



5. สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาระบบคั่นคินสารสนเทศต่าง ๆ นั้น ผู้วิจัยพบว่าปัญหาการคั่นคินสารสนเทศข้ามภาษา เป็นปัญหาที่น่าสนใจโดยเฉพาะการคั่นคินสารสนเทศข้ามภาษาไทย-อังกฤษในส่วนของคำทับศัพท์ การพัฒนาระบบการคั่นคินสารสนเทศข้ามภาษาที่ผ่านมามีหลายแนวทางซึ่งแต่ละแนวทางจะมีข้อดีข้อเสียแตกต่างกันไป ส่วนแนวทางที่ใช้ในการวิจัยครั้งนี้จะเป็นการเข้ารหัสคำทับศัพท์ ซึ่งผู้วิจัยได้พบว่าปัญหาการคั่นคินสารสนเทศข้ามภาษาไทย-อังกฤษในส่วนของคำทับศัพท์นั้นยังสามารถแบ่งออกเป็นปัญหาย่อย ๆ ได้อีกสองปัญหาคือคำทับศัพท์นั้นเป็นได้ทั้งภาษาไทยทับศัพท์ภาษาอังกฤษ และภาษาอังกฤษทับศัพท์ภาษาไทย

ในงานวิจัยชิ้นนี้ได้เน้นเฉพาะในเรื่องของขั้นตอนวิธีในการเข้ารหัสคำทับศัพท์เพื่อการคั่นคินข้ามภาษาไทย-อังกฤษเป็นหลัก เนื่องจากในส่วนประกอบอื่น ๆ ของระบบคั่นคินสารสนเทศนั้นสามารถนำสิ่งที่ได้มีการคิดค้นขึ้นมาแล้วมาใช้ร่วมกับขั้นตอนการเข้ารหัสคำที่พัฒนาขึ้นได้

5.1 สรุปผลการวิจัย

ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการคั่นคินข้ามภาษาไทย-อังกฤษที่ผู้วิจัยได้นำเสนอนี้แบ่งออกเป็นสองขั้นตอนวิธี เพื่อแก้ปัญหาสองปัญหาคือ (1) ภาษาไทยทับศัพท์ภาษาอังกฤษ และ (2) ภาษาอังกฤษทับศัพท์ภาษาไทย โดยทั้งผู้วิจัยได้ใช้ความรู้ทางด้านหลักภาษามาช่วยในการออกแบบการประมวลผลตัวอักษรเบื้องต้นเพื่อให้คำศัพท์ทั้งสองภาษามีรูปแบบที่เหมือนกันก่อนที่ไปนำไปเข้ารหัสคำ การเข้ารหัสคำนั้นผู้วิจัยได้เลือกใช้เทคนิคชาวเด็กซ์เพื่อให้คำที่อ่านออกเสียงเหมือนกันได้รหัสเหมือนกัน หลังจากนั้นจะนำรหัสคำที่ได้มาทำการเปรียบเทียบกัน ซึ่งขั้นตอนวิธีนี้จะใช้เทคนิคของระยะแก้ไขขั้นที่สุดโดยจะเน้นในเรื่องการอ่านออกเสียงของอักขระมากกว่ารูปอักขระมาัดดแปลงเพื่อให้เหมาะสมกับการอ่านออกเสียงของภาษาไทย และผู้วิจัยได้ออกแบบเงื่อนไขที่ใช้ทดสอบว่ารหัสคำทั้งสองว่าเป็นรหัสคำที่ได้มาจากคำทับศัพท์ที่ตรงกันระหว่างภาษาไทยกับภาษาอังกฤษ

5.2 ข้อดีและข้อเสียของขั้นตอนวิธี

ข้อดีของขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ ได้แก่

1. ขั้นตอนวิธีที่นำเสนอมีลักษณะเป็นขั้นตอนวิธีล้วน ๆ ทำงานโดยไม่ต้องอาศัยพจนานุกรมคำศัพท์ ทำให้ไม่มีข้อจำกัดในเรื่องของการใช้คำศัพท์ใหม่
2. ขั้นตอนวิธีที่นำเสนอมีความยืดหยุ่นโดยสามารถปรับเปลี่ยนค่าพารามิเตอร์ที่กำหนด ความถูกต้องของการค้นคืนได้ เพื่อเพิ่มความแม่นยำหรือคำเรียกคืนได้ตามความต้องการของผู้ใช้งาน
3. ขั้นตอนวิธีใช้เทคนิคของชาวคัลเด็กซ์ทำให้เพิ่มคำเรียกคืนในกรณีที่คำหลักมีการสะกดต่างกันแต่อ่านออกเสียงเหมือนกัน
4. หลักเกณฑ์การทับศัพท์ที่ใช้ในขั้นตอนวิธี อ้างอิงจากหลักเกณฑ์การทับศัพท์ของราชบัณฑิตยสถาน และความนิยมในการทับศัพท์ ทำให้สามารถใช้งานได้จริงถึงแม้ว่าผู้ใช้จะสะกดไม่ตรงตามราชบัณฑิตยสถานทุกประการก็ตาม

