว่า จงให้แผ่นดินโลกเกิดสัตว์ที่มีชีวิตตามชนิดของมัน สัตว์ใช้งาน สัตว์เลื้อยคลาน และสัตว์ป่าบนแผ่นดินโลกตามชนิดของมัน ก็เป็นดังนั้น” จะเห็นได้ว่ากลุ่มคำว่า “all things moving on the earth” แปลเป็นคำว่า “สัตว์เลื้อยคลาน” ทำให้การจับคู่คำเป็นไปได้ยาก

3. ภาษาอังกฤษหนึ่งคำสามารถแปลเป็นกลุ่มคำในภาษาไทยได้ และกลุ่มคำที่แปลได้นั้นอาจจะประกอบด้วยคำที่เหมือนกับคำที่แปลได้จากภาษาอังกฤษคำอื่น เช่น คำว่า desk แปลว่า โต๊ะเรียน คำว่า class แปลว่า ห้องเรียน คำว่า study แปลว่า เรียน ทำให้จำนวนการเกิดคำว่า study และคำว่าเรียนไม่เท่ากัน จะเจอคำว่า เรียน มากกว่า study หลายครั้ง คำว่า study จึงจับคู่ได้กับคำอื่นที่ไม่ถูกต้องแทน

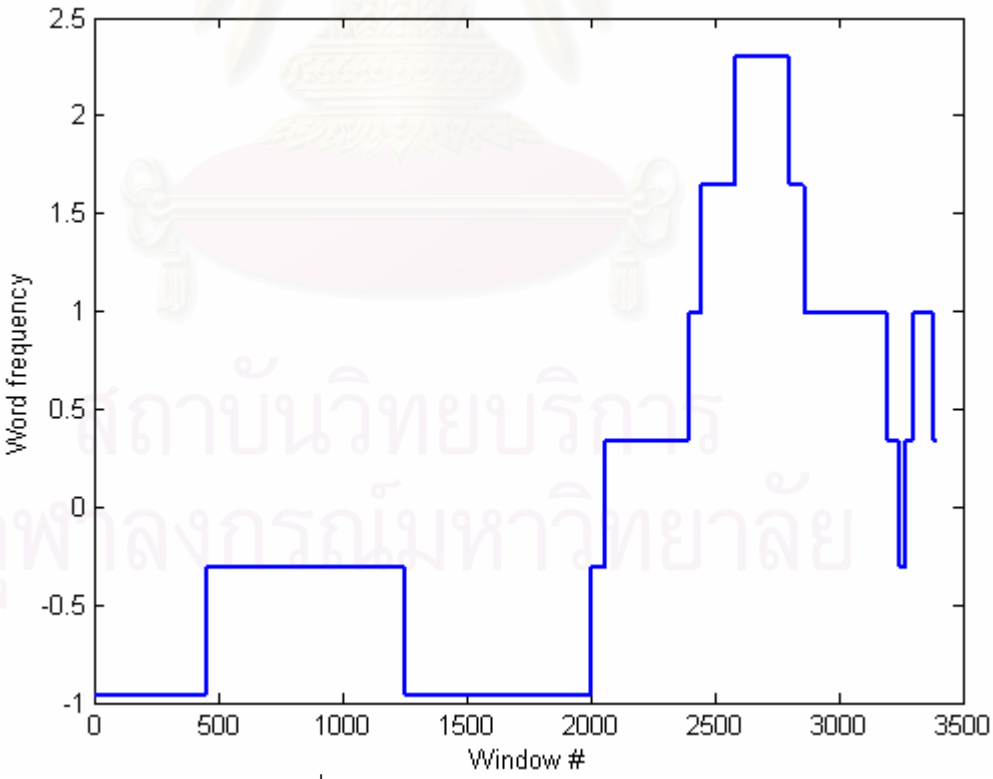
4. การใช้คำราชาศัพท์ เช่น คำว่า say สามารถแปลเป็นคำว่า ตรัส กล่าวหรือ พูด ขึ้นอยู่กับประธานในประโยค ดังนั้นจึงทำให้ความถี่ในการเกิดคำว่า say มีค่าใกล้เคียงกับความถี่ในการเกิดคำว่าตรัส คำว่ากล่าวและคำว่าพูด รวมกัน ดังแสดงในรูปที่ 5.4 ถึง 5.7 ทำให้คำว่า say จับคู่ได้กับคำอื่น แต่ถ้าทั้งข้อความนั้นใช้คำราชาศัพท์ทั้งหมดจะสามารถจับคู่คำว่า say กับคำว่าตรัสได้



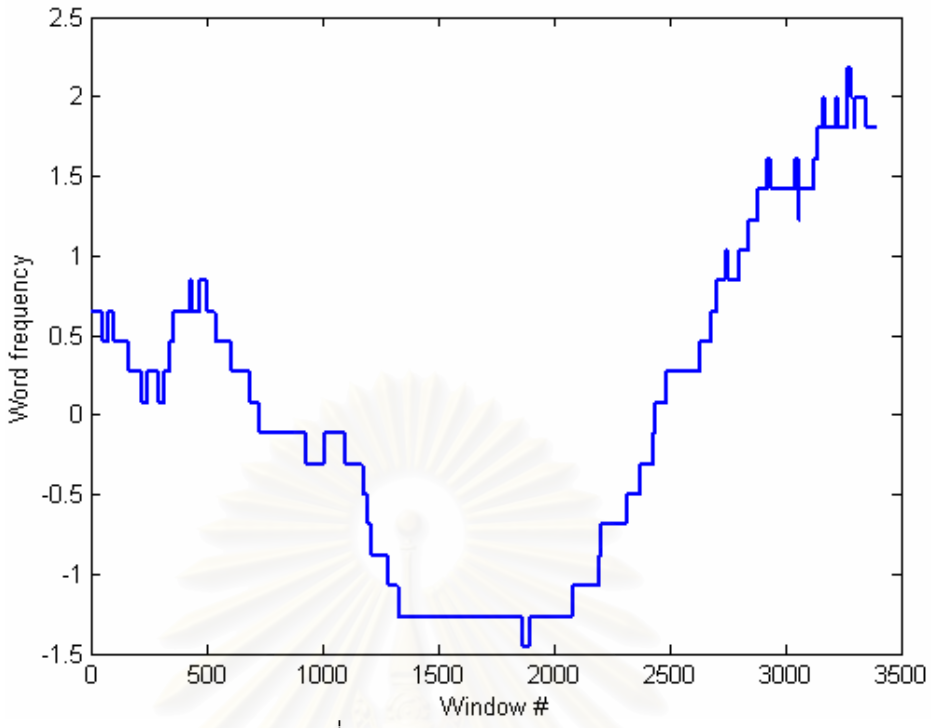
รูปที่ 5.4 กราฟของคำว่า “ตรัส”



รูปที่ 5.5 กราฟของคำว่า “พูด”



รูปที่ 5.6 กราฟของคำว่า “กล่าว”



รูปที่ 5.7 กราฟของคำว่า “say”

5. คำศัพท์ภาษาอังกฤษหลายคำสามารถแปลเป็นภาษาไทยคำเดียวกันได้ เช่น คำว่า earth และ land ถูกแปลเป็นคำว่า แผ่นดิน เหมือนกัน ตัวอย่างแสดงดังตารางที่ 5.16

ตารางที่ 5.16 ตัวอย่างคำศัพท์ภาษาอังกฤษที่สามารถแปลเป็นภาษาไทยคำเดียวกันได้

ภาษาอังกฤษ	ภาษาไทย
And God said, let grass come up on the **earth**	พระเจ้าตรัสว่า จงให้**แผ่นดิน**เกิดต้นหญ้า
On the far side of Jordan in the **land** of Moab, Moses gave the people this law, saying,	โมเสสได้เริ่มอธิบายพระราชบัญญัตินี้ที่ใน**แผ่นดิน**โมอับฟากแม่น้ำจอร์แดนข้างนี้ว่า

6. คำศัพท์แต่ละคำสามารถแปลได้หลายความหมาย มีทั้งที่ความหมายใกล้เคียงกัน เช่น คำว่า land แปลว่า แผ่นดิน หรือประเทศ เป็นต้น และที่ความหมายต่างกันโดยสิ้นเชิง เช่น row แปลว่า แถว หรือ พายเรือ เป็นต้น

7. โครงสร้างประโยคภาษาอังกฤษและภาษาไทยมีความแตกต่างกัน รวมถึงการเรียงประโยคที่แตกต่างกัน ดังตัวอย่างในตารางที่ 5.17 จะเห็นว่าในภาษาไทยมีการกล่าวถึงช่วงเวลาตอนต้นประโยค ในขณะที่ภาษาอังกฤษกล่าวถึงช่วงเวลาในตอนกลางของประโยค ทำให้กราฟที่

ได้มีตำแหน่งที่ความคลาดเคลื่อนกัน ซึ่งความคลาดเคลื่อนนี้มีผลทำให้ผลรวมระยะห่างที่คำนวณได้จากฟังก์ชันระยะห่างมีค่ามากขึ้น ดังนั้นจึงทำให้คำศัพท์ภาษาอังกฤษจับคู่ได้กับคำที่ไม่ถูกต้อง

ตารางที่ 5.17 ตัวอย่างการเรียงประโยคที่แตกต่างกันในภาษาอังกฤษและภาษาไทย

ภาษาอังกฤษ	ภาษาไทย
And the Lord said to Moses in the waste land of Sinai, in the Tent of meeting, **on the first day of the second month**, in the second year after they came out of the land of Egypt,	ณ **วันที่หนึ่งเดือนที่สองปีที่สอง**ตั้งแต่เขาทั้งหลายออกจากประเทศอียิปต์ พระเยโฮวาห์ตรัสกับโมเสสในพลับพลาแห่งชุมนุม ณ ถิ่นทุรกันดารซีนายว่า

8. บางคำถูกละเว้นไม่แปลเพราะแปลรวมไปกับคำอื่นแล้ว

9. ความถูกต้องในการตัดคำมีผลกับประสิทธิภาพในการจับคู่คำ ดังตัวอย่างแสดงในตารางที่ 5.18 การตัดคำผิดนอกจากจะทำให้จับคู่คำศัพท์ภาษาอังกฤษคำนั้นกับคำศัพท์ภาษาไทยที่คู่กันไม่ได้แล้ว บางครั้งยังมีผลกับความถี่และรูปร่างกราฟของคำอื่น ๆ ได้ เช่น จากตัวอย่าง คำว่ายา จะมีความถี่และรูปร่างของกราฟไม่ตรงกับที่ควรจะเป็น เป็นต้น

ตารางที่ 5.18 ตัวอย่างการตัดคำที่ผิด

ภาษาอังกฤษ	ภาษาไทย	ข้อความที่ได้จากการตัดคำ
Jacob	ยาโคบ	ยา-โค-บ
Benjamin	เบนยามิน	เบน-ยา-มิ-น
Hebrew	ฮีบรู	ฮีบ-รู

จากปัจจัยต่าง ๆ ที่กล่าวมา เป็นเหตุให้ความถูกต้องในการจับคู่คำในภาษาอังกฤษและภาษาไทยไม่มากนัก แต่อย่างไรก็ตามก็ยังเป็นประโยชน์อย่างมากกับการจับคู่ประโยคในหัวข้อ 5.3 และ 5.6

5.3 การทดลองจับคู่ประโยคแบบ 1:1

การทดลองจับคู่ประโยคแบบ 1 ประโยคภาษาอังกฤษกับ 1 ประโยคภาษาไทย ทดลองกับข้อมูลเข้าดังนี้ ไบเบิลขนาดสั้น ไบเบิลขนาดกลาง ไบเบิลขนาดยาว ตัวอย่างคู่ประโยคจาก

ดิกชันนารีขนาดสั้น ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง ข้อกฎหมายขนาดสั้น และข้อกฎหมายขนาดกลาง โดยทดลองกับหน้าต่างขนาด 1% 5% 10% 20% และ30% ของความยาวอนุกรมเวลา ทำการลดความยาวอนุกรมเวลาด้วยอัตราส่วน 0.5 ทดลองใช้ฟังก์ชันระยะห่างแมนฮัตตันและยุคลิเดียน โดยเลือกกำจัดกลุ่มคำหยุด 2 กลุ่ม คือ กลุ่มคำหยุดทั่วไปและกลุ่มคำหยุดกูเกิล ทำการทดลองให้คะแนนคู่ประโยค 3 แบบได้แก่

1. การให้คะแนนคู่ประโยคโดยใช้คู่คำอันดับที่ 1
2. การให้คะแนนคู่ประโยคโดยใช้คู่คำอันดับที่ 1 และ 2
3. การให้คะแนนคู่ประโยคโดยใช้คู่คำอันดับที่ 1 2 และ 3

