บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในสถานการณ์ของโลกยุคปัจจุบัน คอมพิวเตอร์ได้เข้ามามีบทบาท และส่งผลกระทบต่อวิถีการดำรงชีวิตของมนุษย์มากขึ้น ดังจะเห็นได้จากจำนวนของผู้ใช้งานคอมพิวเตอร์ที่สูงมากขึ้นเรื่อยๆ ซึ่งการเพิ่มขึ้นของจำนวนผู้ใช้งานดังกล่าวนั้น ก็เป็นสาเหตุหนึ่งของการเพิ่มขึ้นของสารสนเทศในรูปแบบต่างๆ มากมาย ทำให้การแลกเปลี่ยนข้อมูลของมนุษย์ในยุคสมัยนี้เกิดขึ้นกับกองสารสนเทศจำนวนมาก

ด้วยสาเหตุดังกล่าวจึงจำเป็นอย่างยิ่งที่จะต้องมีการพัฒนาระบบค้นคืนสารสนเทศ (Information Retrieval Systems) ขึ้นใช้งาน เพื่อสร้างระบบที่สามารถจัดการกับกองสารสนเทศเหล่านั้นได้อย่างมีประสิทธิภาพ ซึ่งในปัจจุบันก็มีระบบค้นคืนสารสนเทศให้เลือกใช้งานเป็นจำนวนมาก แต่จะแตกต่างกันตรงที่ ความสามารถที่จำเพาะเจาะจงลงไปในแต่ละระบบ ที่จะสร้างขึ้นให้เป็นจุดเด่นของระบบของตนเอง แต่ทุกระบบก็จะมุ่งเน้นไปที่สารสนเทศในหลากหลายรูปแบบ ที่ตัวระบบนั้นจะต้องมีความสามารถในการค้นคืนออกมาให้ได้มากที่สุด ทำให้ผลลัพธ์ที่ได้จากการค้นคืนของระบบนั้น พบว่ามีทั้งสารสนเทศที่เกี่ยวข้อง และที่ไม่เกี่ยวข้องกับความต้องการของผู้ใช้งานสารสนเทศนั้น

ปัจจุบันเอกสารทางวิชาการในประเทศไทย มักจะจัดทำทั้งในรูปแบบของภาษาไทย และในรูปแบบของภาษาอังกฤษ เพื่อประโยชน์ในการเผยแพร่ทั้งภายในและภายนอกประเทศ ซึ่งเอกสารเหล่านี้โดยเฉพาะอย่างยิ่งเอกสารทางด้านวิทยาศาสตร์และวิศวกรรมศาสตร์โดยมากแล้วมักจะปรากฏคำนามเฉพาะ (Proper Noun) และคำศัพท์เทคนิคต่างๆ เป็นจำนวนมาก ซึ่งจะพบได้ทั้งในรูปแบบของคำในภาษาอังกฤษ คำภาษาไทยทับศัพท์คำภาษาอังกฤษ คำภาษาอังกฤษทับศัพท์คำภาษาไทย หรือคำในภาษาไทยเอง ดังนั้นถ้าระบบค้นคืนสารสนเทศไม่สนับสนุนการทำงานข้ามภาษา ก็จะทำให้ประสิทธิภาพในการค้นคืนนั้นลดต่ำลง และเป็นการใช้ประโยชน์จากสารสนเทศที่มีอยู่ในเครือข่ายได้อย่างไม่เต็มที่

การค้นคืนสารสนเทศข้ามภาษา (Cross-Language Information Retrieval) หมายถึงการค้นคืนสารสนเทศซึ่งภาษาที่แสดงในเอกสารไม่ตรงกับภาษาที่แสดงในการสอบถาม [1]

ปัญหาในระบบค้นคืนสารสนเทศมีอยู่หลายประการด้วยกัน โดยเฉพาะอย่างยิ่งปัญหาในเรื่องของการค้นคืนข้ามภาษา ซึ่งพบว่าคำในภาษาหนึ่งอาจจะถูกเขียนในอีกภาษาหนึ่งได้หลาย