อย่างไรก็ตามการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ ก็มีข้อเสียเช่นกัน ซึ่งข้อเสียเหล่านี้ได้แก่

1. ในกรณีที่มีคำหลักมีการอ่านออกเสียงคล้ายกัน เอกสารที่ระบบค้นกลับมาอาจจะไม่ตรงกับความต้องการของผู้ใช้ เนื่องจากขั้นตอนวิธีที่นำเสนอใช้เทคนิคของชาวคัลเด็กซ์ เช่น ผู้ใช้ต้องการค้นคืนคำว่า กมณวัณย์ (KAMONWAN) ระบบค้นคืนคำว่า กุมลาวัณย์ (KAMALAWAN)
2. การเปรียบเทียบรหัสคำมีลักษณะแบบโดยประมาณ (Approximate Search) ทำให้ต้องเปรียบเทียบรหัสคำกับฐานรหัสคำทั้งหมด

อย่างไรก็ตามข้อเสียเหล่านี้สามารถแก้ไขได้ เช่น การปรับเปลี่ยนพารามิเตอร์แอฟฟาเพื่อเพิ่มความแม่นยำของเอกสารที่ระบบค้นกลับมา แต่ก็จะทำให้คำเรียกคืนลดลง และการเลือกใช้โครงสร้างข้อมูลที่เหมาะสมกับปัญหานี้ก็คือ ทริช (Trie)¹ ซึ่งเป็นโครงสร้างข้อมูลที่มีลักษณะพิเศษคือข้อมูลที่มีส่วนเริ่มต้นเหมือนกันจะถูกจัดเก็บไว้ในที่เดียวกันทำให้ลดปริมาณการเปรียบเทียบรหัสคำลง เนื่องจากส่วนที่เหมือนกันของคำศัพท์จะถูกเปรียบเทียบเพียงหนึ่งครั้งเท่านั้น

¹ H. Shang and T. H. Merrettal, *Tries for Approximate String Matching*, IEEE Transactions On Knowledge and Data Engineering., Vol. 8, No. 4, pp. 540-547, August 1996.

5.3 ข้อเสนอแนะ

ผู้วิจัยพบว่า มีแง่มุมและข้อเสนอแนะบางประการของการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษที่น่าสนใจ ซึ่งเป็นแง่มุมซึ่งไม่ได้นำมาใช้ในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ

ข้อเสนอแนะเหล่านี้ได้แก่

1. ควรจะพัฒนาขั้นตอนวิธีเพื่อแก้ปัญหาการค้นคืนข้ามภาษาไทย-อังกฤษ ในส่วนที่เป็นภาษาไทยทับศัพท์ภาษาอังกฤษ และภาษาอังกฤษทับศัพท์ภาษาไทยให้เป็นขั้นตอนวิธีเดียว เพื่อให้เกิดความสะดวกยิ่งขึ้นในการนำไปใช้งาน เนื่องจากพบว่าเอกสารบางอย่างมีทั้งสองปัญหาอยู่ในเอกสารเดียวกัน
2. ควรจะคงคำหลักในภาษาใดภาษาหนึ่ง เพื่อลดภาระการทำดัชนีคำหลักซ้ำใหม่ทั้งหมด ในกรณีที่มีฐานคำหลักเป็นดัชนีอยู่ก่อนแล้ว ขั้นตอนวิธีจะถูกใช้ในลักษณะโปรแกรมเสริม
3. ควรจะใช้การเปรียบเทียบคำหลักในลักษณะข้ามภาษา แทนที่จะนำคำหลักเหล่านั้นมาเข้ารหัสคำก่อน เช่น นำคำว่า ALEXANDER มาเปรียบเทียบกับ อเล็กซานเดอร์ ว่าเป็นคำหลักที่ตรงกันในภาษาอื่นหรือไม่ โดยใช้เทคนิคของการเปรียบเทียบคู่สายอักขระโดยประมาณ และใช้โครงสร้างข้อมูลแบบทรี เพื่อเพิ่มประสิทธิภาพเมื่อนำไปใช้กับงาน

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย