การสกัดวลีสำคัญภาษาไทยอัตโนมัติโดยใช้เทคนิคเอ็นแกรมร่วมกับการเรียนรู้ของเครื่องจักร

Automatic Thai Keyphrase Extraction using N-Gram and Machine Learning Approach

ดร. แกมกาญจน์ สมประเสริฐศรี[1] *Dr. Gamgarn Somprasertsri*[1]

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อออกแบบกระบวนการสกัดวลีสำคัญภาษาไทยอัตโนมัติโดยใช้เทคนิคเอ็นแกรมร่วมกับการเรียนรู้ของเครื่องจักร และเปรียบเทียบประสิทธิภาพในการสกัดวลีสำคัญอัตโนมัติ ระหว่างการใช้เทคนิคเอ็นแกรมร่วมกับโครงข่ายประสาทเทียมและเทคนิคเอ็นแกรมร่วมกับตัวจำแนกเบย์อย่างง่าย ข้อมูลที่ใช้ในการทดลองคือ เอกสารบทความวิชาการ กระบวนการที่นำเสนอนี้ประกอบด้วยงานหลัก 2 ส่วน คือ 1) การสกัดวลีที่คาดว่าเป็นวลีสำคัญโดยใช้แกรมตั้งแต่ 1 ถึง 3-แกรมที่มีความถี่มากกว่า 2 และ 2) การฝึกฝนแบบจำลองโดยใช้ข้อมูลในการสอน คือ ความถี่ของวลี ตำแหน่งที่ปรากฏวลี ค่าความถี่ผกผันของเอกสาร และความยาวของวลี ผลการทดลองพบว่าการใช้เทคนิคเอ็นแกรมร่วมกับโครงข่ายประสาทเทียมในการจำแนกวลีสำคัญมีประสิทธิภาพมากกว่าการใช้เทคนิคเอ็นแกรมร่วมกับตัวจำแนกเบย์อย่างง่าย และการทดลองสกัดวลีสำคัญจากชุดข้อมูลทดสอบมีความถูกต้องร้อยละ 97.31 แสดงให้เห็นว่ากระบวนการที่นำเสนอสามารถสกัดวลีสำคัญภาษาไทยจากเอกสารบทความวิชาการแบบอัตโนมัติได้อย่างมีประสิทธิภาพ ซึ่งจะมีประโยชน์ต่องานด้านการค้นคืนสารสนเทศต่อไป

[1] ผู้ช่วยศาสตราจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
Assistant Professor, Faculty of Informatics, Mahasarakham University

รับต้นฉบับ 12 มิถุนายน 2557 ปรับแก้ไขตามข้อเสนอแนะของผู้ทรงคุณวุฒิ 25 สิงหาคม 2557
รับลงตีพิมพ์ 26 สิงหาคม 2557

Abstract

This study aims to design method for automatic keyphrases extraction of academic paper using n-gram and machine learning approaches including Neural Network and Naïve Bayes and to test the performance of the keyphrases extraction model. Experimental data was Thai academic papers. This method consists of two major steps; first to extracting the keyphrase candidates with were 1 to 3-gram and more than two frequencies. Second, train the model by using information including the phrase frequency, phrase position, inverse document frequency, and phrase length. The results showed that n-gram with Neural Network can extract keyphrases more than n-gram with Naïve Bayes effectively. Furthermore, the data testing was highly accurate (97.31%). Evaluation results show that this proposed method was highly effective for automatic keyphrases extraction of academic paper and being useful for information retrieval.

คำสำคัญ: การสกัดวลีสำคัญภาษาไทย เอ็นแกรม การเรียนรู้ของเครื่องจักร
Keywords: Thai keyphrase extraction, N-gram, Machine learning

บทนำ

ปัจจุบันสารสนเทศมีจำนวนเพิ่มขึ้นอย่างรวดเร็ว จึงเกิดมีระบบค้นคืนสารสนเทศหรือระบบค้นคืนเอกสารที่เป็นการรวบรวมและจัดเก็บสารสนเทศอย่างเป็นระบบ เพื่อให้ผู้ใช้สามารถเข้าถึงสารสนเทศที่ต้องการได้อย่างสะดวกและรวดเร็ว โดยเปรียบเทียบคำค้นที่ผู้ใช้ป้อนเข้ามาในระบบกับตัวแทนของเอกสารที่มีในระบบ คำสำคัญมักถูกใช้เป็นตัวแทนเอกสารในระบบค้นคืนสารสนเทศ ดังนั้นคำสำคัญจึงมีความสำคัญอย่างมากในระบบค้นคืนสารสนเทศ นอกจากการค้นคืนสารสนเทศแล้วคำสำคัญยังสามารถนำไปประยุกต์ใช้ในงานด้านการสรุปย่อเนื้อหาเอกสาร การจัดหมวดหมู่เอกสารแบบอัตโนมัติ และการถามตอบแบบอัตโนมัติด้วย การสกัดคำสำคัญที่เป็นตัวแทนของเอกสารที่กระทำโดยผู้เชี่ยวชาญพิจารณาเลือกจากเอกสารทั้งหมดที่มีจำนวนมากเป็นงานที่ต้องใช้เวลาและต้องมีความรู้เกี่ยวกับเอกสารนั้นๆ ดังนั้นจึงมีการพัฒนากระบวนการสกัดคำสำคัญแบบอัตโนมัติขึ้น

การระบุคำสำคัญที่สามารถเป็นตัวแทนเอกสารที่ดีในการทำดรรชนีคำค้น แบ่งได้เป็น 2 แนวทาง คือ การดึงคำหรือวลีจากเอกสารโดยตรง (Extraction Method) และการกำหนดคำหรือวลีจากเนื้อหาเอกสาร (Assignment Method) ทั้งสองแนวทางสามารถอาศัยหลักการของการเรียนรู้ของเครื่องจักรได้ (Boonchote, 2005) โดยทั่วไปคำสำคัญจะประกอบไปด้วยคำเดี่ยว (Single Word) หรือวลี (Phrase) การเลือกคำสำคัญที่เป็นตัวแทนของเอกสารจะส่งผลต่อประสิทธิภาพของการค้นคืนสารสนเทศ เมื่อเปรียบเทียบการใช้ดรรชนีระดับคำและดรรชนีระดับวลีในการค้นคืน พบว่าในกรณีที่ค้นคืนได้ในดรรชนีระดับวลีจะมีค่าความถูกต้องสูง (Chirarattanachan, 2000) งานวิจัยที่ผ่านมาพบว่ามีการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องจักร (Machine Learning) ในการสกัดวลีสำคัญ (Boonchote, 2005; Damrongson, 2003; Sarkar, Nasipuri, & Ghose, 2010; Turney, 2003)โดยงานของ Damrongson (2003) ได้นำเสนอวิธีการสกัดวลีสำคัญจากเอกสารภาษาอังกฤษแบบอัตโนมัติโดยใช้โครงข่ายประสาทเทียม สำหรับการสกัดวลีสำคัญภาษาไทยจากงานของ Boonchote (2005) เป็นการสกัดวลีสำคัญจากเอกสารบทความทางวิชาการฉบับเต็ม (Full Text) ที่จัดเก็บในรูปแบบอิเล็กทรอนิกส์โดยให้ค่าน้ำหนักคุณสมบัติของคำหรือวลีในการเรียนรู้โดยใช้โครงข่ายประสาทเทียมและเป็นการวิเคราะห์ในระดับหน่วยคำ

โดยทั่วไปพบว่าคำสำคัญที่เป็นตัวแทนเอกสารบทความทางวิชาการมีลักษณะเป็นกลุ่มคำหรือวลีมากกว่าที่จะเป็นคำเดี่ยว และจะถูกกำหนดโดยผู้เขียนซึ่งไม่มีกฎเกณฑ์หรือมาตราฐานสำหรับการกำหนดคำสำคัญแต่ขึ้นอยู่กับผู้เขียน ทำให้คำสำคัญเหล่านั้นอาจจะเป็นหรือไม่เป็นตัวแทนของเอกสารที่ดี อาจมีความผิดพลาดได้ เนื่องจากขาดความแม่นยำ เพราะขึ้นอยู่กับมาตราฐานของผู้เขียนแต่ละคน ซึ่งจะส่งผลต่อระบบการค้นคืนเอกสารนั้นๆ ได้ ในต่างประเทศจึงมีผู้ศึกษาวิจัยเกี่ยวกับการสกัดวลีสำคัญอัตโนมัติจากเอกสารบทความวิชาการหรือบทความวิจัยมากมาย (Kim, Medelyan, Kan, & Baldwin, 2013) แต่การสกัดวลีสำคัญอัตโนมัติภาษาไทยยังมีการศึกษาจำนวนไม่มากเนื่องจากการสกัดคำสำคัญในระดับวลีจะมีความซับซ้อนมากกว่าการสกัดในระดับหน่วยคำ โดยเฉพาะภาษาไทยที่แตกต่างจากภาษาต่างประเทศที่เขียนคำติดกันไม่มีการเว้นช่องว่างระหว่างคำเหมือนภาษาอังกฤษ ส่งผลให้การวิเคราะห์วลีมีปัญหาในเรื่องขอบเขตของวลีที่ไม่ชัดเจน ซึ่งจะส่งผลต่อความถูกต้องของการสกัดวลีสำคัญ การใช้การตัดคำที่ช่วยระบุขอบเขตของวลีที่ผ่านมาอาจจะยังไม่เพียงพอต่อการสกัดวลีสำคัญซึ่งต้องการการวิเคราะห์วลีที่มีความ

ซับซ้อนขึ้น งานวิจัยของ TeCho, Nattee, & Theeramunkong (2008) ได้นำเสนอวิธีการสกัดคำสำคัญโดยใช้ Thai Character Cluster (TCC) ซึ่งเป็นการจัดกลุ่มอักขระภาษาไทยโดยอาศัยระบบการเขียนไทยในการสกัดคำที่คาดว่าเป็นคำสำคัญและใช้การเรียนรู้ของเครื่องจักรในการจำแนกว่าเป็นคำสำคัญ การใช้ Thai Character Cluster (TCC) อาจมีข้อจำกัดในเรื่องเวลาที่ต้องใช้ในการประมวลผลมากกว่าเทคนิคเอ็นแกรม (n-gram) และยังต้องอาศัยความรู้ทางภาษาแต่เทคนิคเอ็นแกรมไม่ต้องอาศัยความรู้ทางด้านภาษา ดังนั้นปัญหาการสกัดวลีสำคัญภาษาไทยจึงนับเป็นปัญหาที่ยังคงต้องการการศึกษาวิจัยต่อไป งานวิจัยนี้จึงมุ่งศึกษาออกแบบกระบวนการสกัดวลีสำคัญภาษาไทยแบบอัตโนมัติจากเอกสารบทความทางวิชาการโดยใช้เทคนิคเอ็นแกรม มาช่วยในการวิเคราะห์คำในระดับวลีและประยุกต์ใช้การเรียนรู้ของเครื่องจักรเพื่อช่วยให้การสกัดวลีสำคัญภาษาไทยมีประสิทธิภาพมากขึ้น

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. การสกัดวลีสำคัญอัตโนมัติ

คำสำคัญ (Keyword) คือ คำที่ถูกกำหนดขึ้นเพื่อเป็นตัวแทนเนื้อหาสารสนเทศ โดยมาจากคำที่ปรากฏในหนังสือ บทความ หรือทรัพยากรสารสนเทศนั้นๆ อาจเป็นคำที่ปรากฏในชื่อเรื่อง บทคัดย่อ สาระสังเขป หรือตัวเนื้อหาของทรัพยากรสารสนเทศนั้นๆ โดยย่อเหลือเพียงคำที่แสดงใจความสำคัญของเนื้อเรื่อง ที่สั้น กะทัดรัด และมีความชัดเจน รวมถึงเป็นคำที่รู้จักกันเป็นสากลสำหรับผู้ค้นหาสารสนเทศ ทั้งนี้คำสำคัญจัดเป็นประเภทของคำที่ใช้แทนเนื้อหาสารสนเทศประเภทหนึ่ง นอกเหนือจากคำศัพท์สัมพันธ์ (Thesaurus) และ หัวเรื่อง (Subject heading) คำสำคัญถูกกำหนดขึ้นเพื่อเป็นเครื่องมือในการจัดเก็บและค้นคืนสารสนเทศ สำหรับเข้าถึงทรัพยากรสารสนเทศ การกำหนดคำสำคัญนั้นอาจทำโดยมนุษย์ หรือโปรแกรมคอมพิวเตอร์ โดยการดึงคำที่ปรากฏในทรัพยากรสารสนเทศซึ่งมีความถี่สูง ซึ่งไม่รวมคำที่ไม่สำคัญ เช่น คำบุพบท คำสันธาน คำวิเศษณ์ เป็นต้น เพื่อให้เหลือเพียงคำที่จะมาเป็นตัวแทนเนื้อหาของทรัพยากรสารสนเทศ (School of Liberal Arts, 2012) โดยทั่วไปคำสำคัญของเอกสารบทความวิชาการจะให้ผู้เขียนกำหนดคำสำคัญไว้ในเอกสาร คำสำคัญเหล่านั้นมักเป็นกลุ่มคำหรือวลีที่ประกอบด้วยคำตั้งแต่สองคำขึ้นไปมากกว่าเป็นคำเดี่ยว ดังนั้นจึงเรียกได้ว่าเป็นวลี

สำคัญ (Keyphrase) วิธีการสกัดคำหรือวลีสำคัญโดยอัตโนมัติที่มีใช้อยู่ปัจจุบันสามารถแบ่งออกเป็น 3 วิธีหลักๆ (Sittichoksataporn, 2012) ได้แก่ วิธีทางสถิติอย่างง่าย (Simple Statistic Approaches) วิธีทางภาษาศาสตร์ (Linguistics Approaches) และวิธีทางการเรียนรู้ของเครื่องจักร (Machine Learning Approaches) จากงานวิจัยที่ผ่านมาการสกัดวลีสำคัญอัตโนมัติโดยใช้วิธีทางการเรียนรู้ของเครื่องจักร สามารถแบ่งงานหลักๆ ได้เป็น 4 ส่วน คือ (Kim, Medelyan, Kan, & Baldwin, 2013)

1) การระบุวลีที่คาดว่าเป็นวลีสำคัญ เป็นส่วนเริ่มต้นของกระบวนการสกัดวลีสำคัญอัตโนมัติ เพื่อตรวจสอบหาวลีทั้งหมดที่คาดว่าเป็นวลีสำคัญ ซึ่งส่วนใหญ่จะอยู่ในรูปแบบคำนามหรือนามวลีที่ปรากฏอยู่ในเอกสาร วิธีการระบุอาจอาศัยหลักการกำกับหน้าที่คำ (POS Tag) ใช้เทคนิคเอ็นแกรม (n-grams) หรือใช้กลุ่มนามวลี (NP-chunks)

2) การกำหนดคุณสมบัติวลีสำคัญ ขั้นตอนนี้เกี่ยวกับการกำหนดและสร้างคุณสมบัติของวลีที่เป็นวลีสำคัญ ว่ามีคุณสมบัติอย่างไรบ้าง เพื่อเป็นข้อมูลที่นำไปใช้ในการเรียนรู้หรือสอนในขั้นตอนการสร้างแบบจำลอง

3) การสร้างแบบจำลอง เป็นขั้นตอนในการสร้างแบบจำลองที่จะใช้สกัดวลีสำคัญโดยแบบจำลองที่สร้างนั้นจะทำการเรียนรู้จากข้อมูลตัวอย่างจากขั้นตอนที่ผ่านมา ปัญหาการสกัดวลีสำคัญสามารถมองได้เป็นปัญหาของการจำแนกประเภทข้อมูล กล่าวคือแบบจำลองที่สร้างขึ้นสามารถจำแนกได้ว่าวลีไหนเป็นวลีสำคัญหรือเป็นวลีที่ไม่สำคัญ

4) การประเมินผลการสกัดวลีสำคัญ หลังจากสร้างแบบจำลองเรียบร้อยแล้วจะนำแบบจำลองที่สร้างขึ้นไปสกัดวลีสำคัญจากเอกสารใหม่ จึงจำเป็นต้องมีการวัดประสิทธิภาพของแบบจำลองในการสกัดวลีสำคัญ

2. เอ็นแกรม

เอ็นแกรมเป็นเทคนิคหนึ่งที่นิยมนำมาใช้ในการสร้างสายอักขระย่อยในงานด้านการประมวลผลภาษาธรรมชาติ โดยแกรม (gram) คือ หน่วยย่อยที่อาจจะเป็นเสียง คำ หรือ อักขระก็ได้ ซึ่งแกรมมีได้หลายขนาดแล้วแต่จะกำหนด ตั้งแต่ 1 จนถึง เอ็น (n) ถ้ากำหนดสายคำ $S= s_1 ,s_2 ,\ldots,s_N$ และคำ $s_i \in \{W_1 ,W_2 ,\ldots,W_m \}$ เมื่อ i = 1, 2, ...,N กำหนด N คือ ความยาวของสายคำ S และ m คือ จำนวนของคำ ดังนั้นรูปแบบ เอ็นแกรมของสายคำ S คือ สายคำย่อยของคำ s_i ที่เรียงต่อกัน ได้ตั้งแต่ 1-แกรม 2-แกรม หรือ 3-แกรม ไปเรื่อยๆ จนถึงเอ็น-แกรม จากงานวิจัยที่ผ่านมา Salton & Buckley (1988) พบว่า

การเลือกใช้เทคนิคเอ็นแกรมในลักษณะของไบแกรม (Bigrams) หรือ 2-แกรม และไตรแกรม (Trigrams) หรือ 3-แกรม เป็นขนาดที่เหมาะสมที่สุดสำหรับการค้นคืน ส่วนการใช้ยูนิแกรม (Unigrams) หรือ 1-แกรม จะให้ผลการค้นคืนมากเกินไปและการกำหนด n มากกว่า 4 จะส่งผลให้เกิดความผิดพลาดได้มากกว่าขนาด 2-แกรม หรือ 3-แกรม งานวิจัยนี้นำเทคนิคเอ็นแกรมมาใช้สำหรับสกัดวลีที่คาดว่าเป็นวลีสำคัญโดยใช้ n เท่ากับ 3

3. การเรียนรู้ของเครื่องจักร

งานวิจัยนี้ใช้การเรียนรู้ของเครื่องจักรแบบมีการสอน ได้แก่ โครงข่ายประสาทเทียม และตัวจำแนกเบย์อย่างง่าย โครงข่ายประสาทเทียมที่นำมาใช้เป็นโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer perceptron) และใช้อัลกอริทึมการแพร่กระจายย้อนกลับ (Back propagation algorithm) ซึ่งเป็นอัลกอริทึมที่นิยมใช้วิธีหนึ่งในการเรียนรู้ของโครงข่ายประสาทเทียมแบบหลายชั้น สำหรับปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนดให้เหมาะสม เพื่อหาค่าต่ำสุดของค่าผิดพลาดระหว่างเอาต์พุตของโครงข่ายที่คำนวณได้กับเอาต์พุตเป้าหมาย สำหรับตัวจำแนกเบย์อย่างง่ายเป็นวิธีจำแนกประเภทข้อมูลที่มีพื้นฐานมาจากทฤษฎีของเบย์ โดยการสร้างแบบจำลองจะอยู่ในรูปแบบความน่าจะเป็น ตัวจำแนกเบย์อย่างง่ายมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนและเหมาะกับกรณีที่ข้อมูลเซตตัวอย่างมีจำนวนมากและมีคุณสมบัติไม่ขึ้นต่อกัน จึงมักนิยมนำไปประยุกต์ใช้งานด้านการจำแนกประเภทข้อความ (Kijsirikul, 2003)