รูปแบบ ตัวอย่างเช่น "Carbohydrate" ในภาษาไทยอาจพบได้หลายแบบทั้ง "คาร์โบไฮเดรต" "คาร์โบไฮเดรท" หรือ "คาร์โบฮัยเดรต" หรือชื่อเฉพาะในภาษาไทย เช่น "โอภาส" อาจปรากฏในเอกสารที่ใช้เผยแพร่ในต่างประเทศในรูปของ "Ophas" หรือ "Opas" เป็นต้น ซึ่งระบบค้นคืนสารสนเทศควรจะมีความสามารถในการค้นคืนเอกสารที่มีคำเหล่านี้ปรากฏอยู่ออกมาให้ได้ทั้งหมดหรือให้ได้มากที่สุด และถึงแม้ว่าจะมีการนำพจนานุกรมสองภาษา (Bilingual Dictionary) มาใช้ในระบบค้นคืนสารสนเทศก็ไม่อาจจะแก้ปัญหานี้ได้มากนัก เนื่องจากมีคำศัพท์เทคนิคและคำทับศัพท์ใหม่ ๆ เกิดขึ้นมากมายในหลากหลายสาขาแทบทุกวัน และคำศัพท์ใหม่ ๆ เหล่านี้ส่วนมากมักจะไม่ปรากฏพบในพจนานุกรม [2] ดังนั้นจึงทำให้การสอบถามด้วยคำหลักในภาษาหนึ่ง อาจจะทำให้พลาดเอกสารที่มีคำหลักที่ตรงกันในอีกภาษาหนึ่งได้

ด้วยเหตุนี้ นักวิจัยจำนวนมากจึงได้หันมาสนใจกับปัญหาของการค้นคืนข้ามภาษา โดยมุ่งเน้นแก้ปัญหาด้วยการใช้การเข้ารหัสคำเพื่อนำมาช่วยในการค้นคืนข้ามภาษา โดยรหัสคำดังกล่าวนั้นจะเป็นสัญลักษณ์ที่แทนเสียงอ่านของคำ เพราะเนื่องจากคำที่อ่านออกเสียงเหมือนกันจะมีรหัสคำที่ตรงกันหรือใกล้เคียงกัน งานวิจัยต่างๆ ที่เกี่ยวข้องกับการเข้ารหัสคำและการค้นคืนข้ามภาษาไทย-ภาษาอังกฤษ ได้แก่ [3] [4] และ [5]

เนื่องจากรหัสคำที่ได้จากขั้นตอนของการเข้ารหัสคำนั้น อาจมีรหัสคำไม่เหมือนกันทุกตัวอักษร ถึงแม้ว่ารหัสคำนั้นจะเป็นรหัสคำที่อ่านออกเสียงตรงกันจากทั้งสองภาษา ดังนั้นในงานวิจัยดังกล่าวที่ผ่านมาข้างต้น จึงได้นำเสนอวิธีในการเปรียบเทียบรหัสคำ ซึ่งต้องใช้วิธีการเปรียบเทียบแบบประมาณ (Approximate Matching) ด้วยเทคนิคระยะแก้ไขสั้นสุด (Minimum Edit Distance) [6] ซึ่งเป็นการคำนวณความคล้ายคลึงกันระหว่างสายอักขระ 2 สาย โดยคำนวณจากต้นทุน (Cost) ที่ใช้ในการเปลี่ยนสายอักขระหนึ่งไปเป็นอีกสายอักขระหนึ่ง ด้วยการเพิ่ม ลบ หรือแทนที่อักขระ ซึ่งในขั้นตอนของการเปรียบเทียบรหัสคำนั้น ในงานวิจัยก่อนหน้าจะทำการแยกส่วนของการเปรียบเทียบออกเป็น 2 ส่วนจากกัน คือ ส่วนของพยัญชนะ และส่วนของสระ ตัวอย่างเช่น "kanokpiravut_" ซึ่งเป็นรหัสคำที่ได้จากการเข้ารหัสจากคำว่า "กนกพีระวุฒิ" ซึ่งในงานวิจัยก่อนหน้าจะทำการสลับตำแหน่งของตัวอักษรที่เป็นสัญลักษณ์แทนเสียงพยัญชนะ และเสียงสระออกจากกัน ด้วยการเลื่อนสัญลักษณ์ที่แทนเสียงของสระไปไว้ด้านหลังกลายเป็น "knkprvt, aoiau" แล้วแยกการเปรียบเทียบในแต่ละส่วนออกจากกัน