การทดลองนี้เป็นการทดลองจับคู่ประโยคแบบ 1:1 ผู้วิจัยจึงเตรียมข้อมูลเข้าทั้งภาษาอังกฤษและภาษาไทยให้มีจำนวนประโยคที่เท่ากัน เช่น ไบเบิลขนาดสั้นจะมีประโยคภาษาอังกฤษและภาษาไทย 31 ประโยคเท่ากัน เป็นต้น ในปัจจุบันการนิยามคำว่าประโยคในภาษาไทยยังเป็นเรื่องที่ยากและไม่ชัดเจนเท่าใดนัก และเนื่องจากไบเบิล ตัวอย่างคู่ประโยคจากดิกชันนารี และข้อกฎหมายล้วนมีลักษณะข้อมูลแบบเป็นข้อ ๆ อยู่แล้ว ดังนั้นในงานวิจัยนี้จึงใช้ 1 ข้อเป็น 1 ประโยค เพื่อความสะดวกสำหรับการประเมินผลความถูกต้องในการจับคู่ประโยคนั่นเอง

ผลการทดลองจับคู่ประโยคโดยให้คะแนนคู่ประโยคทั้ง 3 แบบแสดงในภาคผนวก ข ส่วนตารางที่ 5.19 ถึง 5.25 เป็นค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคที่ได้จากการให้คะแนนทั้ง 3 แบบ

<u>การทดลองที่ 8</u> การทดลองจับคู่ประโยคกับไบเบิลขนาดสั้น
ข้อมูลเข้า: บทที่ 1 ปฐมกาล (Genesis)
จำนวนประโยค: 31 ประโยค

ตารางที่ 5.19 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคกับไบเบิลขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.86	0.69	0.83	0.66
5%	0.91	0.85	0.90	0.81
10%	0.86	0.81	0.76	0.72
20%	0.83	0.81	0.82	0.81
30%	0.70	0.71	0.68	0.70
ค่าเฉลี่ย	0.83	0.77	0.80	0.74

การทดลองที่ 9 การทดลองจับคู่ประโยคกับไบเบิลขนาดกลาง

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

จำนวนประโยค : 175 ประโยค

ตารางที่ 5.20 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคกับไบเบิลขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.21	0.15	0.25	0.16
5%	0.35	0.32	0.38	0.32
10%	0.35	0.32	**0.39**	0.34
20%	0.33	0.27	0.33	0.33
30%	0.32	0.31	0.31	0.35
ค่าเฉลี่ย	0.31	0.27	0.33	0.30

การทดลองที่ 10 การทดลองจับคู่ประโยคกับไบเบิลขนาดยาว

ข้อมูลเข้า : บทที่ 1 ถึง 25 ของปฐมกาล (Genesis)

จำนวนประโยค : 718 ประโยค

ตารางที่ 5.21 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคกับไบเบิลขนาดยาว

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.45	0.32	0.45	0.36
5%	0.52	0.52	0.52	0.51
10%	0.51	0.51	**0.55**	0.51
20%	0.47	0.47	0.52	0.53
30%	0.39	0.36	0.42	0.40
ค่าเฉลี่ย	0.47	0.44	0.49	0.46

การทดลองที่ 11 การทดลองจับคู่ประโยคกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 70 ตัวอย่าง

จำนวนประโยค : 70 ประโยค

ตารางที่ 5.22 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.35	0.37	0.30	0.36
5%	0.49	0.40	0.43	0.35
10%	0.52	0.50	0.43	0.42
20%	0.29	0.33	0.27	0.32
30%	0.30	0.30	0.30	0.28
ค่าเฉลี่ย	0.39	0.38	0.35	0.35

การทดลองที่ 12 การทดลองจับคู่ประโยคกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 304 ตัวอย่าง

จำนวนประโยค : 304 ประโยค

ตารางที่ 5.23 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.36	0.26	0.27	0.22
5%	0.44	0.39	0.36	0.30
10%	0.36	0.37	0.31	0.33
20%	0.20	0.17	0.21	0.20
30%	0.11	0.10	0.11	0.10
ค่าเฉลี่ย	0.29	0.26	0.25	0.23

การทดลองที่ 13 การทดลองจับคู่ประโยคกับข้อกฎหมายขนาดสั้น

ข้อมูลเข้า : ข้อกฎหมาย 50 ข้อ

จำนวนประโยค : 50 ประโยค

ตารางที่ 5.24 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคกับข้อกฎหมายขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.45	0.37	0.45	0.39
5%	0.51	0.46	0.53	0.48
10%	0.50	0.47	0.53	0.47
20%	0.51	0.51	0.52	0.49
30%	0.56	0.52	**0.59**	0.52
ค่าเฉลี่ย	0.51	0.47	0.52	0.47

การทดลองที่ 14 การทดลองจับคู่ประโยคกับข้อกฎหมายขนาดกลาง

ข้อมูลเข้า : ข้อกฎหมาย 102 ข้อ

จำนวนประโยค : 102 ประโยค

ตารางที่ 5.25 ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคกับข้อกฎหมายขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล		กำจัดกลุ่มคำหยุดทั่วไป	
ขนาดของหน้าต่าง	แมนฮัตตัน	ยุคลิเดียน	แมนฮัตตัน	ยุคลิเดียน
1%	0.45	0.43	0.48	0.47
5%	0.50	0.46	**0.52**	0.44
10%	**0.52**	0.47	0.46	0.43
20%	0.46	0.43	0.49	0.47
30%	0.40	0.44	0.44	0.46
ค่าเฉลี่ย	0.47	0.45	0.48	0.45

5.4 วิเคราะห์ผลการทดลองจับคู่ประโยคแบบ 1:1

จากผลการทดลองในตารางที่ 5.19 ถึง 5.25 พบว่า ในการทดลองที่ 8 ซึ่งใช้ไบเบิลขนาดสั้นเป็นข้อมูลเข้า ให้ค่าเฉลี่ยความถูกต้องในการจับคู่ประโยคมากที่สุด โดยเฉพาะที่หน้าต่างขนาด 5% ของความยาวอนุกรมเวลา ใช้ฟังก์ชันระยะห่างแบบแมนฮัตตัน และกำจัดกลุ่มคำหยุดภูเกิล ให้ผลดีถึง 91% ที่เป็นเช่นนี้เนื่องจาก หน้าต่างขนาด 5% ของความยาวอนุกรมเวลา และการใช้ฟังก์ชันระยะห่างแบบแมนฮัตตัน เป็นค่าพารามิเตอร์ที่ทำให้การจับคู่คำมีความถูกต้องสูงกว่าค่าอื่น ๆ (ดังที่กล่าวไปแล้วในหัวข้อ 4.3.1 และ 4.3.3) ดังนั้นเมื่อการจับคู่คำมีความถูกต้องมากก็ย่อมส่งผลให้การจับคู่ประโยคดีตามไปด้วย

ความถูกต้องในการจับคู่ประโยค ส่วนใหญ่เป็นไปในทิศทางเดียวกับความถูกต้องในการจับคู่คำ คือ ถ้าจับคู่คำได้มาก ก็จะทำให้การจับคู่ประโยคได้ผลดี นั่นคือ

1. ถ้าข้อมูลเป็นประเภทเดียวกันเช่น ประเภทตัวอย่างคู่ประโยคจากดิกชันนารีเหมือนกัน หรือประเภทข้อกฎหมายเหมือนกัน ข้อมูลเข้าที่มีขนาดสั้นมักจะให้ผลดีกว่าข้อมูลเข้าที่มีขนาดยาวกว่า ตัวอย่างเช่น การทดลองที่ 11 ให้ผลดีกว่าการทดลองที่ 12 และ การทดลองที่ 13 ให้ผลดีกว่าการทดลองที่ 14 เป็นต้น

แต่อย่างไรก็ตาม ข้อมูลเข้าที่ยาวสามารถที่จะให้ผลการจับคู่ประโยคที่ดีกว่าข้อมูลเข้าประเภทเดียวกันแต่มีความยาวสั้นกว่าได้ ดังการทดลองที่ 9 และ 10 ซึ่งการทดลองที่ 10 ใช้ข้อมูลไบเบิลที่ยาวกว่าการทดลองที่ 9 แต่ให้ผลการจับคู่ประโยคที่ดีกว่า โดยที่เปอร์เซ็นต์ความถูกต้องของการจับคู่คำไม่ต่างกันเท่าใดนัก (ดูเปอร์เซ็นต์ความถูกต้องในการจับคู่คำได้จากการทดลองที่ 2 และ 3) ที่เป็นเช่นนี้เนื่องจากความถูกต้องในการจับคู่คำนับจากทั้งกรณีที่จับคู่ได้กับคำที่ถูกต้องและจับคู่ได้กับคำที่อยู่ในประโยคเดียวกับคำที่ถูกต้อง โดยกรณีที่จับคู่ได้กับคำที่ถูกต้องนั้นส่วนใหญ่มาจากการที่คำศัพท์ภาษาอังกฤษและคู่ของคำศัพท์นั้นมีความถี่ในการเกิดมาก และมีรูปร่างที่คล้ายกันอย่างเห็นได้ชัด ซึ่งข้อมูลที่ใช้ในการทดลองที่ 10 มีความยาวมากพอที่จะทำให้กราฟมีรูปร่างคล้ายกันอย่างเด่นชัด ในขณะที่ข้อมูลที่ใช้ในการทดลองที่ 9 จะจับคู่ได้ถูกต้องเนื่องจากกรณีจับคู่ได้กับคำที่อยู่ในประโยคเดียวกับคำที่ถูกต้องเป็นส่วนใหญ่

ดังนั้นการทดลองที่ 10 จึงให้ความถูกต้องในการจับคู่ประโยคดีกว่าการทดลองที่ 9 เพราะความถูกต้องในการจับคู่คำที่ได้จากกรณีจับคู่ได้กับคำที่ถูกต้อง จะมีความแม่นยำมากกว่ากรณีจับคู่ได้กับคำที่อยู่ในประโยคเดียวกับคำที่ถูกต้อง

2. ข้อมูลที่ใช้คำศัพท์หลากหลาย มีคำศัพท์ที่แตกต่างกันเป็นจำนวนมาก จะให้ผลด้อยกว่าข้อมูลที่มีคำศัพท์ที่แตกต่างกันจำนวนน้อยกว่า เช่น ในการทดลองที่ 12 และ 14 ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง มีความยาวของข้อความมากกว่าข้อกฎหมายขนาดกลางไม่

มากนัก แต่มีการใช้คำศัพท์ที่แตกต่างกันในภาษาอังกฤษเป็นจำนวนมากกว่าถึงหนึ่งเท่า คือ ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลางใช้ 1,096 คำ ในขณะที่ข้อกฎหมายขนาดกลางใช้เพียง 550 คำ ทำให้การใช้ข้อกฎหมายขนาดกลางเป็นข้อมูลเข้าจึงได้ผลดีกว่า

นอกจากนี้ ปัจจัยอีกข้อหนึ่งที่มีผลต่อความถูกต้องในการจับคู่ประโยคได้แก่ ความยาวของประโยคภาษาอังกฤษ เนื่องจากถ้าประโยคมีความยาวมาก จะมีคำที่นำไปคิดคะแนนมากกว่าประโยคภาษาไทยที่เป็นคู่ก็จะมีโอกาสได้คะแนนมากและคะแนนจะต่างจากประโยคที่ผิดอย่างชัดเจน แต่ถ้าประโยคภาษาอังกฤษมีขนาดสั้น มีคำที่นำไปคิดคะแนนเพียงไม่กี่คำ อาจมีประโยคภาษาไทยหลายประโยคที่มีคะแนนใกล้เคียงกัน และทำให้จับคู่ประโยคผิดได้

5.5 การเปรียบเทียบกับวิธีอื่นที่ไม่ใช้อนุกรมเวลา

ในปัจจุบันการจับคู่คำสามารถทำได้หลายวิธี เช่น การเปิดพจนานุกรม และการใช้สถิติเป็นต้น วิธีการใช้สถิตินั้นได้มีผู้พัฒนาโปรแกรมที่เป็นโอเพนซอส (Open Source) ขึ้นมา โดยใช้ชื่อว่า กิซาพลัสพลัส (GIZA++) ซึ่งสามารถดาวน์โหลดได้จากอินเทอร์เน็ต [32, 33] กิซาพลัสพลัสเป็นโปรแกรมที่พัฒนาต่อมาจากโปรแกรมที่ชื่อว่ากิซา (GIZA) ซึ่งเป็นส่วนหนึ่งของชุดเครื่องมือการแปลภาษาด้วยเครื่องโดยวิธีทางสถิติ (Statistical Machine Translation Toolkit) กิซาพลัสพลัสใช้ในการสร้างโมเดลการแปลจากข้อความขนาน และทำการจับคู่คำโดยใช้อีเอ็มอัลกอริทึม (Expectation Maximization : EM) สามารถใช้กับคู่ภาษาใด ๆ ก็ได้

ในหัวข้อนี้ผู้วิจัยได้ใช้ข้อมูลหลายประเภทและหลายขนาดมาทดลองจับคู่คำด้วยโปรแกรมกิซาพลัสพลัส ได้ผลการทดลองดังตารางที่ 5.26 และ 5.27

ตารางที่ 5.26 ผลการทดลองจับคู่คำด้วยโปรแกรมกิซ่าพลัสพลัสแบบกำจัดกลุ่มคำหยุดกูเกิล

ข้อมูลเข้า	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้องจากกิซาพลัสพลัส	% ความถูกต้องจากวิธีการที่นำเสนอ
ไบเบิลขนาดสั้น	62	7	1	0.60	0.76
ไบเบิลขนาดกลาง	210	52	9	0.54	0.35
ไบเบิลขนาดยาว	443	181	45	0.61	0.29
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น	137	9	1	0.36	0.47
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง	487	45	3	0.48	0.27
ข้อกฎหมายขนาดสั้น	170	5	0	0.58	0.60
ข้อกฎหมายขนาดกลาง	293	33	3	0.59	0.47

ตารางที่ 5.27 ผลการทดลองจับคู่คำด้วยโปรแกรมกิซาพลัสพลัสแบบกำจัดกลุ่มคำหยุดทั่วไป

ข้อมูลเข้า	อันดับที่ 1	อันดับที่ 2	อันดับที่ 3	% ความถูกต้องจากกิซาพลัสพลัส	% ความถูกต้องจากวิธีการที่นำเสนอ
ไบเบิลขนาดสั้น	42	7	0	0.66	0.86
ไบเบิลขนาดกลาง	176	35	6	0.58	0.36
ไบเบิลขนาดยาว	391	153	40	0.62	0.30
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น	106	6	0	0.36	0.48
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง	425	40	1	0.49	0.27
ข้อกฎหมายขนาดสั้น	153	5	0	0.59	0.59
ข้อกฎหมายขนาดกลาง	255	29	3	0.60	0.49

จากผลการทดลองพบว่า สำหรับข้อมูลขนาดสั้นวิธีการที่นำเสนอจะให้ผลการจับคู่คำที่ดีกว่ากิซาพลัสพลัส แต่สำหรับข้อมูลขนาดกลางและยาวกิซาพลัสพลัสจะให้ผลการจับคู่คำที่ดีกว่า เนื่องจากกิซาพลัสพลัสเป็นการใช้วิธีทางสถิติ จึงให้ผลดีเมื่อข้อมูลเข้ามีขนาดใหญ่

เมื่อนำคู่คำที่ได้จากการจับคู่คำด้วยโปรแกรมกิซาพลัสพลัสไปจับคู่ประโยค ได้ผลการทดลองดังตารางที่ 5.28 และ 5.29

ตารางที่ 5.28 ผลการทดลองจับคู่ประโยคด้วยคู่คำที่ได้จากโปรแกรมกิซาพลัสพลัสแบบกำจัดกลุ่มคำหยุดกูเกิล

ข้อมูลเข้า	จำนวนประโยคที่จับคู่ได้ถูกต้อง	% ความถูกต้องจากกิซาพลัสพลัส	% ความถูกต้องจากวิธีการที่นำเสนอ
ไบเบิลขนาดสั้น	29	0.94	0.91
ไบเบิลขนาดกลาง	123	0.70	0.35
ไบเบิลขนาดยาว	549	0.76	0.52
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น	46	0.66	0.52
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง	265	0.87	0.44
ข้อกฎหมายขนาดสั้น	37	0.74	0.56
ข้อกฎหมายขนาดกลาง	83	0.81	0.52

ตารางที่ 5.29 ผลการทดลองจับคู่ประโยคด้วยคู่คำที่ได้จากโปรแกรมกิซาพลัสพลัสแบบกำจัดกลุ่มคำหยุดทั่วไป

ข้อมูลเข้า	จำนวนประโยคที่จับคู่ได้ถูกต้อง	% ความถูกต้องจากกิซาพลัสพลัส	% ความถูกต้องจากวิธีการที่นำเสนอ
ไบเบิลขนาดสั้น	28	0.90	0.90
ไบเบิลขนาดกลาง	126	0.72	0.39
ไบเบิลขนาดยาว	572	0.80	0.55
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น	39	0.56	0.43
ตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง	255	0.84	0.36
ข้อกฎหมายขนาดสั้น	43	0.86	0.59
ข้อกฎหมายขนาดกลาง	83	0.81	0.52

จากผลการทดลองจับคู่ประโยคพบว่า ส่วนใหญ่การใช้คู่คำจากกิซาพลัสพลัสจะให้ผลการทดลองที่ดีกว่าวิธีการที่นำเสนอ เนื่องจากในการจับคู่คำ คิดความถูกต้องโดยนับจากคำที่อยู่ในประโยคที่ถูกต้อง ดังนั้นคำที่จับคู่ได้จึงอาจเป็นคำที่มีความหมายคู่กันหรือคำที่มีตำแหน่งใกล้เคียงก็ได้ ซึ่งการจับคู่ประโยคโดยใช้คำที่มีความหมายคู่กันจะให้ผลดีกว่าการใช้คำที่มีตำแหน่งใกล้เคียง และคู่คำที่ได้จากกิซาพลัสพลัสเป็นกรณีคำที่มีความหมายคู่กันมากกว่าตำแหน่งใกล้เคียงจึงให้ผลการจับคู่ประโยคดีกว่า แต่สำหรับไบเบิลขนาดสั้นการจับคู่ประโยคได้ผลดีเท่ากัน คือ 90 เปอร์เซ็นต์ ซึ่งถือเป็นแนวทางที่ดี สำหรับการพัฒนาต่อไปในอนาคต เนื่องจากวิธีการที่นำเสนอนี้ยังมิได้มีการวิเคราะห์เกี่ยวกับโครงสร้างของประโยค หรือใช้ความรู้เกี่ยวกับภาษาศาสตร์มากนัก ดังนั้นถ้าในอนาคตมีการนำความรู้ทางด้านภาษาศาสตร์เข้ามาช่วย ย่อมทำให้ผลการจับคู่ประโยคดีขึ้นได้อีกอย่างแน่นอน

อย่างไรก็ตาม ข้อจำกัดอย่างหนึ่งของกิซาพลัสพลัสคือ การที่จะจับคู่คำด้วยวิธีนี้ ข้อมูลเข้าที่ใช้จะต้องทราบขอบเขตหรือคู่ประโยค แต่สำหรับวิธีที่นำเสนอ ไม่จำเป็นต้องทราบขอบเขตหรือคู่ประโยค และข้อดีอีกอย่างของวิธีที่นำเสนอคือ ไม่ยุ่งยากซับซ้อนแต่ก็สามารถให้ผลการจับคู่ประโยคที่ดีได้สำหรับไบเบิลขนาดสั้น

5.6 การทดลองจับคู่ประโยคแบบ 1:N

การทดลองจับคู่ประโยคแบบ 1 ประโยคภาษาอังกฤษกับ N วรรคภาษาไทย ทำการทดลองโดยใช้ข้อมูลสองประเภท ความยาวหลายขนาด เนื่องจากข้อมูลเข้าที่เป็นตัวอย่างคู่

ประโยคจากดิกชันนารีนั้นเป็นข้อมูลที่ได้รับมาจากหน่วยงานวิจัยอื่น ซึ่งมีการเตรียมข้อมูลโดยการตัดคำมาเรียบร้อยแล้ว จึงไม่ทราบตำแหน่งของการสิ้นสุดวรรค ในการทดลองแบบ 1:N นี้จึงไม่ได้ทำการทดลองกับตัวอย่างคู่ประโยคจากดิกชันนารี จากการทดลองจับคู่ประโยคแบบ 1:1 ส่วนใหญ่การใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันจะให้ผลที่ดีกว่าแบบยุคลิเดียน ดังนั้นในการทดลองจับคู่ประโยคแบบ 1: N จึงเลือกใช้ฟังก์ชันระยะห่างแมนฮัตตัน โดยทดลอง 3 แบบได้แก่

1. การให้คะแนนคู่ประโยคโดยใช้คู่คำอันดับที่ 1
2. การให้คะแนนคู่ประโยคโดยใช้คู่คำอันดับที่ 1 และ 2
3. การให้คะแนนคู่ประโยคโดยใช้คู่คำอันดับที่ 1 2 และ 3

โดยในการทดลองจะให้ 1 คะแนนกับประโยคที่สามารถจับคู่ได้ถูกต้อง และประโยคที่จับคู่ได้เนื้อหาครบถ้วน แม้จะมีวรรคเกินมาบ้าง ส่วนประโยคที่จับคู่ได้เพียงบางส่วน หรือมีบางวรรคขาดหายไปจะให้คะแนน 0.5 การหาค่าเฉลี่ยนั้นจะได้จากการนำคะแนนทั้งกรณีที่จับคู่ได้เนื้อหาครบและกรณีจับคู่ได้เพียงบางส่วนมาบวกกันและหารด้วยจำนวนประโยคทั้งหมด ซึ่งผลการทดลองแสดงดังตารางที่ 5.30 ถึง 5.39

<u>การทดลองที่ 15</u> การทดลองจับคู่ประโยคแบบ 1:N กับไบเบิลขนาดสั้น
ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)
จำนวนประโยคภาษาอังกฤษ : 31 ประโยค
จำนวนวรรคภาษาไทย : 108 วรรค

ตารางที่ 5.30 ผลการจับคู่ประโยคแบบ1:N กับไบเบิลขนาดสั้น โดยกำจัดกลุ่มคำหยุดกูเกิล

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	7	1	0.26	8	2.5	0.34	6	3	0.29
5%	5	4.5	**0.31**	4	5.5	0.31	5	5.5	0.34
10%	4	5.5	**0.31**	4	7	**0.35**	3	8	**0.35**
20%	4	5.5	**0.31**	4	7	**0.35**	3	6.5	0.31
30%	4	4	0.26	4	5	0.29	5	5.5	0.34
	ค่าเฉลี่ย		0.29	ค่าเฉลี่ย		0.33	ค่าเฉลี่ย		0.33

ตารางที่ 5.31 ผลการจับคู่ประโยคแบบ1:N กับไบเบิลขนาดสั้น โดยกำจัดกลุ่มคำหยุดทั่วไป

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	5	1	0.19	5	1.5	0.21	6	2.5	0.27
5%	4	3.5	0.24	4	4	0.26	5	4	0.29
10%	4	4.5	0.27	4	6	0.32	4	6.5	**0.34**
20%	4	6.5	**0.34**	4	7	**0.35**	2	6	0.26
30%	4	4	0.26	4	4.5	0.27	4	4	0.26
	ค่าเฉลี่ย		0.26	ค่าเฉลี่ย		0.28	ค่าเฉลี่ย		0.28

<u>การทดลองที่ 16</u> การทดลองจับคู่ประโยคแบบ 1:N กับไบเบิลขนาดกลาง

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

จำนวนประโยคภาษาอังกฤษ : 175 ประโยค

จำนวนวรรคภาษาไทย : 547 วรรค

ตารางที่ 5.32 ผลการจับคู่ประโยคแบบ 1:N กับไบเบิลขนาดกลาง โดยกำจัดกลุ่มคำหยุดกูเกิล

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	8	3	0.06	9	3	0.07	11	3.5	0.08
5%	13	4	0.10	13	3.5	0.09	14	4	0.10
10%	13	3.5	0.09	14	2	0.09	17	3	0.11
20%	14	3.5	0.10	14	1.5	0.09	13	2.5	0.09
30%	16	3.5	**0.11**	16	5	**0.12**	17	4	**0.12**
	ค่าเฉลี่ย		0.09	ค่าเฉลี่ย		0.09	ค่าเฉลี่ย		0.10

ตารางที่ 5.33 ผลการจับคู่ประโยคแบบ1:N กับไบเบิลขนาดกลาง โดยกำจัดกลุ่มคำหยุดทั่วไป

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	8	11	0.11	7	10	0.10	7	9	0.09
5%	9	15.5	0.14	12	15.5	**0.16**	10	17.5	**0.16**
10%	11	17	**0.16**	12	14.5	0.15	12	16	**0.16**

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
20%	11	13	0.14	12	6	0.10	12	8.5	0.12
30%	13	10.5	0.13	10	12	0.13	11	10.5	0.12
	ค่าเฉลี่ย		0.14	ค่าเฉลี่ย		0.13	ค่าเฉลี่ย		0.13

<u>การทดลองที่ 17</u> การทดลองจับคู่ประโยคแบบ 1:N กับไบเบิลขนาดยาว

ข้อมูลเข้า : บทที่ 1 ถึง 25 ของปฐมกาล (Genesis)

