1. บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ระบบค้นคืนสารสนเทศเป็นเครื่องมือที่สำคัญอย่างยิ่งในการบริหารสารสนเทศที่มีอยู่จำนวนมากในรูปของสื่ออิเล็กโทรนิกส์ โดยเฉพาะเวิลด์ไวด์เว็บ และซีดีรอม เนื่องจากสื่อดังกล่าวสามารถจัดเก็บสารสนเทศได้เป็นจำนวนมาก และสามารถเข้าถึงได้ง่ายทั้งจากระยะใกล้และไกล

โดยปกติแล้วการวัดประสิทธิผลของระบบค้นคืนสารสนเทศใด ๆ มักจะวัดจากค่าแม่นยำ (Precision) และ ค่าเรียกคืน (Recall) [1] ซึ่ง ค่าแม่นยำ หมายถึงการวัดความสามารถของระบบในการที่จะขจัดเอกสารที่ไม่เกี่ยวข้องออกไป ค่าที่ได้จะเป็นอัตราส่วนระหว่างจำนวนของเอกสารที่เกี่ยวข้องที่คืนกลับมา กับจำนวนเอกสารทั้งหมดที่กลับคืนมา ส่วนค่าเรียกคืนหมายถึงการวัดความสามารถของระบบในการคืนเอกสารที่เกี่ยวข้องกลับมา ค่าที่ได้จะเป็นอัตราส่วนระหว่างจำนวนของเอกสารที่เกี่ยวข้องที่คืนกลับมา กับจำนวนทั้งหมดของเอกสารที่เกี่ยวข้อง

ปัญหาเกิดขึ้นเมื่อผู้ใช้ป้อนคำหลัก (Keyword) ด้วยภาษาใดภาษาหนึ่ง ในขณะที่คำหลักในเอกสารจัดเก็บด้วยภาษาอื่น ตัวอย่างเช่น ผู้ใช้ต้องการสืบค้นคำว่า "ALEXANDER" แต่ระบบไม่ได้คืนเอกสารที่มีคำว่า "อเล็กซานเดอร์" (คำทับศัพท์ที่ตรงกัน) ทำให้ค่าเรียกคืนของระบบค้นคืนสารสนเทศน้อยกว่าที่ควรจะเป็น ถ้าระบบดังกล่าวไม่สนับสนุนการทำงานแบบข้ามภาษา

การค้นคืนสารสนเทศข้ามภาษา (Cross-Language Information Retrieval) หมายถึง การค้นคืนสารสนเทศ โดยภาษาที่ใช้ในข้อคำถามแตกต่างจากภาษาที่ใช้ในการจัดเก็บเอกสาร[2] โดยทั่วไปแล้วมักพบว่าเอกสารทางวิชาการที่จัดทำเป็นภาษาไทยมักจะมีคำนามเฉพาะ (Proper Noun) และคำศัพท์ทางเทคนิคจำนวนมากปรากฏอยู่ในรูปของคำในภาษาอังกฤษหรือคำทับศัพท์ การใช้พจนานุกรมสองภาษา (Bilingual Dictionary) ในลักษณะของอรรถาภิธาน (Thesaurus) กับระบบค้นคืนสารสนเทศไม่สามารถแก้ไขปัญหาดังกล่าวได้มากนัก เนื่องจากคำทับศัพท์ส่วนมากมักไม่

---

[1] W. B. Frakes and R. Baeza-Yates, Information Retrieval : Data Structures and Algorithms (Englewood Cliffs, N.J. : Prentice-Hall, 1992).

[2] D. Oard and B. Dorr, A Survey of Multilingual Text Retrieval, Technical Report UMIACS-TR-96-19 CD-TR-3615, University of Maryland, College Park, April 1996.

ปรากฏในพจนานุกรม [3] ดังนั้น การสอบถามด้วยคำหลักภาษาหนึ่งจะทำให้พลาดเอกสารที่มีคำหลักตรงกันในภาษาอื่น ถ้าระบบค้นคืนสารสนเทศไม่สนับสนุนการทำงานในแบบข้ามภาษา

งานวิจัยนี้จะเน้นการค้นคืนคำทับศัพท์ข้ามภาษาไทย-อังกฤษ โดยจะแบ่งขั้นตอนวิธีออกเป็น 2 ส่วน คือ ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ กรณีคำทับศัพท์นั้นเป็นภาษาไทยทับศัพท์ภาษาอังกฤษ เช่น "คลินตัน" ทับศัพท์ "CLINTON" และอีกส่วนหนึ่งคือ กรณีคำทับศัพท์นั้นเป็นภาษาอังกฤษทับศัพท์ภาษาไทย เช่น "SOMPORN" ทับศัพท์ "สมพร" เนื่องจากทั้ง 2 กรณีใช้วงความรู้แตกต่างกันมากในการแก้ปัญหา ตัวอย่างเช่น กรณีภาษาไทยทับศัพท์ภาษาอังกฤษ ตามหลักภาษาแล้วจะพยายามถอดทุกตัวอักษรของภาษาอังกฤษมาเป็นตัวอักษรภาษาไทย และอักษรตัวที่ไม่ออกเสียงในภาษาไทยก็ให้ใส่เครื่องหมายทัณฑฆาตกำกับไว้ [4] เช่น "WINDSOR" ถอดอักษรเป็น "วินด์เซอร์" แต่ถ้าเป็นกรณีภาษาอังกฤษทับศัพท์ภาษาไทยมักจะถอดตามเสียงที่อ่านได้เช่น "บุญเสริมทรัพย์" ถอดอักษรเป็น "BOONSERMSAP" จะเห็นว่าทั้ง "สร" และ "ทร" ถอดอักษรเป็น "S"