แต่ในงานวิจัยนี้ได้มีการปรับเปลี่ยนวิธีการดังกล่าว โดยจะทำการเข้ารหัสใหม่เพื่อแยกรหัสเสียงของเสียงพยัญชนะต้น เสียงสระ และเสียงของพยัญชนะที่เป็นตัวสะกดออกจากกัน เพื่อทำให้การเข้ารหัสคำนั้นสามารถแบ่งแยกได้อย่างชัดเจนว่าเป็นรหัสคำในส่วนไหน เพื่อให้การค้นคืนนั้นสามารถทำการเปรียบเทียบถูกต้องมากยิ่งขึ้น และการวิจัยนี้มีข้อสมมุติฐานว่าขั้นตอนที่

นำเสนอนั้น จะสามารถทำการสืบค้นคำทับศัพท์ข้ามภาษาไทย-ภาษาอังกฤษ ได้โดยไม่จำเป็นต้องใช้พจนานุกรม

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาโดยการตัดพยางค์ของรหัสเสียง

1.3 ขอบเขตของการวิจัย

1. คำทับศัพท์ที่ใช้เป็นคำทับศัพท์ระหว่างภาษาไทยและภาษาอังกฤษเท่านั้น
2. คำศัพท์ในภาษาอังกฤษที่ใช้ไม่รวมถึงคำย่อ (Abbreviation) และคำรัสพจน์ (Acronym)ยึดหลักเกณฑ์การออกเสียงของคำตามหลักของราชบัณฑิตยสถาน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำไปใช้ในระบบการสืบค้นสารสนเทศ ให้มีสามารถในค้นคืนข้ามภาษาไทย และภาษาอังกฤษได้โดยไม่ต้องอาศัยพจนานุกรม และสามารถรองรับคำที่ศัพท์ที่เกิดขึ้นใหม่ได้ รวมทั้งสามารถใช้เป็นแนวทางในการสร้างระบบการค้นคืนข้ามภาษาในภาษาอื่นหรือการค้นคืนด้วยวิธีการที่ดียิ่งขึ้นได้

1.5 วิธีดำเนินการวิจัย

1. ศึกษาทฤษฎีพื้นฐานทางด้านภาษาศาสตร์ โดยเฉพาะอย่างยิ่งเรื่องของการตัดพยางค์
2. ศึกษาวิธีการเข้ารหัสคำโดยใช้นิวรอลเน็ตเวิร์ก
3. รวบรวมและจัดเก็บชุดข้อมูลคำทับศัพท์เพื่อใช้ในการทดลอง และกำหนดรหัสคำของแต่ละคำศัพท์
4. แปลงข้อมูลคำศัพท์ให้อยู่ในรูปที่ใช้สำหรับฝึกสอนนิวรอลเน็ตเวิร์ก
5. ฝึกสอนและทดสอบนิวรอลเน็ตเวิร์ก เพื่อใช้เป็นตัวสร้างรหัสคำ
6. ออกแบบขั้นตอนวิธี และสร้างลำดับของกฎที่จะนำมาใช้ในการแยกกลุ่มของรหัสคำให้เป็นรหัสของพยางค์ โดยใช้ความรู้ทางด้านภาษาศาสตร์ และการตัดพยางค์
7. ทำการทดลองและปรับปรุงผลการทดลอง
8. สรุปผลการทดลอง และจัดทำวิทยานิพนธ์

1.6 ผลงานที่ตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในงานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ 2549 (The National Computer Science and Engineering Conference: NCSEC'06) เมื่อวันที่ 25-27 ตุลาคม พ.ศ. 2549 ในบทความเรื่อง

"การเข้ารหัสและนับจำนวนพยางค์ของรหัสคำด้วยแบ็คพรอพาเกชั่นนิวรอลเน็ตเวิร์ก" (Word Encoding and Syllable Enumeration of Phonetic Codes Using by A Back-Propagation Neural Networks) [7] โดยผู้นำเสนอคือ โอภาส วงษ์ทวีทรัพย์ บุญเสริม กิจศิริกุล และ สมชาย ประสิทธิ์จูตระกูล