จำนวนประโยคภาษาอังกฤษ : 718 ประโยค

จำนวนวรรคภาษาไทย : 2,306 วรรค

ตารางที่ 5.34 ผลการจับคู่ประโยคแบบ1:N กับไบเบิลขนาดยาว โดยกำจัดกลุ่มคำหยุดกูเกิล

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	71	15.5	0.12	85	18	0.14	95	17	0.16
5%	57	72	**0.18**	82	68.5	**0.21**	88	74.5	**0.23**
10%	107	14.5	0.17	109	19	0.18	101	20.5	0.17
20%	57	50.5	0.15	79	51	0.18	85	55	0.19
30%	65	17.5	0.10	85	14	0.14	86	15.5	0.14
	ค่าเฉลี่ย		0.14	ค่าเฉลี่ย		0.17	ค่าเฉลี่ย		0.18

ตารางที่ 5.35 ผลการทดลองจับคู่ประโยคแบบ1:N กับไบเบิลขนาดยาว โดยกำจัดกลุ่มคำหยุดทั่วไป

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	49	73	0.17	64	62.5	0.18	66	69.5	0.19
5%	59	51	0.15	87	46	0.19	84	50	0.19
10%	71	72.5	**0.20**	74	68	**0.20**	75	70	**0.20**
20%	46	72	0.16	65	63	0.18	74	64	0.19
30%	51	41.5	0.13	74	35.5	0.15	72	33.5	0.15
	ค่าเฉลี่ย		0.16	ค่าเฉลี่ย		0.18	ค่าเฉลี่ย		0.18

<u>การทดลองที่ 18</u> การทดลองจับคู่ประโยคแบบ 1:N กับข้อกฎหมายขนาดสั้น

ข้อมูลเข้า : ข้อกฎหมาย 50 ข้อ

จำนวนประโยคภาษาอังกฤษ : 50 ประโยค

จำนวนวรรคภาษาไทย : 153 วรรค

ตารางที่ 5.36 ผลการจับคู่ประโยคแบบ1:N ข้อกฎหมายขนาดสั้น โดยกำจัดกลุ่มคำหยุดกูเกิล

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	8	1	0.18	7	0.5	0.15	7	1.5	0.17
5%	12	5.5	0.35	12	4	0.32	12	5	0.34
10%	7	6.5	0.27	7	7	0.28	10	8.5	**0.37**
20%	11	7.5	**0.37**	10	9	**0.38**	10	7	0.34
30%	10	6	0.32	10	6.5	0.33	9	8	0.34
	ค่าเฉลี่ย		0.30	ค่าเฉลี่ย		0.29	ค่าเฉลี่ย		0.31

ตารางที่ 5.37 ผลการจับคู่ประโยคแบบ1:N กับข้อกฎหมายขนาดสั้น โดยกำจัดกลุ่มคำหยุดทั่วไป

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	8	1.5	0.19	7	1	0.16	6	1.5	0.15
5%	9	5	0.28	12	4	0.32	12	4	**0.32**
10%	7	7	0.28	7	6.5	0.27	8	8	**0.32**
20%	12	4.5	0.33	11	5.5	0.33	8	7	0.30
30%	10	7.5	**0.35**	10	7	**0.34**	9	7	**0.32**
	ค่าเฉลี่ย		0.29	ค่าเฉลี่ย		0.28	ค่าเฉลี่ย		0.28

<u>การทดลองที่ 19</u> การทดลองจับคู่ประโยคแบบ 1:N กับข้อกฎหมายขนาดกลาง

ข้อมูลเข้า : ข้อกฎหมาย 102 ข้อ

จำนวนประโยคภาษาอังกฤษ : 102 ประโยค

จำนวนวรรคภาษาไทย : 387 วรรค

ตารางที่ 5.38 ผลการจับคู่ประโยคแบบ1:N ข้อกฎหมายขนาดกลาง โดยกำจัดกลุ่มคำหยุดกูเกิล

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	15	1	0.16	17	0.5	0.17	14	1.5	0.15
5%	16	9.5	**0.25**	11	8.5	0.19	12	10.5	0.22
10%	9	13.5	0.22	13	13.5	**0.26**	12	15.5	**0.27**
20%	16	0.5	0.16	10	2	0.12	11	2.5	0.13
30%	8	13	0.21	9	14	0.23	9	14.5	0.23
	ค่าเฉลี่ย		0.20	ค่าเฉลี่ย		0.19	ค่าเฉลี่ย		0.20

ตารางที่ 5.39 ผลการจับคู่ประโยคแบบ1:N กับข้อกฎหมายขนาดกลาง โดยกำจัดกลุ่มคำหยุดทั่วไป

	ใช้คู่คำอันดับที่ 1			ใช้คู่คำอันดับที่ 1, 2			ใช้คู่คำอันดับที่ 1, 2, 3		
ขนาดหน้าต่าง	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย	ครบ	ขาด	เฉลี่ย
1%	17	1	0.18	16	0.5	0.16	15	1	0.16
5%	16	7	0.23	14	8	0.22	13	6.5	0.19
10%	8	11.5	0.19	10	9	0.19	10	10	0.20
20%	12	15	**0.26**	12	14	**0.25**	10	16	**0.25**
30%	8	14.5	0.22	11	13	0.24	10	15	**0.25**
	ค่าเฉลี่ย		0.22	ค่าเฉลี่ย		0.21	ค่าเฉลี่ย		0.21

5.7 วิเคราะห์ผลการทดลองจับคู่ประโยคแบบ 1:N

จากผลการทดลองในตารางที่ 5.30 ถึง 5.39 พบว่าไบเบิลขนาดสั้นให้ผลการจับคู่ประโยคแบบ 1:N ดีกว่าข้อมูลเข้าชนิดอื่น ๆ รองลงมาได้แก่ กฎหมายขนาดสั้น และกฎหมายขนาดกลาง ส่วนการใช้ ไบเบิลขนาดยาวก็ให้ผลดีกว่าไบเบิลขนาดกลาง ที่ผลเป็นเช่นนี้เนื่องมาจากเหตุผลเดียวกับที่ได้กล่าวถึงไปแล้วในการวิเคราะห์ผลการจับคู่ประโยคแบบ 1:1 คือ ถ้าจับคู่คำได้มากผลการจับคู่ประโยคก็จะดีตามไปด้วย

แต่จากผลการทดลองจะพบว่าเปอร์เซ็นต์ความถูกต้องในการจับคู่ประโยคแบบ 1:1 จะให้ผลดีกว่าการจับคู่แบบ 1:N มาก เนื่องจากไม่ต้องใช้พารามิเตอร์อัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและคำเนื้อหาในภาษาไทย เพราะพารามิเตอร์ตัวนี้มีผลทำให้จับคู่ผิดได้มาก เพราะ ประโยคภาษาอังกฤษและภาษาไทยมีจำนวนคำเนื้อหาไม่เท่ากัน ขึ้นอยู่กับผู้แปลด้วยว่าจะใช้คำเนื้อหาเท่ากับประโยคในอีกภาษาหนึ่งหรือไม่ ซึ่งตามปกติไม่มีการนับ ดังนั้นค่าอัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและคำเนื้อหาในภาษาไทย ที่ใช้ในการทดลองนี้เป็นค่าที่หาได้จากการทดลองในหัวข้อ 4.3.6 ซึ่งเป็นเพียงค่าเฉลี่ยเท่านั้น การกำหนดค่าอัตราส่วนไว้น้อยจะทำให้ได้จำนวนวรรคน้อย ดังนั้นประโยคที่ได้อาจจะมีวรรคที่ขาดไป ทั้ง ๆ ที่ถ้ากำหนดอัตราส่วนให้มากขึ้น วรรคที่ถูกต้องจะไม่ขาดไป ในทางตรงกันข้าม ถ้ากำหนดอัตราส่วนไว้มากเกินไปจะทำให้ได้จำนวนวรรคมาก ดังนั้นจึงอาจจะมีวรรคที่ไม่ถูกต้องเกินมาได้ แต่อย่างไรก็ตามยังได้เนื้อหาที่ครบประโยค

นอกจากการหาจำนวนวรรคในประโยคภาษาไทย จะเป็นส่วนหนึ่งที่ทำให้การจับคู่แบบ 1 ประโยคภาษาอังกฤษต่อ N วรรคในภาษาไทยเป็นเรื่องที่ยากแล้ว ยังมีปัจจัยอื่นอีกที่มีผลต่อความถูกต้องในการจับคู่ประโยค ได้แก่ การกำจัดวรรคที่ไม่ถูกต้อง วรรคต่าง ๆ ที่มีคะแนนสูงสุด 5 อันดับแรกที่เก็บไว้เพื่อจะทำการจับคู่กับประโยคภาษาอังกฤษ ย่อมมีทั้งวรรคที่ถูกต้องและวรรคที่ไม่ถูกต้อง ดังนั้นจึงต้องมีการกำจัดวรรคที่คาดว่าจะไม่ถูกต้องออกไป ซึ่งการกำจัดผิดเพียงหนึ่งวรรคต่อประโยค ก็จะมีผลทำให้ความถูกต้องในการจับคู่ประโยคลดลง การกำจัดวรรคที่อยู่ห่างจากวรรคอื่นมาก ๆ สามารถช่วยกำจัดวรรคที่ไม่ถูกต้องได้ และการรวมเอาวรรคที่อยู่ระหว่างวรรคอื่นที่จับคู่ได้ก็ช่วยให้ได้ประโยคที่ครบถ้วนมากยิ่งขึ้น

การจับคู่ประโยคแบบ 1:N เป็นการศึกษาต่อเนื่องจากการจับคู่ประโยคแบบ 1:1 เนื่องจากในปัจจุบันยังไม่มีงานวิจัยที่ทำการจับคู่ 1 ประโยคภาษาอังกฤษกับ N วรรคในภาษาไทย และแม้ว่าผลการทดลองจับคู่แบบ 1:N อาจจะไม่ดีเท่าใดนัก แต่ผู้วิจัยคาดว่าการศึกษานี้น่าจะเป็นประโยชน์สำหรับงานวิจัยอื่น ๆ ในอนาคตได้ไม่มากก็น้อย

บทที่ 6

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้เป็นการศึกษาเกี่ยวกับการจับคู่คำและการจับคู่ประโยคในคลังข้อความขนานภาษาอังกฤษและภาษาไทย โดยใช้อนุกรมเวลา เนื่องจากในคลังข้อความขนาน คำศัพท์ในภาษาอังกฤษและภาษาไทยแต่ละคำจะมีความถี่และตำแหน่งในการปรากฏต่างกัน ทำให้สามารถสร้างอนุกรมเวลาของคำศัพท์แต่ละคำได้โดยที่รูปร่างไม่ซ้ำกัน แต่คำศัพท์ที่เป็นคู่กันในภาษาอังกฤษและภาษาไทยมีแนวโน้มที่รูปร่างของอนุกรมเวลาจะคล้ายกัน ดังนั้นผู้เขียนจึงนำรูปร่างของอนุกรมเวลามาใช้ในการจับคู่คำในคลังข้อความขนาน และนำคำที่จับคู่ได้ไปใช้ในการจับคู่ประโยคต่อไป

6.1 สรุปผลการวิจัย

งานวิจัยนี้ได้ทดลองจับคู่คำและจับคู่ประโยคในคลังข้อความขนานภาษาอังกฤษและภาษาไทยโดยใช้อนุกรมเวลา ซึ่งทำการทดลองกับคลังข้อความขนาน 3 ประเภทได้แก่ คัมร์ภีไบเบิล ตัวอย่างคู่ประโยคจากดิกชันนารี และข้อกฎหมาย โดยข้อความแต่ละประเภทก็จะใช้ความยาวแตกต่างกันมีทั้งสั้น กลาง และยาว พร้อมทั้งปรับพารามิเตอร์ต่าง ๆ ที่คาดว่าจะมีผลต่อความถูกต้องในการจับคู่คำและจับคู่ประโยค โดยพารามิเตอร์ต่าง ๆ มีดังต่อไปนี้

1. ขนาดของหน้าต่าง จากการทดลองพบว่า ขนาดหน้าต่างที่มีความเหมาะสมจะอยู่ในช่วง 5% - 10% ของความยาวอนุกรมเวลา

2. อัตราส่วนการลดความยาวของอนุกรมเวลา อัตราส่วน 0.5 เป็นค่าที่เหมาะสมสำหรับการลดความยาวของอนุกรมเวลาเพราะนอกจากจะช่วยลดเวลาในการคำนวณฟังก์ชันระยะห่างแล้ว ยังช่วยให้ผลการจับคู่คำดีขึ้นอีกด้วย

3. วิธีที่ใช้ในการวัดความเหมือนของอนุกรมเวลา การวัดโดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันจะให้ผลที่ดีกว่าแบบยุคลิเดียนเล็กน้อย ส่วนแบบไดนามิกไทม์วอร์ปปิงจะให้ผลด้อยกว่าแบบแมนฮัตตันและยังใช้เวลาในการคำนวณนานกว่ามาก จึงควรเลือกใช้แบบแมนฮัตตันจะเหมาะสมที่สุด

4. จำนวนของคำหยุด คำหยุดในภาษาอังกฤษมีอยู่เป็นจำนวนมาก ดังนั้นการกำจัดคำหยุดทุกคำ บางครั้งก็ทำให้ได้ผลด้อยกว่าการกำจัดเฉพาะบางคำ โดยถ้าเป็นข้อมูลขนาดสั้นควรกำจัดกลุ่มคำหยุดกูเกิล ส่วนข้อมูลขนาดกลางและยาวควรกำจัดกลุ่มคำหยุดทั่วไป

5. อันดับของคู่คำที่ใช้ในการให้คะแนนคู่ประโยค สำหรับข้อมูลเข้าที่มีการใช้คำศัพท์ที่แตกต่างกันมาก หรือมีอัตราส่วนระหว่างจำนวนคำที่แตกต่างกันต่อจำนวนคำทั้งหมดมาก การใช้ทั้ง 3 อันดับจะให้ผลดีกว่าการใช้อันดับเดียว

6. อัตราส่วนระหว่างจำนวนคำเนื้อหาในภาษาอังกฤษและคำเนื้อหาในภาษาไทย ค่า 0.8 เป็นอัตราส่วนที่เหมาะสมค่าหนึ่ง แต่อย่างไรก็ตาม การปรับค่าอัตราส่วนให้มากกว่าหรือน้อยกว่า 0.8 ก็อาจจะส่งผลดีกับเปอร์เซ็นต์ความถูกต้องได้เช่นกัน

จากการวิจัยเรื่องการจับคู่ประโยค สรุปได้ว่าผลในการจับคู่ประโยคทั้งแบบ 1:1 และ 1:N นั้นแปรผันตามความถูกต้องในการจับคู่คำเป็นหลัก ถ้าจับคู่คำได้มากก็จะจับคู่ประโยคได้มากตามไปด้วย การจับคู่คำในคลังข้อความขนานโดยใช้อนุกรมเวลา จะให้ผลดีสำหรับข้อความขนาดสั้นประมาณ 1 หน้า โดยการจับคู่แบบ 1:1 จะให้ผลดีกว่าการจับคู่แบบ 1:N ทั้งนี้เปอร์เซ็นต์ความถูกต้องในการจับคู่ประโยคจะดีเพียงใดต้องขึ้นอยู่กับปัจจัยอื่น ๆ ด้วยได้แก่ ประเภทของข้อมูล ความยาวของข้อมูล ความหลากหลายในการใช้คำศัพท์ กลุ่มของคำหยุดที่เลือกใช้ และสำหรับการจับคู่แบบ 1:N จะขึ้นอยู่กับวิธีการหาจำนวนวรรคของประโยคภาษาไทย และการกำจัดวรรคที่ไม่ถูกต้องด้วย ซึ่งทำให้การจับคู่ประโยคแบบ 1:N เป็นเรื่องที่ยาก และแม้ว่าผลการจับคู่ประโยคแบบ 1:N จะไม่มากเท่าใดนัก แต่ถ้าในอนาคตมีการนำอัลกอริทึมอื่น ๆ เช่น การวิเคราะห์โครงสร้างประโยค การใช้โปรแกรมตัดประโยคสำเร็จรูป มาสนับสนุน ย่อมจะช่วยให้การจับคู่ประโยคมีถูกต้องมากยิ่งขึ้นอย่างแน่นอน

6.2 ข้อเสนอแนะ

1. ค่าความถูกต้องจะเพิ่มมากขึ้น ถ้าเลือกกลุ่มของคำหยุดที่จะกำจัดออกให้เหมาะสมกับข้อความที่นำมาใช้ เพราะสำหรับข้อความที่สั้น คำหยุดสามารถช่วยให้เปอร์เซ็นต์ความถูกต้องในการจับคู่ประโยคเพิ่มขึ้นได้

2. ถ้ามีอัลกอริทึมที่ดีช่วยในการหาจำนวนวรรคของประโยคภาษาไทย จะทำให้ความถูกต้องในการจับคู่ประโยคแบบ 1:N มีค่าเพิ่มขึ้น

3. สามารถนำผลจากการจับคู่คำด้วยอนุกรมเวลานี้ไปช่วยในงานด้านการแปลได้ โดยเป็นส่วนเพิ่มเติมจากการแปลโดยใช้ดิกชันนารี เพราะความหมายของคำบางคำก็ไม่ปรากฏในดิกชันนารี แต่สามารถอาศัยความหมายที่ได้จากการจับคู่คำด้วยอนุกรมเวลาช่วยได้

รายการอ้างอิง

[1] Veronis, J. From the Rosetta stone to the information society: a survey of parallel text processing. In **Parallel Text Processing Alignment and Use of Translation Corpora**, 1-17. Netherlands : Kluwer Academic Publishers, 2000.

[2] National Electronics and Computer Technology Center (NECTEC). **Smart Word Analysis for Thai (SWATH)**[Online]. Available from : http://www.links.nectec.or.th/download.php [2007,July 8].

[3] SIL International 2000. **KTAGGER**[Online]. Available from : http://www.sil.org/computIng/catalog/show_software.asp?id=22 [2007,July 8].

[4] Supnithi, T. **Introduction: Enter a world of inter-cultural.** Bangkok : National Electronics and Computer Technology Center, 2005.

[5] Melby, A. K. Sharing of translation memory databases derived from aligned parallel text. In **Parallel Text Processing Alignment and Use of Translation Corpora**, 347-368. Netherlands : Kluwer Academic Publishers, 2000.

[6] Charoenpornsawat, P., Sornlertlamvanich, V., and Charoenporn, T. Improving Translation Quality of Rule-based Machine Translation. In **International Conference on Computational Linguistics**, 81-87. Taiwan, 2002.

[7] Hutchins, J. Towards a definition of example-based machine translation. In **MT Summit X: Proceedings of Workshop on Example-Based Machine Translation, Phuket, 2005**, 63-70.

[8] Callison-Burch, C., Koehn, P., and Osborne, M. Improved Statistical Machine Translation Using Paraphrases. In **Proceedings NAACL**, 2006.

[9] Streiter, O., Carl, M., and Iomdin, L. A Virtual Translation Machine for Hybrid Machine Translation. In **Proceedings of the Dialogue'2000 International Seminar in Computational Linguistics and Applications**, Russia, 2000.

[10] Phaholphinyo, S., Modhiran, T., Kritsuthikul, N., and Supnithi, T. A practical of memory-based approach for improving accuracy of MT. In **Proceedings of Conference MT Summit X**, 2005.

[11] Baeza – Yates, R., and Ribeiro - Neto, B. **Modern Information Retrieval**. USA: PEARSON Addison Wesley, 1999.

[12] กลุ่มคำหยุดทั่วไป, Available from : http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words [2007,July 8].

[13] กลุ่มคำหยุดกูเกิล, Available from : http://www.ranks.nl/tools/stopwords.html [2007,July 8].

[14] กลุ่มคำหยุดภาษาไทย, Available from : http://www.cs.sci.ku.ac.th/~fscichu/ThaiIr/publications.html [2007,July 8].

[15] Corver, N., and Van Riemsdijk, H. **Semi-lexical categories : the function of content words and the content of function words.**, 2001.

[16] Han, J., and Kamber, M. **Data Mining Concepts and Techniques**. USA: Morgan Kaufmann Publishers, 2001.

[17] Sun, R., and Lee Giles, C. **Sequence learning : paradigms, algorithms, and applications.**, 2001.

[18] Melamed, I. D. Pattern recognition for mapping bitext correspondence. In **Parallel Text Processing Alignment and Use of Translation Corpora**, 25-46. Netherlands : Kluwer Academic Publishers, 2000.

[19] Simard, M., Foster, G.F., and Isabelle, P. Using Cognates to Align Sentences in Bilingual Corpora. In **Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation**, 67-81. ,1992.

[20] Haruno, M., and Yamazaki, T. High-performance bilingual text alignment using statistical and dictionary information. In **Proceedings of ACL'96**, 131-138. USA, 1996.

[21] Cardenas, C. **Translating Text Using Bursty Sequences**. Master Project, Department of Computer Science and Engineering University of California Riverside, 2005.

[22] Wu, D. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. In **Computational Linguistics**, 377-403. USA: MIT Press, 1997.

[23] Gispert, A., and others. Improving Statistical Word Alignments with Morpho-syntactic Transformations. In **Proceedings of 5th International Conference on Natural Language Processing (FinTAL)**, 368-379. Finland, 2006.

[24] Aroonmanakun, W. Collocation and Thai Word Segmentation. In **Proceedings of SNLP-Oriental COCOSDA**, 2002.

[25] Meknavin, S., Charoenpornsawat, P., and Kijsirikul, B. Feature-based Thai Word Segmentation. In **Proceedings of the Natural Language Processing Pacific Rim Symposium**, 41-48. Thailand, 1997.

[26] Charoenpornsawat, P., and Sornlertlamvanich, V. Automatic sentence break disambiguation for Thai. In **International Conference on Computer Processing of Oriental Languages (ICCPOL)**, 231-235. Korea, 2001.

[27] Mitrapiyanurak, P., and Sornlertlamvanich, V. The Automatic Thai Sentence Extraction. In **Proceedings of 4th Symposium on Natural Language Processing**, 23-28. Thailand, 2000.

[28] Kawtrakul, A., and Boonkwan, P. An Integrated Tool for Translation-Memory Maintenance. In **Papillon 2004 Workshops on Multilingual Lexical Databases**, 2004.

[29] คัมภีร์ไบเบิล, Available from : http://bibledatabase.org/cgi-bin/bib_search/bible.cgi?BIBLE=50&BOOK=69&CHAP=40&SEARCH=jesus%20king%20lord&Read=Read&FIRST=OK [2006,August 1].

[30] ตัวอย่างคู่ประโยคจากดิกชันนารี, Available from : Human Language Technology Laboratory (National Electronics and Computer Technology Center).

[31] กฎหมายไทย, Available from : http://www.parliament.go.th/files/library/t-b01.htm [2008,January 8].

[32] GIZA++, Available from : http://code.google.com/p/giza-pp/ [2008, January 10].

[33] Och, F.J., and Ney, H. A Systematic Comparison of Various Statistical Alignment Models. **Computational Linguistics** volume 29 March 2003 : 19-51.

ภาคผนวก

ภาคผนวก ก

คำหยุด

กลุ่มคำหยุดทั่วไป มี 319 คำ

a	anything	but	empty	further
about	anyway	by	enough	get
above	anywhere	call	etc	give
across	are	can	even	go
after	around	cannot	ever	had
afterwards	as	cant	every	has
again	at	co	everyone	hasnt
against	back	computer	everything	have
all	be	con	everywhere	he
almost	became	could	except	hence
alone	because	couldnt	few	her
along	become	cry	fifteen	here
already	becomes	de	fify	hereafter
also	becoming	describe	fill	hereby
although	been	detail	find	herein
always	before	do	fire	hereupon
am	beforehand	done	first	hers
among	behind	down	five	herself
amongst	being	due	for	him
amoungst	below	during	former	himself
amount	beside	each	formerly	his
an	besides	eg	forty	how
and	between	eight	found	however
another	beyond	either	four	hundred
any	bill	eleven	from	i
anyhow	both	else	front	ie
anyone	bottom	elsewhere	full	if

in	my	our	somehow	those
inc	myself	ours	someone	though
indeed	name	ourselves	something	three
interest	namely	out	sometime	through
into	neither	over	sometimes	throughout
is	never	own	somewhere	thru
it	nevertheles	part	still	thus
its	s	per	such	to
itself	next	perhaps	system	together
keep	nine	please	take	too
last	no	put	ten	top
latter	nobody	rather	than	toward
latterly	none	re	that	towards
least	noone	same	the	twelve
less	nor	see	their	twenty
ltd	not	seem	them	two
made	nothing	seemed	themselves	un
many	now	seeming	then	under
may	nowhere	seems	thence	until
me	of	serious	there	up
meanwhile	off	several	thereafter	upon
might	often	she	thereby	us
mill	on	should	therefore	very
mine	once	show	therein	via
more	one	side	thereupon	was
moreover	only	since	these	we
most	onto	sincere	they	well
mostly	or	six	thick	were
move	other	sixty	thin	what
much	others	so	third	whatever
must	otherwise	some	this	when

whence	wherein	whither	why	yet
whenever	whereupon	who	will	you
where	wherever	whoever	with	your
whereafter	whether	whole	within	yours
whereas	which	whom	without	yourself
whereby	while	whose	would	yourselves