4. งานวิจัยที่เกี่ยวข้อง

จากงานวิจัยที่ผ่านมา พบว่า Damrongson (2003)ได้นำเสนอกระบวนการสกัดวลีสำคัญจากเอกสารภาษาอังกฤษ โดยพิจารณาคำที่เรียงติดกันไม่เกิน 3 คำ ที่ปรากฏมากกว่าหนึ่งครั้งในส่วนของเนื้อหาเอกสาร การกำหนดคุณสมบัติวลีในงานวิจัยนี้ใช้หลักการทางสถิติ เป็นการคำนวณค่าน้ำหนักแยกแต่ละส่วนของเอกสาร ได้แก่ ชื่อเรื่อง บทคัดย่อ เนื้อหา บทสรุป และเอกสารอ้างอิง ค่าที่คำนวณได้จะใช้เป็นข้อมูลนำเข้าในส่วนการสร้างแบบจำลองโครงข่ายประสาทเทียม แบบจำลองที่สร้างขึ้นสามารถเลือกวลีสำคัญของเอกสารได้ถูกต้องตรงกับที่ผู้เขียนกำหนดไว้โดยเฉลี่ยประมาณร้อยละ 40 สำหรับการสกัดวลีสำคัญภาษาไทย Boonchote (2005) ได้เสนอเทคนิคการสกัดวลีสำคัญภาษาไทยโดยใช้โครงข่ายประสาทเทียม โดยการให้

ค่าน้ำหนักคุณสมบัติของคำหรือวลีที่มีความสัมพันธ์กับไวยากรณ์และโครงสร้างทางภาษาศาสตร์ ข้อมูลค่าน้ำหนักคำหรือวลีในการเรียนรู้ได้แก่ ค่าน้ำหนักความถี่ของการเกิดขึ้น ค่าน้ำหนักตำแหน่งในประโยค ค่าน้ำหนักตำแหน่งในย่อหน้า ค่าน้ำหนักตำแหน่งในเอกสาร และค่าน้ำหนักหน้าที่คำ ผลการทดสอบสกัดวลีสำคัญสำหรับเอกสารชุดฝึกสอนและชุดทดสอบที่เป็นเอกสารเดียวมีความถูกต้องร้อยละ 90.26 สำหรับเอกสารชุดฝึกสอนและชุดทดสอบที่เป็นเอกสารโดเมนเดียวกันมีความถูกต้องร้อยละ 79.09 และสำหรับเอกสารชุดฝึกสอนและชุดทดสอบที่เป็นเอกสารคนละโดเมนมีความถูกต้องร้อยละ 82.80 นอกจากนี้ยังมีงานวิจัยของ TeCho, Nattee, & Theeramunkong (2008) ที่นำเสนอวิธีการสกัดคำสำคัญภาษาไทยโดยใช้ Thai Character Cluster (TCC) ร่วมกับการเรียนรู้ของเครื่องจักร วิธีการที่นำเสนอนี้จะเริ่มจากการสกัดคำที่คาดว่าเป็นคำสำคัญโดยใช้ Thai Character Cluster (TCC) ซึ่งเป็นการจัดกลุ่มอักขระภาษาไทย ในขั้นตอนนี้จะอาศัย LEXiTRON dictionary และ TCCs grammars หลังจากนั้นคำที่คาดว่าเป็นคำสำคัญจะถูกนำไปสกัดคุณสมบัติหรือข้อมูลเพื่อนำไปฝึกฝนแบบจำลองในการจำแนกคำสำคัญ โดยข้อมูลในการสอนที่ใช้ในงานวิจัยนี้มีทั้งหมด 7 ข้อมูล ได้แก่ จำนวนของ TCCs จำนวนของคำใน TCCs จำนวนของ TCC ที่แตกต่างกันทางด้านซ้ายมือและด้านขวามือ จำนวนของอักขระ ความถี่ ค่าความถี่ผกผันของเอกสาร และค่า TF- IDF สำหรับการเรียนรู้ของเครื่องจักรที่ใช้ในการทดลองได้แก่ ตัวจำแนกเบย์อย่างง่าย ตัวจำแนกที่ใช้หลักการพิจารณาจุดศูนย์กลาง (Centroid-based Classifier) และวิธีการหาสมาชิกที่ใกล้กันที่สุด (K-nearest Neighbor Algorithm) ผลการทดสอบสกัดคำสำคัญมีความถูกต้องร้อยละ 79.50 จากการใช้วิธีการหาสมาชิกที่ใกล้กันที่สุด

วิธีการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบกระบวนการสกัดวลีสำคัญภาษาไทยอัตโนมัติและเปรียบเทียบประสิทธิภาพการสกัดวลีสำคัญตามกระบวนการที่ได้ออกแบบ ขั้นตอนการดำเนินงานวิจัยแสดงรายละเอียดดังนี้

1. การเตรียมข้อมูล

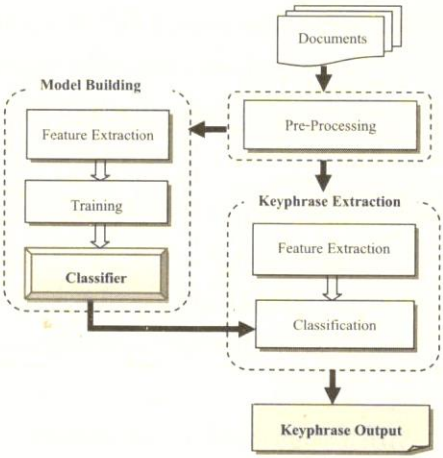
ข้อมูลที่ใช้ในงานวิจัยนี้ แบ่งเป็น 2 ชุด ดังนี้

1. ข้อมูลชุดที่ใช้สร้างแบบจำลองและทดสอบ เพื่อเปรียบเทียบประสิทธิภาพระหว่างการใช้เทคนิคเอ็นแกรมกับโครงข่ายประสาทเทียมและการใช้เทคนิคเอ็นแกรมกับตัวจำแนกเบย์อย่างง่าย ซึ่งได้แก่ เอกสารบทความวิชาการจากวารสารสารสนเทศศาสตร์และวารสารวิจัยสมาคมห้องสมุดแห่งประเทศไทย จำนวน 6 บทความที่มีการกำหนดคำสำคัญโดยผู้แต่ง ข้อมูลที่นำมาใช้สำหรับสกัดวลีสำคัญแบบอัตโนมัติ ได้แก่ ชื่อเรื่อง บทคัดย่อ บทนำ และบทสรุป เนื่องจากวลีสำคัญมักถูกพิจารณามาจากส่วนประกอบดังกล่าว การเตรียมข้อมูลในส่วนนี้จะทำการกำจัดคำภาษาอังกฤษ รูปภาพ ตารางและการอ้างอิงเนื้อหาทิ้ง ชุดข้อมูลประกอบด้วยคำทั้งหมดจำนวน 8,404 คำ และวลีสำคัญจำนวน 22 วลี

2. ข้อมูลชุดทดสอบเพื่อวัดประสิทธิภาพของกระบวนการสกัดวลีสำคัญอัตโนมัติที่นำเสนอในงานวิจัยนี้ เป็นข้อมูลชุดทดสอบที่ใช้วัดประสิทธิภาพการสกัดวลีสำคัญอัตโนมัติในงานวิจัยของ Boonchote (2005) ซึ่งเป็นบทความวิชาการจากวารสารวิศวสารลาดกระบัง ข้อมูลที่นำมาใช้ได้แก่ ชื่อเรื่อง บทคัดย่อ บทนำ และบทสรุป ซึ่งประกอบด้วยคำทั้งหมด 422 คำ และมีวลีสำคัญจำนวน 2 วลี

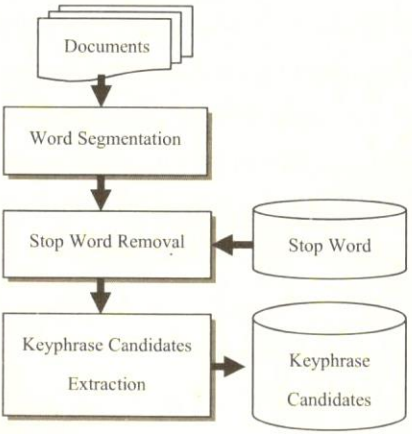
2. การสร้างแบบจำลองสกัดวลีสำคัญอัตโนมัติ

การสกัดวลีสำคัญอัตโนมัติ คือการพิจารณาวลีที่ปรากฏอยู่ในเอกสารส่วนของ ชื่อเรื่อง บทคัดย่อ บทนำ และบทสรุป โดยสร้างแบบจำลองเพื่อจำแนกว่าวลีใดเป็นวลีสำคัญของเอกสาร หลักการทำงานประกอบไปด้วยงานย่อย 3 ส่วน แสดงดังรูปที่ 1



รูปที่ 1 กระบวนการสกัดวลีสำคัญแบบอัตโนมัติ

1. ส่วนการประมวลผลเบื้องต้น (Pre-Processing) ในส่วนนี้จะทำหน้าที่ในการประมวลผลข้อมูลก่อนการสร้างแบบจำลอง และการสกัดวลีสำคัญ ประกอบไปด้วย การตัดคำ การกำจัดคำหยุด และการสกัดวลีที่คาดว่าเป็นวลีสำคัญ ดังดังรูปที่ 2



รูปที่ 2 ส่วนการประมวลผลเบื้องต้น

1.1 การตัดคำ (Word segmentation) เนื่องจากภาษาไทยมีการเขียนโดยไม่เว้นช่องว่างระหว่างคำ การตัดคำจึงเป็นขั้นตอนแรกของส่วนการประมวลผลเบื้องต้น งานวิจัยนี้เลือกใช้โปรแกรม LexTo (2012) ตัวอย่างผลลัพธ์การตัดคำที่ได้แสดงดังรูปที่ 3

งานวิจัย|นี้|มี|วัตถุประสงค์|เพื่อ|ออกแบบ|และ|พัฒนา|กระบวนการ|และ|ระบบ|นำทาง| ความรู้| |และ|ประเมิน| |ประสิทธิภาพ|ของ|ระบบ|นำทาง|ความรู้|เพื่อ|การ|เข้าถึง|เนื้อหา| ใน|ทรัพยากร|สารสนเทศ|ประเภท|สื่อสิ่งพิมพ์| |โดย|ใช้| |แหล่งข้อมูล|จาก|รายการ| สารบัญ|และ|ดรรชนี|ท้าย|เล่ม|ของ|สื่อสิ่งพิมพ์|เป็นตัวแทน|ของ|ประเด็น|เนื้อหา|

รูปที่ 3 ตัวอย่างผลลัพธ์จากการตัดคำ

1.2 การกำจัดคำหยุด (Stop word removal) ข้อมูลที่ผ่านการตัดคำจากขั้นตอนที่แล้วจะถูกนำมากำจัดคำหยุดซึ่งเป็นคำที่ไม่มีนัยสำคัญ คำหยุดที่ใช้ในงานวิจัยนี้ปรับปรุงมาจากงานวิจัยของ Kesorn (2010) โดยเลือกใช้เฉพาะคำหยุดที่ไม่มีนัยสำคัญในงานวิจัยนี้

1.3 การสกัดวลีที่คาดว่าเป็นวลีสำคัญ (Keyphrase candidates extraction) เป็นการวิเคราะห์ข้อมูลในระดับวลีเพื่อสกัดวลีที่คาดว่าเป็นวลีสำคัญ โดยนำเทคนิคเอ็นแกรมมาใช้ ซึ่งกำหนดใช้ค่าเอ็นเท่ากับ 3 ในขั้นตอนนี้ในแต่ละประโยคที่ผ่านการตัดคำและกำจัดคำหยุดแล้วจะสกัดข้อมูลตั้งแต่ 1-แกรม 2-แกรม และ 3-แกรม จากข้อมูลตัวอย่างแกรมที่สกัดได้แสดงดังตารางที่ 1 การสกัดวลีที่คาดว่าเป็นวลีสำคัญจากแกรมที่ได้จะพิจารณาจากความถี่และทำการตัดแกรมที่น้อยกว่าทิ้งไป ตัวอย่างเช่น *"ระบบนำทางความรู้"* มีความถี่มากกว่าเกณฑ์ที่กำหนดจะถือว่าเป็นวลีที่คาดว่าเป็นวลีสำคัญ ดังนั้น *"ระบบ"* และ *"ระบบนำทาง"* จะถูกตัดทิ้งไป ไม่ถูกสกัดเป็นวลีที่คาดว่าเป็นวลีสำคัญ

ตารางที่ 1 ตัวอย่างข้อมูล 1-แกรม 2-แกรม และ 3-แกรม ของวลีที่คาดเป็นวลีสำคัญ

1-แกรม	2-แกรม	3-แกรม
งานวิจัย	งานวิจัย วัตถุประสงค์	งานวิจัย วัตถุประสงค์ ออกแบบ
วัตถุประสงค์	วัตถุประสงค์ ออกแบบ	วัตถุประสงค์ ออกแบบ พัฒนา
ออกแบบ	ออกแบบ พัฒนา	ออกแบบ พัฒนา กระบวนการ
พัฒนา	พัฒนา กระบวนการ	พัฒนา กระบวนการ ระบบ
กระบวนการ	กระบวนการ ระบบ	กระบวนการ ระบบ นำทาง
ระบบ	ระบบ นำทาง	ระบบ นำทาง ความรู้
นำทาง	นำทาง ความรู้	
ความรู้		