งานวิจัยนี้มีข้อสมมุติฐานว่าขั้นตอนวิธีที่นำเสนอจะสามารถทำการสืบค้นข้ามภาษาไทย-อังกฤษได้โดยไม่ต้องอาศัยพจนานุกรม การข้ามภาษาจะเป็นลักษณะของคำทับศัพท์ และผู้ใช้ต้องแจ้งให้ขั้นตอนวิธีทราบว่าจะข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษหรือภาษาอังกฤษทับศัพท์ภาษาไทย ส่วนขั้นตอนการทำงานหลัก ๆ ดังนี้ คือ เมื่อผู้ใช้งานได้ระบุข้อคำถามให้กับระบบค้นคืนข้ามภาษาแล้ว ระบบจะทำการเข้ารหัสคำในข้อคำถามแล้วนำรหัสคำที่ได้ไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี รหัสคำใดที่ผ่านเงื่อนไขการเปรียบเทียบจะถือว่าคำหลักนั้นเป็นคำหลักที่ตรงกันในอีกภาษาหนึ่ง

---

[3] K. Knight and J. Graehl, Machine Transliteration, Annual Meeting of the Association for Computational Linguistics (ACL-97/EACL-97).

[4] ราชบัณฑิตยสถาน, หลักเกณฑ์การทับศัพท์ ฉบับราชบัณฑิตยสถาน (2535).

1.2 วัตถุประสงค์ของการวิจัย

เพื่อออกแบบและพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ

1.3 ขอบเขตของการวิจัย

1. คำทับศัพท์ที่ใช้เป็นการทับศัพท์ระหว่างคำภาษาอังกฤษกับคำภาษาไทยเท่านั้น
2. คำศัพท์ในภาษาอังกฤษที่ใช้ไม่รวมถึงคำย่อและรัสพจน์ (Acronym)
3. คำทับศัพท์อังกฤษ-ไทย ที่ใช้ในการทดสอบขั้นตอนวิธี จะต้องใช้หลักเกณฑ์การทับศัพท์ของราชบัณฑิตยสถาน

1.4 ขั้นตอนและวิธีดำเนินการวิจัย

1. ศึกษาขั้นตอนวิธีการค้นคืนสารสนเทศข้ามภาษา
2. ศึกษาหลักภาษาในการถอดอักษร และหลักเกณฑ์การทับศัพท์
   2.1. จากภาษาอังกฤษเป็นภาษาไทย
   2.2. จากภาษาไทยเป็นภาษาอังกฤษ
3. ศึกษาขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษและภาษาไทย
4. ออกแบบและพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษา
5. ออกแบบวิธีการทดสอบขั้นตอนวิธี
6. ทดสอบและปรับปรุงคุณภาพของขั้นตอนวิธี
7. สรุปผลการวิจัย และจัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

งานวิจัยนี้สามารถนำไปใช้เป็นแนวทางในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาอื่น ๆ ต่อไป เช่น ไทย-ญี่ปุ่น ไทย-ฝรั่งเศส ไทย-เยอรมัน เป็นต้น

1.6 ผลงานที่ตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในงานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ 2541 (The National Computer Science and Engineering Conference: NCSEC'98) เมื่อวันที่ 19-21 ตุลาคม พ.ศ. 2541 ในบทความเรื่อง "Thai-English Cross-Language Transliterated Word Retrieval Using Soundex Technique" โดยผู้นำเสนอคือ Somchai Prasitjutrakul และ Prayut Suwanvisat

1.7 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้ถูกแบ่งออกเป็น 5 บท ดังนี้ คือ บทที่ 1 ซึ่งเป็นบทนำ บทที่ 2 จะกล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้อง เช่น หลักการถ่ายอักษร ขั้นตอนวิธีซาวด์เด็กซ์ และการค้นคืนสารสนเทศข้ามภาษา เป็นต้น ส่วนบทที่ 3 จะกล่าวถึงการออกแบบ การทดลอง และผลการทดลองของขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาในส่วนของภาษาไทยทับศัพท์ภาษาอังกฤษ ในบทที่ 4 เป็นการออกแบบ การทดลอง และผลการทดลองในส่วนของภาษาอังกฤษทับศัพท์ภาษาไทย และท้ายสุดคือบทที่ 5 จะเป็นบทสรุปของการวิจัย รวมทั้งข้อเสนอแนะต่าง ๆ ในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำเพื่อการค้นข้ามภาษาไทย-อังกฤษให้ดียิ่งขึ้น