กลุ่มคำหยุดภาษาไทย มี 77 คำ

ที่	ทาง	ออก	จึง	ลง
ใน	กล่าว	นั้น	หาก	ละ
ว่า	โดย	หรือ	แก่	ในช่วง
และ	ซึ่ง	เมื่อ	เช่น	เดียว
จะ	ต้อง	ขณะ	ทุก	ระหว่าง
มี	จำ	เปิด	ไว้	เฉพาะ
ได้	ก็	แห่ง	บาง	ต่าง
ของ	แต่	ร่วม	เพียง	อย่างไรก็
ให้	ยัง	เพราะ	พร้อม	ใช้
เป็น	ขึ้น	ไร	ใต้	เพิ่ม
นี้	อย่าง	กว่า	ดู	เนื่องจาก
ไม่	ทั้ง	มาก	อาจ	ใด
จาก	เพื่อ	ด้าน	หลาย	นิ
ไป	เข้า	นอก	ตาม	
มา	แล้ว	ใหม่	ดังกล่าว	
ด้วย	อยู่	ก่อน	พบ	

ภาคผนวก ข

ผลการทดลองที่เกี่ยวข้อง

ผลการทดลองจับคู่ประโยคแบบ 1:1

ผลจากการทดลองที่ 8

ข้อมูลเข้า : บทที่ 1 ปฐมกาล (Genesis)

จำนวนประโยค : 31 ประโยค

ตารางที่ ข-1 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตันกับไบเบิลขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	26	27	27	26	27	24
5%	29	28	28	28	28	28
10%	26	27	27	21	24	26
20%	28	24	25	26	25	25
30%	21	22	22	20	22	21

ตารางที่ ข-2 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับไบเบิลขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	12	26	26	13	24	24
5%	26	28	25	22	28	25
10%	24	26	25	17	25	25
20%	26	26	23	24	26	25
30%	22	22	22	20	23	22

ผลจากการทดลองที่ 9

ข้อมูลเข้า : บทที่ 1 ของปฐมกาล (Genesis) อพยพ (Exodus) เลวีนิติ (Leviticus) กันดารวิถี (Numbers) และพระบัญญัติ (Deuteronomy)

จำนวนประโยค : 175 ประโยค

ตารางที่ ข-3 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตันกับไบเบิลขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	38	35	35	43	42	45
5%	60	58	64	65	65	68
10%	65	59	58	72	67	64
20%	63	55	54	58	58	58
30%	59	58	52	56	55	54

ตารางที่ ข-4 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับไบเบิลขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	27	25	26	26	28	29
5%	54	56	56	54	54	60
10%	59	57	52	60	62	59
20%	51	45	46	59	57	59
30%	58	52	55	62	62	58

ผลจากการทดลองที่ 10

ข้อมูลเข้า : บทที่ 1 ถึง 25 ของปฐมกาล (Genesis)

จำนวนประโยค : 718 ประโยค

ตารางที่ ข-5 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตันกับไบเบิลขนาดยาว

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	317	320	324	321	326	327
5%	362	377	389	364	378	385
10%	378	368	356	406	389	385
20%	321	346	336	355	382	377
30%	273	289	286	290	304	305

ตารางที่ ข-6 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับไบเบิลขนาดยาว

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	231	231	232	254	262	267
5%	373	378	376	366	363	359
10%	375	372	360	372	366	367
20%	348	336	331	382	382	372
30%	258	265	259	291	289	289

ผลจากการทดลองที่ 11

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 70 ตัวอย่าง

จำนวนประโยค : 70 ประโยค

ตารางที่ ข-7 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตันกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	23	25	25	20	22	21
5%	38	35	29	32	29	29
10%	37	37	35	31	30	30
20%	23	19	19	19	18	19
30%	20	21	22	22	21	19

ตารางที่ ข-8 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	26	26	26	25	25	26
5%	29	28	27	24	23	26
10%	35	35	34	28	29	31
20%	22	23	24	22	22	23
30%	19	23	22	19	20	19

ผลจากการทดลองที่ 12

ข้อมูลเข้า : ตัวอย่างคู่ประโยคจากดิกชันนารี 304 ตัวอย่าง

จำนวนประโยค : 304 ประโยค

ตารางที่ ข-9 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตันกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	116	107	106	82	80	84
5%	131	135	135	108	107	109
10%	107	113	112	92	95	100
20%	61	61	59	63	67	65
30%	30	35	38	25	32	40

ตารางที่ ข-10 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับตัวอย่างคู่ประโยคจากดิกชันนารีขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	72	82	81	63	67	70
5%	122	120	111	99	89	88
10%	107	111	120	95	101	106
20%	49	51	59	61	62	60
30%	31	28	36	28	26	34

ผลจากการทดลองที่ 13

ข้อมูลเข้า : ข้อกฎหมาย 50 ข้อ

จำนวนประโยค : 50 ประโยค

ตารางที่ ข-11 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแบบแมนฮัตตันกับข้อกฎหมายขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	22	23	23	22	23	22
5%	26	24	26	28	25	26
10%	25	24	26	28	26	26
20%	27	24	25	28	26	26
30%	26	29	29	28	29	31

ตารางที่ ข-12 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับข้อกฎหมายขนาดสั้น

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	20	17	18	21	18	19
5%	23	23	23	24	23	25
10%	25	24	22	25	24	21
20%	27	25	24	26	25	23
30%	25	26	27	25	25	28

ผลจากการทดลองที่ 14

ข้อมูลเข้า : ข้อกฎหมาย 102 ข้อ

จำนวนประโยค : 102 ประโยค

ตารางที่ ข-13 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างแมนฮัตตันกับข้อกฎหมายขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	48	45	45	50	49	48
5%	53	48	51	55	53	52
10%	55	54	50	45	49	48
20%	49	46	45	55	48	47
30%	45	45	47	43	45	48

ตารางที่ ข-14 ผลการทดลองจับคู่ประโยคแบบ 1:1 โดยใช้ฟังก์ชันระยะห่างยุคลิเดียนกับข้อกฎหมายขนาดกลาง

	กำจัดกลุ่มคำหยุดกูเกิล			กำจัดกลุ่มคำหยุดทั่วไป		
ขนาดของหน้าต่าง	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3	ใช้คู่คำอันดับที่ 1	ใช้คู่คำอันดับที่ 1, 2	ใช้คู่คำอันดับที่ 1, 2, 3
1%	46	45	42	50	47	46
5%	48	48	44	47	45	43
10%	51	49	45	45	44	44
20%	49	43	39	51	50	43
30%	46	46	43	48	46	46

ภาคผนวก ค

ผลงานตีพิมพ์

งานประชุมวิชาการนานาชาติ 2nd International Conference on Advances in Information Technology 2007 ณ โรงแรมเอเชีย กรุงเทพมหานคร ระหว่างวันที่ 1-2 พฤศจิกายน 2550 ในบทความเรื่อง Parallel Text Alignment Using Bursty Sequences

Parallel Text Alignment Using Bursty Sequences

Sirinun Sintuwatin and Chotirat Ann Ratanamahatana
Department of Computer Engineering
Chulalongkorn University, Bangkok, Thailand
sirinun381@hotmail.com, ann@cp.eng.chula.ac.th

Abstract

As applications based on parallel corpora (parallel text) has increasingly expanded, especially in the areas of cross-language information retrieval, machine/human translation, natural language processing, and multilingual lexicography, parallel-text processing has become the heart of the development. In this paper, we propose a novel word alignment technique that would be the basis for sentence alignment procedure. We exploit a notion of time series representation, recording the position and frequency of word appearance, without any requirement of any linguistic knowledge, e.g. grammar/syntax, structure, translation (from dictionary lookup, etc. Our intuition lies in the belief that similar words in any multilingual parallel text should possess similar frequency and the position of word occurrences. We demonstrate its utility and effectiveness using the bible text in English and Thai languages.

Keywords: Parallel Text Processing, Word Alignment, Sentence alignment, English - Thai, Time Series, Bursty Sequences

1. Introduction

As multilingualism has evolved dramatically over the past decade, the need of parallel text processing has arisen accordingly, where learning non-mother tongue languages seems to be quite burdensome to most population. Though current machine translation techniques can partly facilitate the cross-language translation, the accuracy and quality of current translation software is still substandard. Part of the reasons is that most of them translate word by word, with some help of syntax and language structure. However, since in typical languages, synonym poses some problems in translation, each word may have several meanings, so it is difficult to select the correct or appropriate one that corresponds to the context. Instead, imagine if we could translate sentence by sentence, the quality of the translation should be better. This has motivated the emergence of sentence-based alignment in parallel corpora (where corpora of texts are accompanied by their translation).

Existing sentence alignment techniques include sentence length correlation, semantic load base on POS tagging, word correspondence base on cognate, and word correspondence based on anchor word list [1]. In this work, we are proposing an intuitive and simple sentence alignment technique based on anchor word list correspondence, where no linguistic knowledge is needed. Instead, we exploit a basic concept similar to dictionary lookup; if we know exact meanings between pairs of words (subject to surrounding context) in two languages, the sentence alignment could be achieved more easily and accurately.

In this paper, we propose a time series approach to represent the anchor word list; each word will be represented by one time series. Specifically, the word occurrence frequency and its position of each word within each sliding window are recorded as a bursty sequence. These sequences are utilized in the word alignment process, where the similarity among time series will be calculated and pattern matching is performed.

The rest of the paper is organized as follows. Section 1 explains what parallel text is and its advantage, as well as our motivation. Section 2 gives some background and related work. Our proposed work in Section 3 includes text preprocessing, time series extraction, similarity measurement. Sections 4 and 5 include the experiment evaluation results and the discussion, respectively. And finally, we conclude our work and discuss some future directions in Section 6.

1.1 Parallel Text

Parallel text is a text in one language accompanied with its translation in another language [2] such as Harry Potter in English – Thai, Bible in English – Spanish, etc. In reality, there exist countless sets of parallel corpora, where we could exploit this information in various ways and applications. These include 1) an alignment methodology at various levels such as words, sentences, and paragraphs, where each unit in two

different languages are matched correspondingly; 2) Application of parallel texts in fields such as lexicography, terminology, translation, and cross-language information retrieval, where Parallel text is used in various areas of linguistic research; 3) Evaluation of alignment method; and 4) Exploration of available resources [3].

1.2 Motivation

Intuitively, in parallel corpora, each word in one language should have similar frequency and positions of occurrences with another language. Since this property can become very useful for the alignment, we can extract these features from the parallel corpora. We test our hypothesis by a simple experiment. By counting frequency of occurrences of each word within a sliding window starting from the beginning of the text and moving one word at a time until the end, we can record these values as sequences of numbers, i.e. time series data or a bursty sequence. However, typically, two different language texts will have different lengths or total number of words. Therefore, we decided to rescale both sequences to the same size to simplify the comparison. Theoretically, each word should have unique corresponding time series, and two time series of the same word in two different languages should have similar shape. We use the whole Bible text in English and Thai (each with almost one million words) as a motivating example. Figure 1 and Figure 2 show the time series of the word "god" in English language and Thai language (พระเจ้า) from the Bible using the approach explained above. It is apparent that both time series do have very similar shape. However, if we obtain a time series of another word with different meanings, its shape should be less similar.

Figure 3 illustrates a time series of the word "fly", which evidently has different shape. For word alignment between two languages, we can simply apply a similarity measurement to compute the similarity or distance between each word pair to discover an appropriate match.

2. Background and Related Work

Several text alignment methods in parallel corpora have been proposed, where an alignment can be done in word, phrase, or sentence levels. Using sentence length as a feature is one of the simplest sentence alignment approaches, assuming that short sentences tend to have short translations, and vice versa. The position of sentence appearance is also crucial as the probability of a first sentence matching with the last sentence in the parallel text is quite low [3, 4].

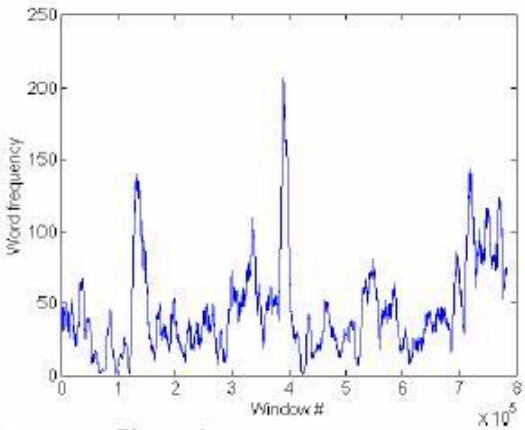


Figure 1. Time series of "god"

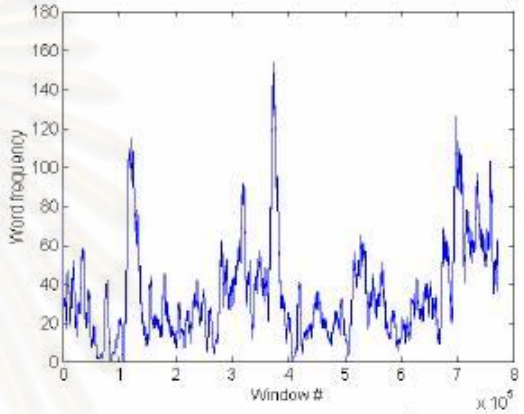


Figure 2. Time series of "พระเจ้า"

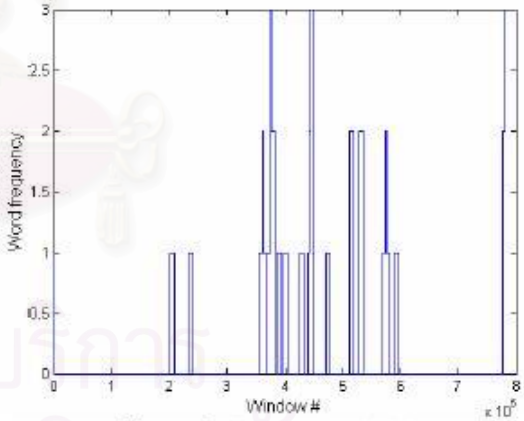


Figure 3. Time series of "fly"

However, the length of the sentence along generally does not work well in many languages; for example, a Thai translation text of English usually is longer due to Thai's expressive nature of writing. Therefore, some researchers proposed a cognate-based alignment method. Cognates may be symbols or words with similar spelling such as

"langage" in French and "language" in English language [5]. This technique is suitable for languages that are in the same family or root. On the other hand, it would fail in some languages that are not in this category such as English-Thai, English-Japanese, English-Chinese, etc., where the alphabets are different. Another sentence alignment method is lexical anchoring, where anchor words may be obtained from a bilingual dictionary [6, 7] or from word alignment process [8]. Word alignment techniques include stochastic inversion transduction grammars [9], statistic, and linguistic [10]. However, the main problem in many languages, such as Thai, is that no spacing is needed between words or sentences, making natural language processing for these language types much harder when accurate word or sentence segmentation is desirable [11, 12]. Various word segmentation and sentence alignment methods have been proposed [13], some using surface feature of languages and bilingual dictionary [14][15].

3. Our proposed work

Our work is based on the concept of anchor word lists since current Thai sentence segmentation is still not as accurate as the word segmentation, hence obtaining correct length of the sentence may be difficult. Our proposed work can align 1 English sentence to several Thai phrases, so sentence segmentation is unnecessary.

Another one of the most important tasks is to obtain the parallel text of any two languages. In this work, we have chosen English-Thai parallel text, where one is the translation of another. Then, some preprocessing is needed since each language has different structures; for example, Thai language does not have any inflexion, while English language does, and therefore needs stemming.

3.1 English text preprocessing

Inflexion in English language appears as variation in subject, tense, and number; it has rules for adding -s, -es, -d, -ed, or to changing its verb forms, including the morpheme derivation such as –ly, -ness, -ize, etc. Before an alignment could be made, "stemming" must be performed; we uses KTAGGER stemming software [16]. Other preprocessing include removals of special characters such as quotations, brackets, etc. otherwise words with and without these symbols will be treated as two different words, hence degrade the word alignment quality. However, period and question mark must not be removed since they indicate the sentence boundary (with some exception in cases such as periods in abbreviation, etc.). English language already has spacing between words so it does not need to do word segmentation.

3.2 Thai text preprocessing

Since Thai language has no word boundary, word segmentation is needed; in this work, we have chosen SWATH [17], word segmentation software developed by National Electronics and Computer Technology Center (NECTEC), Thailand. Similarly to English text preprocessing, all special characters in Thai text are also removed. However, since Thai language does not have any sentence boundary marker other than a space that is occasionally put between phrases, the period and question mark characters are both removed at this stage. Note that since Thai language does not have any sentence boundary marker, all spaces are recorded to help with further sentence alignment process. Generally, comparing with typical English sentences, Thai sentence tends to be longer or contains multiple sentences or phrases. Therefore, in our sentence alignment stage, sentence segmentation will not be necessary. Instead, our goal is to align each English sentence with one or more Thai phrases (or zero if there is no match).

3.3 Time series extraction

The heart of our approach is the idea of representing words in the text with bursty sequences/time series, where their shape similarity will be measured to achieve the parallel text word alignment. As briefly described in Section 1.2, each data point in the time series simply is the word frequency within a current sliding window, starting from the beginning until the end of the text, as illustrated in Figure 4. Figures 5 and 6 show some examples after conversion of text to time series; x-axis is the sliding window number, and y-axis is the selected word frequencies within each sliding window.

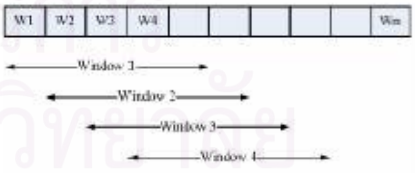


Figure 4. Sliding windows

As mentioned above, the text length between English and Thai text are not equal, therefore making time series to be of different length as well. For example, English bible has around 790,000 words while Thai bible has almost one million

words. To simplify the similarity calculation, we rescale both time series to have the same length, by downsampling the longer one to the size of the shorter.

Another important parameter is the size of the sliding window. It is apparent from Figures 5 and 6 that smaller window size will result in noisier and (slightly) longer time series and larger window size will result in smoother and shorter time series, which could essentially affect the similarity measurement result. Specifically, Figure 5 is a time series of the word "earth" from English bible with the window size of 8000, and Figure 6 is of the same data with the window size of 2000.

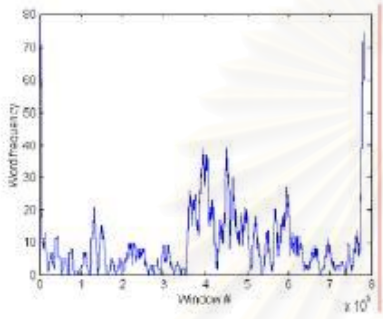


Figure 5. Time series of a word "earth" with the window size 8000

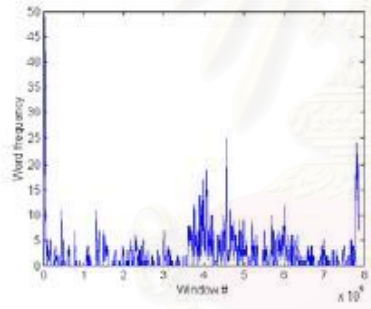


Figure 6. Time series of a word "earth" with the window size 2000

Note that since we essentially are interested in the shapes of the time series, normalization can help control the wide range of values in the y-axis. Therefore, we z-normalize all our time series data before any further distance calculation. Z-score normalization is suitable for this type of data whose minimum and maximum values are not known. Z-score normalization is defined as follows:

$$Z_i = \frac{X_i - M}{SD}$$

where Z_i is a normalized value of X_i, M is the mean value of time series X, and SD is standard deviation of time series X.

3.4 Word Alignment

In this work, we simply use Euclidean distance metric to measure similarity among time series. As mentioned earlier, all time series are rescaled in the x-axis such that all time series of both languages do have the same length.

3.4.1 Time series reduction. We decrease the length of a longer time series down to the length of a shorter one, simply by finding the length difference between English and Thai time series then removing that many points from the longer one. For example, given an English time series of length 500 data points a Thai time series of length 550 points, 50 data points will be removed from the Thai time series. The data positions that will be removed are ones that evenly divide $\lfloor 550/50 \rfloor$, i.e., all the positions p that $(p \ mod \ 11) = 0$.

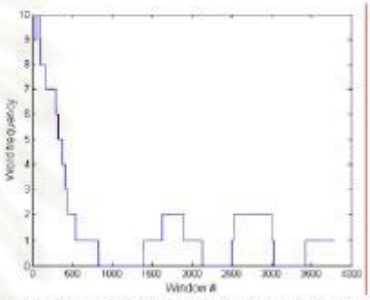


Figure 7. A raw time series of a word "day"

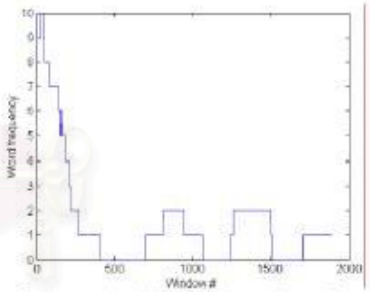


Figure 8. A resulted time series after a 50 percent reduction of the time series in Figure 7

To further speed up the overall distance calculation, all long time series (both English and Thai) may be resized to shorter ones. Figure 7 shows a raw time series of a word "day" before any reduction, and Figure 8 is after its 50 percent reduction in length, where its shape is still similar to the raw data.

3.4.2 Similarity Measurement. After the time series extraction is completed for all the words in both languages, the Euclidean distance metric is used to measure similarities among words in

English-Thai pairing; the Euclidean distance is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} \left(X_i - Y_i\right)^2}$$

where X and Y are two time series of length n, with $X = X_1 X_2 \ldots X_n$ and $Y = Y_1 Y_2 \ldots Y_n$. 3-nearest neighbors are performed to maintain the best 3 matches (3 smallest Euclidean distances) as the high-potential candidates for a correct word alignment.

As part of our preliminary experiments, we test the result of the alignment using all the words in the text vs. the result when stop words are excluded from the calculation. Our result suggests that stop words should be ignored in this problem, where the alignment accuracy scores higher. This seems sensible because stop words are generally viewed as noise, as they are often found in every sentence and commonly regarded as 'functional words' which do not carry meaning and disrupt the ability to achieve effective outcome. Some examples of stop words include 'a', 'an', 'of', 'I', 'it', 'the', etc. To further speed up the calculation, we can also ignore the pairings that have radical difference in word frequencies. For example, the frequency of a particular word in English is 40, but that in Thai is only 2, as they have such a low chance to be matched as a corresponding translation of each other.

4. Experiment

The perhaps most prevalent parallel text is the Bible. Therefore, we choose a bible, King James Version in English and Thai. Our proposed method is tested on 3 sizes of the dataset.

- The first chapter of Genesis, with 816 English words and 812 Thai words (about 1-page text).
- The first chapter of 5 books: Genesis, Exodus, Leviticus, Numbers, and Deuteronomy, with 4287 English words (about 5-page text) and 4499 Thai words (about 5.5-page text), and
- Twenty-five chapters of Genesis, with 17055 English words (about 21.5 pages) and 19125 words (about 24 pages).

The experiments are carried out as follows

1. Data preprocessing: word segmentation, stemming, and removal of special characters, numbers, brackets, etc.

2. Generating time series for every word (excluding stop words) using predefined window size. We also keep the information about each word's appearances in sentences for evaluation purpose. For example, the word 'morning' appears in sentences number 5, 10, and 48, and these numbers are stored in a hash key.

3. Length reduction (if necessary), and

4. Calculating the distance between each English-Thai word pairings. A list of 3-nearest-neighbor (Thai words) is maintained for each English word.

To evaluate our results, we count the number of correct alignment pair. It is important to note that our ultimate goal is to get lists of anchor words for sentence alignment. Therefore, the alignment that matches words in English and Thai that are contained within correct sentences will be considered correct.

More specifically, we will report the correctness by using a mean reciprocal rank, which takes probability of correctness into consideration. A reciprocal rank of a query's responses is the multiplicative inverse of the rank of the correct answer; examples are shown in Table 1.

Table 1. Reciprocal rank

Rank	Number of Correct word alignment	Reciprocal rank
1	55	1
2	6	1/2
3	2	1/3

Given total number of words = 69, Mean reciprocal rank = [(1*55)+(1/2*6)+(1/3*2)] / 69 = 0.85.

Experiment 1
Input: First chapter of Genesis
TextLength = Min(|EnglishText|, |ThaiText|) = 812
Length of time series: 812 – Window size +1
Window size: 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of TextLength
Number of unique English words (excluded stop words): 69

Table 2. Experiment results on short parallel text

Window size	Rank 1	Rank 2	Rank 3	Mean Reciprocal Rank
1%	56	5	0	0.85
5%	57	5	3	0.88
10%	60	5	0	0.91
20%	59	6	1	0.90
30%	61	5	0	0.92
40%	63	3	0	0.93
50%	59	3	3	0.89
60%	59	5	0	0.89
70%	63	3	1	0.94
80%	65	0	0	0.94
90%	60	5	1	0.91

Table 2 suggests that short parallel corpora of about 1-page length can give high word alignment accuracy. The mean reciprocal rank increases as the window size increases, and decreases if the window size gets too large.

Experiment 2
Input: First chapter of 5 books, i.e., Genesis, Exodus, Leviticus, Numbers, and Deuteronomy
TextLength = Min(|EnglishText|, |ThaiText|) = 4287
Length of time series: 4287 – Window size +1
Window size : 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of TextLength
Number of unique English words (excluded stop words): 338

Table 3. Experiment results on medium-length parallel text

Window size	Rank 1	Rank 2	Rank 3	Mean Reciprocal Rank
1%	84	22	18	0.30
5%	135	33	16	0.46
10%	148	19	24	0.49
20%	168	28	9	0.55
30%	152	24	20	0.50
40%	150	25	27	0.51
50%	151	30	15	0.51
60%	195	29	12	0.63
70%	177	27	16	0.58
80%	189	19	10	0.60
90%	231	12	5	0.71

Comparing the results in Table 3 with those in Table 2, while both demonstrate the same behavior, alignment accuracy in short parallel text tends to be much better than the accuracy in medium-length parallel corpora. We also found that generally when window size is larger than 50%, some English words match with too many Thai words in each rank (many ties occur). Taking that into account, the best window size of this experiment appears to be at 20%. In addition, to observe an effect of time series length to the alignment accuracy, we did another experiment similar to Experiment 2, but with different reduction ratio on time series length, as shown in Experiment 3 below.

Experiment 3
Input: same as Experiment 2
TextLength: 4287
Length of time series: 4287 – Window size +1
Window size: 20% (from Experiment 2)
Reduction Ratio: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Number of unique English words: 338

Table 4. Experiment results on medium-length parallel text with various time series reduction

Reduction Ratio	Rank 1	Rank 2	Rank 3	Mean Reciprocal Rank
0.1	159	32	10	0.53
0.2	161	31	9	0.53
0.3	160	30	11	0.53
0.4	162	33	6	0.53
0.5	166	31	12	0.55
0.6	160	33	10	0.53
0.7	159	33	10	0.53
0.8	165	31	10	0.54
0.9	161	32	8	0.53
1.0	168	28	9	0.55

From Table 4, we can see that reduction in the length of time series generally does not have detrimental effect on the alignment, suggesting that we could largely reduce the time series length before similarity calculation to significantly speed up the computation.

Experiment 4
Input: 25 chapters of Genesis
TextLength = Min(|EnglishText|, |ThaiText|) = 17055
Window size: 20% (from Experiment 2)
Reduction Ratio: 0.1 and 0.3
Number of unique English words (excluded stop words): 771

Table 5. Experiment results on long parallel text

Reduction Ratio	Rank 1	Rank 2	Rank 3	Mean Reciprocal Rank
0.1	288	83	55	0.45
0.3	313	73	56	0.48

Experiment results demonstrate that we could reduce the time series length without hurting too much of the alignment accuracy. However, as expected, larger corpus contains higher word variety, hence reduce the accuracy, comparing with the experiments on smaller corpus.

5. Discussion

Our experiment results suggest that our proposed method works quite well with short parallel text as it contributes less noise to the time series. On large parallel text, we have shown that reducing its time series length is one way to smooth out the noise, hence increasing the alignment accuracy to some degree. In addition, the window size could largely affect the accuracy. Generally, large window size gives better results (with similar

smoothing effect on time series), but too large a window will hurt the accuracy since it smoothes and removes out too much information.

This work has been inspired from our earlier hypothesis that frequencies and positions of words in parallel corpora should be very similar. With some analyses of our experiment results, we have found several factors that could have contributed to the incorrect alignments.

1. There are numerous English words that have different meaning but they translate to the same word in Thai; both have similar overall meaning, but the choice of words may cause the confusion, such as

English	Thai
And God said, Let there be light: and there **was** light	พระเจ้าตรัสว่า "จงให้มีความสว่าง" แล้วความสว่างก็เกิดขึ้น
And God said, Let grass **come up** on the earth	พระเจ้าตรัสว่า "จงให้แผ่นดินเกิดต้นหญ้า
And God said, Let the earth give **birth** to all sorts of living things	พระเจ้าตรัสว่า "จงให้แผ่นดินโลกเกิดสัตว์ที่มีชีวิตตามชนิดของมัน

Here, "birth", "come up", and "was" are all translated to "เกิด. in Thai, where normally "was" and 'come up' do not have direct meaning of 'เกิด'.

2. Phrase translate to only one compound word such as

English	Thai
And God said, Let the earth give birth to all sorts of living things, cattle and **all things moving on the earth**, and beasts of the earth after their sort: and it was so.	พระเจ้าตรัสว่า "จงให้แผ่นดินโลกเกิดสัตว์ที่มีชีวิตตามชนิดของมัน สัตว์ใช้งาน สัตว์เลื้อยคลาน และสัตว์ป่าบนแผ่นดินโลกตามชนิดของมัน" ก็เป็นดังนั้น

The phrase "all things moving on the earth" gets translated to "สัตว์เลื้อยคลาน" (meaning "reptile") which is not quite a correct translation, making it extremely difficult to get correct alignment in this case.

3. Some word has been ignored because it is translated together with other words such as

English	Thai
On the far side of Jordan in the land of Moab, Moses gave the **people** this law, saying,	โมเสสได้เริ่มอธิบายพระราชบัญญัตินี้ที่ในแผ่นดินโมอับฟากแม่น้ำจอร์แดนข้างนี้ว่า

The word "people" has been ignored and gets disappeared in the translation, where its implicit meaning is assumed in the Thai translation (that the law must be given to the "people", hence the omission).

4. Word segmentation is crucial to the alignment quality. The word segmentation software used in this work does not work well with proper names, giving wrong segmentation which makes it hard to align one whole English proper name with several segmented portion of the name in Thai, such as

English	Thai	Thai Word segmentation
Jacob	ยาโคบ	ยา-โค-บ
Benjamin	เบนยามิน	เบน-ยา-มิ-น
Hebrew	ฮีบรู	ฮีบ-รู

Jacob and Benjamin will generate a total of 6 time series, "ยา", "โค", "บ", "เบน", "มิ", and "น", instead of 2 time series. Moreover, "ยา" itself has its own meaning "medicine" or "medication". Hence, the time series of the word "ยา" will be incorrect and may not correctly align to "medicine".

5. Several English words have almost the same meaning when translate to Thai, for example

English	Thai
At the first God made the heaven and the **earth**. And the **earth** was waste and without form	ในเริ่มแรกนั้นพระเจ้าทรงสร้างจักรวาลสร้างฟ้าและแผ่นดินโลก แผ่นดินโลกนั้นก็ปราศจากรูปร่างและว่างเปล่าอยู่
On the far side of Jordan in the **land** of Moab, Moses gave the people this law, saying,	โมเสสได้เริ่มอธิบายพระราชบัญญัตินี้ที่ในแผ่นดินโมอับฟากแม่น้ำจอร์แดนข้างนี้ว่า

Both "earth" and "land" get translated to "แผ่นดิน", but we can not align "earth" and "land" to "แผ่นดิน" because the time series "earth" and "land" are different from that of "แผ่นดิน", as illustrated in Figures 9 - 11. However, if we have realized this phenomenon and combine the two time series of "earth" and "land" together, it could lead to a perfect alignment with the time series of "แผ่นดิน".

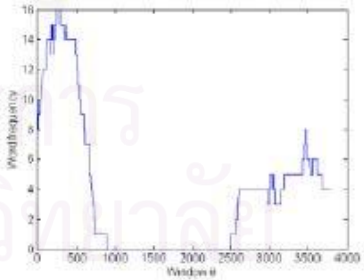


Figure 9. Time series of "แผ่นดิน"

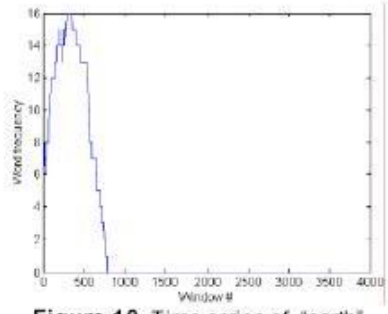


Figure 10. Time series of "earth"

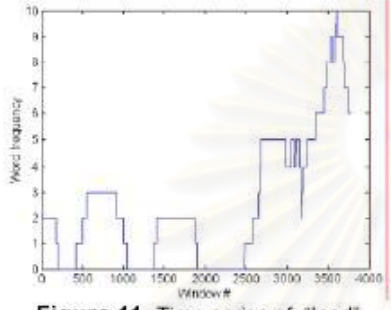


Figure 11. Time series of "land"

6. Conclusion and future work

We have proposed a novel, yet intuitive and simple word alignment technique, based on bursty sequences and their similarity matching. We exploit a notion of time series representation, recording the position and frequency of word appearances. Our technique is attractive in that no linguistic knowledge is needed. The experiment results on parallel English – Thai Bible text reconfirm our hypothesis. In particular, the word alignment performance is highly satisfactory, especially in small parallel text. We do hope that this proposed technique would become a basis for future sentence alignment procedure.

7. References

[1] H. M. Caseli, M. G. V. Nunes, "Evaluation of Sentence Alignment Methods on Portuguese-English Parallel Texts," 2003.

[2] "parallel text," wikipedia, 1 July 2007.

[3] J. Veronis, "From the Rosetta stone to the information society," in *Parallel text processing. Alignment and use of translation corpora*, vol. 13: Kluwer Academic Publishers, 2000, pp. 3-17.

[4] I. D. Melamed, "Pattern recognition for mapping bitext correspondence," in *Parallel text processing. Alignment and use of translation corpora*, vol. 13: Kluwer Academic Publishers, 2000, pp. 25-47.

[5] M. Simard, G.F. Foster, P. Isabelle, "Using Cognates to Align Sentences in Bilingual Corpora " in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992, pp. 67-81.

[6] F. Debili, E. Sammouda, "Appariement des Phrases de Textes Bilingues," in *The 14th International Conference on Computational Linguistics*, France, 1992, pp. 517-538.

[7] M. Haruno, T. Yamazaki, "High-performance bilingual text alignment using statistical and dictionary information," *Journal of Natural Language Engineering*, pp. 1-14, 1997.

[8] C. Cardenas, "Translating Text Using Bursty Sequences ": University of California Riverside, 2005.

[9] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," in *Computational Linguistics*, 1997, pp. 377-404.

[10] A. Gispert, D. Gupta, M. Popovic, P. Lambert, J. B. Mariño, M. Federico, H. Ney, R. E. Banchs, "Improving Statistical Word Alignments with Morpho-syntactic Transformations," in *Proc. of 5th Int. Conf. on Natural Language Processing (FinTAL)*, Turku (Finland), 2006, pp. 368-379.

[11] S. Longchupole, "Thai Syntactical Analysis system by Method of Splitting Sentences from Paragraph for Machine Translation," in *King Mongkut 's institute of technology Ladkrabang*, 1995.

[12] P. Mittrapiyanuruk, V. Sornlertlumvanich, "The Automatic Thai Sentence Extraction," in *The 4th Symposium on Natural Language Processing 2000* Thailand, 2000.

[13] S. Meknavin, P. Charoenpornsawat, B. kijsirikul, "Feature-based Thai Word Segmentation," in *Proceedings of Natural Language Processing Pacific Rim Symposium*, 1997, pp. 41-46.

[14] N. Tannin, K. Chancharoen, B. Sirinaowakul, "Alignment for Thai-English Sentence," in *The 1998 IEEE Asia-Pacific Conference on Circuits and Systems*, Thailand, 1998.

[15] K. Limpirojanarit, J. Pattayakorn, J. Kiatsompob, "A sentence alignment tool for English-Thai parallel corpus," Chulalongkorn University, 2006.

[16] "KTAGGER", SIL International, 2000, A part-of-speech tagger based on PC-KIMMO , Available at: http://www.sil.org/computIng/catalog/show_software.asp?id=22.

[17] "Smart Word Analysis for Thai (SWATH)", National Electronics and Computer Technology Center (NECTEC), Available at: http://www.links.nectec.or.th/download.php.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวศิรินันท์ สินธุวาทิน เกิดเมื่อวันที่ 4 มกราคม พ.ศ. 2526 ที่จังหวัดเชียงใหม่ สำเร็จการศึกษาหลักสูตรวิศวกรรมศาสตรบัณฑิต (วศ.บ.) สาขาวิชาคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่ เมื่อปีการศึกษา 2547 และเข้าศึกษาต่อหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2549 ขณะศึกษาได้มีโอกาสนำเสนอผลงานเรื่อง Parallel Text Alignment Using Bursty Sequences ในงานประชุมวิชาการนานาชาติ (The 2nd International Conference on Advances in Information Technology 2007) ปัจจุบันทำงานอยู่ที่ บริษัทรอยเตอร์ซอฟแวร์ไทยแลนด์ ตำแหน่งซอฟต์แวร์เอนจิเนียร์