2. ส่วนการสร้างแบบจำลอง (Model building) เป็นขั้นตอนในการสร้างโมเดลหรือแบบจำลองที่จะใช้เป็นตัวจำแนก (Classifier) ซึ่งสามารถจำแนกว่าวลีที่คาดว่าเป็นวลีสำคัญจะเป็นวลีสำคัญหรือไม่เป็นวลีสำคัญ โดยทำการเปรียบเทียบข้อมูลนั้นกับคุณสมบัติของต้นแบบวลีที่เป็นวลีสำคัญและวลีที่ไม่เป็นวลีสำคัญ รายละเอียดขั้นตอนมีดังนี้

2.1 การสกัดคุณสมบัติ (Feature extraction) เป็นขั้นตอนในการสกัดคุณสมบัติที่จะเป็นข้อมูลสำหรับการเรียนรู้ที่จะใช้สอนตัวจำแนกเพื่อทำนายว่าวลีที่คาดว่าเป็นวลีสำคัญจะเป็นวลีสำคัญหรือไม่เป็นวลีสำคัญ สำหรับการสกัดคุณสมบัตินี้ในส่วนของการสกัดวลีสำคัญจำเป็นต้องมีการสกัดคุณสมบัติเช่นกัน แต่เป็นการสกัดคุณสมบัติจากวลีที่ต้องการจำแนกว่าเป็นวลีสำคัญหรือไม่เป็นวลีสำคัญ แต่การสกัดคุณสมบัติในส่วนการสร้างแบบจำลองนี้จะสกัดจากวลีที่ทราบแล้วว่าเป็นวลีสำคัญและวลีที่ไม่เป็นวลีสำคัญ สำหรับคุณสมบัติที่ใช้ในงานวิจัยนี้แสดงดังตารางที่ 2 ซึ่งประกอบด้วย ความถี่ของวลีที่ปรากฏในเอกสาร ตำแหน่งที่ปรากฏของวลีในเอกสาร ความยาวของวลี และค่าความถี่ผกผันของเอกสาร (Inverse Document Frequency: IDF) ส่วนตารางที่ 3 แสดงตัวอย่างข้อมูลคุณสมบัติที่สกัดได้ในส่วนการสร้างแบบจำลอง

ตารางที่ 2 คุณสมบัติที่ใช้ในงานวิจัย

คุณสมบัติ	คำอธิบาย
ความถี่	จำนวนครั้งของวลีที่ปรากฏในเอกสาร
ชื่อเรื่อง	วลีปรากฏในตำแหน่งที่เป็นชื่อเรื่อง
บทคัดย่อ	วลีปรากฏในตำแหน่งที่เป็นบทคัดย่อ
บทนำ	วลีปรากฏในตำแหน่งที่เป็นบทนำ
บทสรุป	วลีปรากฏในตำแหน่งที่เป็นบทสรุป
ความยาว	ความยาวของวลี
IDF	Inverse document frequency
เป้าหมาย	เป็นวลีสำคัญหรือไม่เป็นวลีสำคัญ

ตารางที่ 3 ตัวอย่างผลลัพธ์ที่ได้จากการสกัดคุณสมบัติ

ความถี่	ตำแหน่งที่ปรากฏของวลี				ความยาว	IDF	เป้าหมาย
	ชื่อเรื่อง	บทคัดย่อ	บทนำ	บทสรุป			
5	1	1	0	0	3	1.585	0
7	1	1	0	0	3	1.585	1
2	0	1	0	0	2	2.585	0
12	0	1	1	1	3	1	0
3	1	1	0	0	2	1.585	1
6	0	1	1	0	1	1.585	0
5	0	1	0	1	1	1.585	0

2.2 การฝึกฝน (Training) ในขั้นตอนนี้เป็นการฝึกฝนเพื่อสร้างแบบจำลองจากข้อมูลคุณสมบัติที่สกัดได้จากขั้นตอนที่ผ่านมา งานวิจัยนี้ทำการสร้างแบบจำลองโครงข่ายประสาทเทียมและตัวจำแนกเบย์อย่างง่ายสำหรับการสกัดวลีสำคัญภาษาไทยจากเอกสารบทความวิชาการ

3. ส่วนการสกัดวลีสำคัญ (Keyphrase extraction) ในส่วนนี้เป็นการสกัดวลีสำคัญภาษาไทยจากเอกสารบทความวิชาการ โดยการใช้แบบจำลองที่ได้จากการฝึกฝนในส่วนของการสร้างแบบจำลอง เป็นตัวจำแนกว่าวลีที่คาดว่าจะเป็นวลีที่กำลังพิจารณานั้นเป็นวลีสำคัญหรือเป็นวลีไม่สำคัญ ในส่วนนี้จำเป็นต้องสกัดคุณสมบัติเหมือนในส่วนของการสร้างแบบจำลองแต่ไม่มีการกำหนดเป้าหมายว่าวลีนั้นเป็นวลีสำคัญหรือไม่

3. การประเมินผลการสกัดวลีสำคัญ

ในขั้นตอนการดำเนินการวิจัยนี้ เป็นการประเมินผลการสกัดวลีสำคัญอัตโนมัติตามกระบวนการที่ได้ออกแบบโดย ใช้ตัววัดประสิทธิภาพ ได้แก่ ค่าความระลึก ค่าความแม่นยำ ค่าเอฟเมเซอร์ และค่าความถูกต้อง วิธีการทดสอบใช้การตรวจสอบไขว้ (Cross-Validation) ซึ่งเป็นวิธีการทดสอบที่จะทำการแบ่งชุดเอกสารออกเป็น k ชุด แล้วทำการทดสอบ โดยการทดสอบแต่ละครั้งจะใช้เอกสาร 1 ชุด เป็นชุดทดสอบ ส่วนที่เหลือ k-1 ชุด จะใช้เป็นชุดการสอน สับเปลี่ยนชุดทดสอบโดยใช้ทุกชุดเป็นชุดทดสอบจนครบ การทดสอบแต่ละครั้งถือเป็นการทดลองหนึ่งครั้ง ซึ่งเท่ากับได้ทำการทดลอง k ครั้ง แล้วนำผลที่วัดได้จากการทดลองทั้ง k ครั้งมาหาค่าเฉลี่ยของตัววัดประสิทธิภาพ โดยการวิจัยในครั้งนี้ใช้การทดสอบ k เท่ากับ 10 เพื่อเปรียบเทียบผลของการสกัดวลีสำคัญอัตโนมัติระหว่างการใช้โครงข่ายประสาทเทียมและตัวจำแนกเบย์อย่างง่าย แล้วเลือกแบบจำลองที่ดีที่สุดนำไปทดลองสกัดวลีอัตโนมัติจากข้อมูลชุดที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบในงานวิจัยของ Boonchote (2005) เพื่อวัดประสิทธิภาพของกระบวนการสกัดวลีสำคัญอัตโนมัติที่นำเสนอในงานวิจัยนี้ การวัดประสิทธิภาพในการสกัดวลีสำคัญภาษาไทยสามารถคำนวณค่าได้ดังนี้ (Baeza-Yates, & Ribeiro-Neto, 1999)

ค่าความระลึก (Recall: R) $= \frac{TP}{TP + FN}$

ค่าความแม่นยำ (Precision: P) $= \frac{TP + FP}{TP + FN}$

ค่าเอฟเมเซอร์ (F-measure) $= \frac{2\,PR}{R + P}$

ค่าความถูกต้อง (Accuracy) $= \frac{TP + TN}{FN + TN + FP}$

เมื่อ TP = จำนวนวลีที่เป็นวลีสำคัญและระบบจำแนกว่าเป็นวลีสำคัญ TN = จำนวนวลีที่ไม่เป็นวลีสำคัญและระบบจำแนกว่าไม่เป็นวลีสำคัญ FP = จำนวนวลีที่ไม่เป็นวลีสำคัญและระบบจำแนกว่าเป็นวลีสำคัญ FN = จำนวนวลีที่เป็นวลีสำคัญและระบบจำแนกว่าไม่เป็นวลีสำคัญ

ผลการวิจัยและอภิปราย

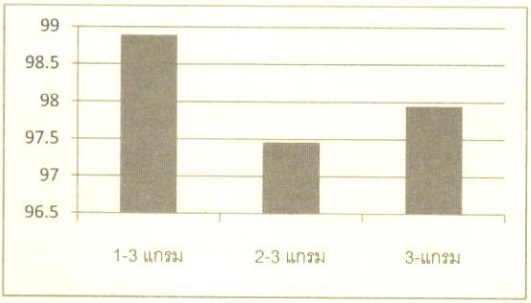
งานวิจัยนี้ได้แบ่งการทดลองออกเป็น 2 ส่วน คือ 1) การทดลองเพื่อเปรียบเทียบประสิทธิภาพของกระบวนการสกัดวลีอัตโนมัติที่ออกแบบโดยใช้เทคนิคเอ็นแกรมในการสกัดวลีที่คาดว่าเป็นวลีสำคัญและการจำแนกวลีสำคัญโดยใช้โครงข่ายประสาทเทียมและตัวจำแนกเบย์อย่างง่าย 2) การทดลองสกัดวลีอัตโนมัติจากชุดข้อมูลทดสอบจากงานวิจัยที่ผ่านมาของ Boonchote (2005) เพื่อวัดประสิทธิภาพของกระบวนการสกัดวลีสำคัญอัตโนมัติที่นำเสนอในงานวิจัยนี้ ผลการทดลองที่ 1 เปรียบเทียบประสิทธิภาพในการสกัดวลีสำคัญด้วยโครงข่ายประสาทเทียมกับตัวจำแนกเบย์อย่างง่าย โดยกำหนดค่าความถี่ที่แตกต่างกันสำหรับใช้เป็นเกณฑ์ในการสกัดวลีที่คาดว่าเป็นวลีสำคัญ แสดงดังตารางที่ 4

ตารางที่ 4 ประสิทธิภาพการสกัดวลีสำคัญของแบบจำลองเมื่อใช้ความถี่แตกต่างกัน

ค่าความถี่	แบบจำลอง	ค่าความแม่นยำ	ค่าความระลึก	ค่าเอฟเมเซอร์	ค่าความถูกต้อง (ร้อยละ)
> 1	โครงข่ายประสาทเทียม	0.83	0.56	0.67	98.71
	ตัวจำแนกเบย์อย่างง่าย	0.10	0.33	0.15	91.52
> 2	โครงข่ายประสาทเทียม	**0.78**	**0.78**	**0.78**	**98.89**
	ตัวจำแนกเบย์อย่างง่าย	0.11	0.33	0.17	91.67
> 3	โครงข่ายประสาทเทียม	0.78	0.78	0.78	98.51
	ตัวจำแนกเบย์อย่างง่าย	0.15	0.44	0.23	89.93

จากข้อมูลในตารางที่ 4 พบว่า ค่าความแม่นยำ ค่าความระลึก ค่าเอฟเมเซอร์และค่าความถูกต้องในการสกัดวลีสำคัญโดยใช้โครงข่ายประสาทเทียมมีประสิทธิภาพมากกว่าการใช้ตัวจำแนกเบย์อย่างง่าย ซึ่งสอดคล้องกับงานวิจัยของ Sarkar, Nasipuri, & Ghose (2010) ที่สกัดวลีสำคัญจากเอกสารบทความวิทยาศาสตร์ โดยใช้โครงข่ายประสาทเทียม และเมื่อเปรียบเทียบผลการทดลองระหว่างการใช้ค่าความถี่ที่แตกต่างกันสำหรับใช้เป็นเกณฑ์ในการสกัดวลีที่คาดว่าเป็นวลีสำคัญพบว่า การใช้ความถี่มากกว่า 2 สำหรับโครงข่ายประสาทเทียมมีประสิทธิภาพสูงกว่าการใช้ความถี่มากกว่า 1 และ 3 โดยให้ค่าความแม่นยำ ค่าความระลึก และค่าเอฟเมเซอร์ เท่ากับ 0.78 มีความถูกต้องร้อยละ 98.89 ส่วนตัวจำแนกเบย์อย่างง่ายพบว่า การใช้ความถี่มากกว่า 2 ให้ความถูกต้องมากกว่าการใช้ความถี่มากกว่า 1 และ 3 คือ ร้อยละ 91.67 แต่การใช้ความถี่มากกว่า 3 ให้ค่าความแม่นยำ ค่าความระลึก และค่าเอฟเมเซอร์สูงสุด งานวิจัยนี้จึงเลือกใช้โครงข่ายประสาทเทียมเป็นแบบจำลองในการสกัดวลีสำคัญ และกำหนดเกณฑ์ความถี่มากกว่า 2 ในการสกัดวลีที่คาดว่าเป็นวลีสำคัญ

การใช้เทคนิคเอ็นแกรมในการสกัดวลีที่คาดว่าเป็นวลีสำคัญ นอกจากการหาความถี่ที่เหมาะสมเพื่อกำหนดเป็นเกณฑ์การสกัดวลีที่คาดว่าเป็นวลีสำคัญแล้ว การเลือกจำนวนแกรมที่ใช้ให้เหมาะสมก็มีความสำคัญต่อการสกัดวลีสำคัญเช่นกัน ดังนั้นในการทดลองที่ 1 นี้ จึงได้ทดลองใช้จำนวนแกรมที่แตกต่างกันในการสกัดวลีสำคัญ ผลการทดลองแสดงดังรูปที่ 5 ซึ่งพบว่าการใช้แกรมตั้งแต่ 1-แกรม ถึง 3-แกรม ให้ค่าความถูกต้องในการสกัดวลีสำคัญสูงกว่าการเลือกใช้เฉพาะ 2-แกรม หรือ 3-แกรม



รูปที่ 5 ความถูกต้องในการสกัดวลีสำคัญโดยใช้จำนวนแกรมต่างกัน

สำหรับผลการทดลองที่ 2 การทดลองสกัดวลีสำคัญอัตโนมัติจากชุดข้อมูลทดสอบจากงานวิจัยที่ผ่านมาแสดงดังตารางที่ 5 สามารถคำนวณค่าความถูกต้องได้ ร้อยละ 97.31 ซึ่งได้ค่าความถูกต้องที่สูงกว่างานวิจัยที่ผ่านมา ทั้งนี้อาจเนื่องมาจากกระบวนการสกัดวลีสำคัญอัตโนมัติที่นำเสนอในงานวิจัยนี้ ได้นำเทคนิคเอ็นแกรมเข้ามาช่วยในการวิเคราะห์ในระดับวลีซึ่งเป็นวิธีที่เหมาะสมช่วยเพิ่ม True positive และลด False-positive

ตารางที่ 5 ผลการทดลองสกัดวลีสำคัญจากข้อมูลชุดทดลอง

	จำนวนวลีที่เป็นวลีสำคัญ	จำนวนวลีที่ไม่เป็นวลีสำคัญ
จำนวนวลีที่ระบบจำแนกว่าเป็นวลีสำคัญ	2	0
จำนวนวลีที่ระบบจำแนกว่าไม่เป็นวลีสำคัญ	5	179

สรุป

การวิจัยครั้งนี้เป็นการศึกษาเพื่อออกแบบกระบวนการสกัดวลีสำคัญภาษาไทยโดยใช้เทคนิคเอ็นแกรมและวิธีการเรียนรู้ของเครื่องจักร และวัดประสิทธิภาพของกระบวนการในการสกัดวลีสำคัญ กระบวนการสกัดวลีสำคัญที่นำเสนอนี้เริ่มจากการวิเคราะห์ข้อมูลในระดับวลีด้วยเทคนิคเอ็นแกรมในการสกัดวลีที่คาดว่าเป็นวลีสำคัญโดยใช้จำนวนแกรมตั้งแต่ 1-แกรม 2-แกรม และ 3-แกรม กำหนดเกณฑ์ความถี่ที่มากกว่า 2 และทำการตัดแกรมที่น้อยกว่าทิ้ง แล้วทำการฝึกฝนแบบจำลองโดยใช้ข้อมูลคุณสมบัติ คือ ความถี่ของวลี ตำแหน่งที่ปรากฏวลี (ชื่อเรื่อง บทคัดย่อ บทนำ บทสรุป) ค่าความถี่ผกผันของเอกสาร และความยาวของวลี แบบจำลองที่ได้จากการฝึกฝนโดยใช้วิธีการเรียนรู้ของเครื่องจักร จะถูกนำมาใช้จำแนกวลีที่คาดว่าเป็นวลีสำคัญ หรือเป็นวลีที่ไม่สำคัญ ผลการทดลองพบว่า การใช้โครงข่ายประสาทเทียมในการจำแนกวลีสำคัญมีประสิทธิภาพมากกว่าการใช้ตัวจำแนกเบย์อย่างง่าย โดยจำนวนแกรมที่เหมาะสมในการสกัดวลีที่คาดว่าเป็นวลีสำคัญคือ 1 ถึง 3 แกรม และมีความถี่มากกว่า 2 นอกจากนี้ ยังพบว่าเมื่อนำแบบจำลองที่สร้างขึ้นไปทดลองสกัดวลีสำคัญอัตโนมัติจากชุดข้อมูลทดสอบจากงานวิจัยที่ผ่านมาของ Boonchote (2005) ได้ค่าความถูกต้องที่สูงกว่า แสดงให้เห็นว่ากระบวนการที่นำเสนอโดยใช้เทคนิคเอ็นแกรมร่วมกับวิธีการเรียนรู้ของเครื่องจักรสามารถสกัดวลีสำคัญ

ภาษาไทยจากเอกสารบทความวิชาการแบบอัตโนมัติได้ ซึ่งกระบวนการที่นำเสนอในงานวิจัยนี้ยังสามารถนำไปประยุกต์ใช้กับการสกัดวลีสำคัญสำหรับภาษาอื่นได้ เนื่องจากเทคนิคเอ็นแกรมไม่จำเป็นต้องอาศัยความรู้ทางด้านภาษา นอกจากนี้ยังเป็นแนวทางในการนำไปประยุกต์กับการสกัดวลีสำคัญจากเอกสารรูปแบบอื่นๆ ได้ต่อไป

กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ให้ทุนสนับสนุนในการดำเนินการวิจัยนี้

เอกสารอ้างอิง

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). **Modern Information Retrieval**. New York: ACM Press.

Boonchote, K. (2005). **Thai keyphrases extraction using artificial neural network**. (In Thai). Master of Science in Management of Information Technology, Prince of Songkla University.

Chirarattanachan, C. (2000). **Thai noun phrase analysis for information retrieval system**. (In Thai). Master of Engineering in Computer Engineering, Kasetsart University.

Damrongson, T. (2003). **Automatic keyphrase extraction and its application in content based article retrieval**. (In Thai). Master of Engineering in Computer Engineering, Kasetsart University.

Kesorn, W. (2010). **Similarity measurement of Thai documents using natural language processing**. (In Thai). Master of Science in Computer Science, Chiang Mai University.

Kijsirikul, B. (2003). **Artificial intelligence**. (In Thai). Bangkok: Chulalongkorn University.

Kim, N.S., Medelyan, O., Kan, M.Y., & Baldwin, T. (2013). Automatic keyphrase extraction from scientific articles. **Language Resources and Evaluation,** 47, 723-742.

LexTo. Retrieved 21 June 2012, from http://www.sansarn.com/lexto/

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, 24(5), 513-523.

Sarkar, K., Nasipuri, M., & Ghose, S. (2010). A new approach to keyphrase extraction using neural networks. **International Journal Computer Science,** 7(3), 16-25.

School of Liberal Arts. (2012). **Information analysis**. (In Thai). 3d ed. Nonthaburi: Sukhothai Thammathirat Open University.

Sittichoksataporn, J. (2012). **Prototype of e-saraban ontology for semantic search: A case study administration office of Faculty of Medicine at Prince of Songkla University**. (In Thai). Master of Science in Management of Information Technology, Prince of Songkla University.

TeCho, J., Nattee, C., & Theeramunkong, T. (2008). A Corpus-based approach for keyword identification using supervised learning techniques. **Proceedings of the 5th International Conference on ECIT-CON**, 33-36.

Turney, P.D. (2003). Coherent keyphrase extraction via web mining. **Proceedings of the 18th International Joint Conference on Artificial Intelligence**, 434-439.