บทที่ 4

การค้นคืนข้ามภาษา

ขั้นตอนการค้นคืนข้ามภาษาประกอบด้วยการนำรหัสคำของคำที่เป็นข้อคำถาม ไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักที่ถูกเก็บไว้ในฐานข้อมูลทุกคำ โดยถ้าผลการเปรียบเทียบออกมาอยู่ในเกณฑ์ที่กำหนด (ที่ตั้งไว้และยอมรับได้) ก็จะถือว่ารหัสคำทั้งสองนั้นมีเสียงอ่านที่ตรงกัน และระบบจะค้นคืนคำนั้นกลับออกไปให้กับผู้ใช้งาน สำหรับเนื้อหาในบทนี้จะอธิบายถึงวิธีการที่ใช้ในการเปรียบเทียบรหัสคำ เกณฑ์ที่ใช้ในการเปรียบเทียบรหัสคำ

4.1 การคำนวณความต่างของรหัสคำ

เนื่องจากรหัสคำที่ได้จากขั้นตอนการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กนั้น แม้ว่าคู่คำจากทั้งสองภาษาจะออกเสียงเหมือนกัน ก็อาจมีรหัสไม่เหมือนกันทุกตัวอักษร ดังนั้นในการเปรียบเทียบรหัสคำ จึงต้องใช้วิธีการเปรียบเทียบแบบประมาณ (Approximate Matching) ซึ่งอาศัยการคำนวณความแตกต่าง (Distance) ของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นสุด (Minimum Edit Distance) ซึ่งเป็นการคำนวณความคล้ายคลึงกันระหว่างสายอักขระ 2 สาย โดยคำนวณหาจากจำนวนครั้งที่น้อยที่สุดที่ใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้รหัสคำทั้งสองเหมือนกัน การคำนวณค่าความแตกต่างนี้อาศัยเทคนิคกำหนดการพลวัต (Dynamic Programming) ซึ่งวิธีการคำนวณสามารถเขียนให้อยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ดังนี้

$$\text{Edit } (P_0, W_0) = 0$$

$$\text{Edit } (P_j, W_0) = j$$

$$\text{Edit } (P_0, W_k) = k$$

$$\text{Edit } (P_j, W_k) = \min \{ \text{Edit } (P_{j-1}, W_k) + 1,$$
$$\text{Edit } (P_j, W_{k-1}) + 1,$$
$$\text{Edit } (P_{j-1}, W_{k-1}) + R(p_j, w_k) \ \}$$

โดยที่

$P_j = p_1\, p_2\, p_3 \ldots\, p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1\, w_2\, w_3 \ldots\, w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$R(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

$= 1$ ถ้า p_j ไม่เท่ากับ w_k

4.2 เกณฑ์การเปรียบเทียบรหัสคำ

สำหรับขั้นตอนของการค้นคืนข้ามภาษา เพื่อให้ได้ค่าความเที่ยง (Precision) และค่าเรียกคืน (Recall) ที่ดี จึงได้ใช้การเปรียบเทียบรหัสคำแบบประมาณโดยได้มีการกำหนดเกณฑ์ที่จะใช้ในการเปรียบเทียบขึ้นมา สำหรับเกณฑ์ที่จะใช้ในการเปรียบเทียบรหัสคำนั้น จะหาได้จากค่าความแตกต่างระหว่างรหัสคำ โดยถ้าค่าความแตกต่างที่ได้มีค่าไม่เกินเกณฑ์ (ค่าคงที่ d) ก็จะสรุปได้ว่ารหัสคำทั้งสองเป็นรหัสคำที่มาจากคำหลักที่ตรงกันในอีกภาษา สำหรับการค้นคืนข้ามภาษาในงานวิจัยนี้ได้ลองทำการทดลองทั้งหมด 4 กรณีด้วยกัน คือ

กรณีที่ 1 ใช้การเปรียบเทียบแบบประมาณกับรหัสคำแบบดั้งเดิม (ใช้ตารางที่ 3.1 และตารางที่ 3.2 ในการเข้ารหัส) โดยในขั้นตอนนี้จะใช้นิวรอลตัวที่ให้ผลการเรียนรู้ดีที่สุดมาใช้ในการทดสอบการเข้ารหัสคำ เมื่อได้รหัสเสียงของคำของทุกตัวอักษรแล้ว จะนำรหัสเสียงทั้งหมดมาเรียงต่อกันตามลำดับเพื่อสร้างเป็นรหัสคำอ่าน โดยสุดท้ายเมื่อได้รหัสคำอ่านแล้ว จะทำการตัดรหัสที่ไม่ออกเสียง (_) ออก และทำการย้ายรหัสเสียงของสระไปต่อท้ายเสียงของพยัญชนะ ดังตัวอย่างที่ 1 ด้านล่างนี้

ตัวอย่างที่ 1 ต้องการทดสอบคำว่า "เปล่งสมบัติ" และ "plengsombut" ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

เปล่งสมบัติ --> เปลงสมบัติ --> eplgsombat_ --> plgsmbteoa

plengsombut --> pleg_sombat --> plgsmbteoa

การค้นคืน

Edit("plgsmbteoa" , "plgsmbteoa") = 0

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า "plengsombut" จากคำว่า "เปล่งสมบัติ" ได้

กรณีที่ 2 ใช้การเปรียบเทียบแบบประมาณกับการเข้ารหัสคำแบบใหม่ (ใช้ตารางที่ 3.4 ในการเข้ารหัส) ซึ่งรหัสคำอ่านในกรณีนี้จะมี 3 ส่วนคือ รหัสคำของเสียงพยัญชนะที่เป็นพยัญชนะต้น รหัสคำของเสียงสระ และรหัสคำของเสียงพยัญชนะที่เป็นตัวสะกด โดยในขั้นตอนนี้จะใช้นิวรอลตัวที่ให้ผลการเรียนรู้ดีที่สุดมาใช้ในการทดสอบการเข้ารหัสคำ เมื่อได้รหัสเสียงของคำของทุกตัวอักษรแล้ว จะนำรหัสเสียงทั้งหมดมาเรียงต่อกันตามลำดับเพื่อสร้างเป็นรหัสคำอ่าน โดยสุดท้ายเมื่อได้รหัสคำอ่านแล้ว จะทำการตัดรหัสที่ไม่ออกเสียง (_) ออก และทำการย้ายรหัสเสียงให้เรียงเป็นรหัสเสียงของพยัญชนะต้น รหัสของเสียงสระ และรหัสของเสียงตัวสะกด ดังตัวอย่างที่ 2 ด้านล่างนี้

ตัวอย่างที่ 2 ต้องการทดสอบคำว่า "รุ่งแสงมนูญ" และ "roongsangmanoon" ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

รุ่งแสงมนูญ --> รุงแสงมนูญ --> rugwsgMaNun --> rsMNuwauggn

roongsangmanoon --> ru_g_swg_MaNu_n --> rsMNuwauggn

การค้นคืน

Edit("rsMNuwauggn" , "rsMNuwauggn") = 0

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า "roongsangmanoon" จากคำว่า "รุ่งแสงมนูญ" ได้

กรณีที่ 3 ใช้การเปรียบเทียบแบบประมาณกับการเข้ารหัสคำแบบใหม่ คล้ายๆ กับกรณีที่ 2 แต่จะแตกต่างกันตรงที่จะไม่ย้ายตำแหน่งของรหัสคำอ่าน แต่จะใช้วิธีการประมวลผลรหัสคำอ่าน โดยทำการจัดวางตำแหน่งใหม่ ให้ตำแหน่งของรหัสคำอ่านเขียนเรียงเป็นพยางค์ๆ ต่อกันโดยให้ในแต่ละพยางค์มีรหัสคำ 3 ส่วนเขียนเรียงกัน คือ รหัสของเสียงพยัญชนะต้น รหัสเสียงสระ และรหัสเสียงตัวสะกด ดังตัวอย่างที่ 3 ด้านล่างนี้

ตัวอย่างที่ 3 ต้องการทดสอบคำว่า "รุ่งแสงมนูญ" และ "roongsangmanoon" ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

รุ่งแสงมนูญ --> รุงแสงมนูญ --> rugwsgMaNun --> rugswgMaNun

roongsangmanoon --> ru_g_swg_MaNu_n --> rugswgMaNun

การค้นคืน

Edit("rugswgMaNun" , "rugswgMaNun") = 0

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า "roongsangmanoon" จากคำว่า "รุ่งแสงมนูญ" ได้

กรณีที่ 4 คล้ายๆ กับกรณีที่ 3 แต่จะใช้วิธีการตัดแบ่งพยางค์ของรหัสคำอ่านให้เป็นหน่วยย่อยๆ โดยในหน่วยย่อยนั้น จะเป็นหน่วยย่อยของรหัสเสียงพยางค์ ดังตัวอย่างที่ 4 ด้านล่างนี้

ตัวอย่างที่ 4 ต้องการทดสอบคำว่า "รุ่งแสงมนูญ" และ "roongsangmanoon" ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

รุ่งแสงมนูญ --> รุงแสงมนูญ --> rugwsgMaNun --> rugswgMaNun

--> rug-swg-Ma-Nun

roongsangmanoon --> ru_g_swg_MaNu_n --> rugswgMaNun

--> rug-swg-Ma-Nun

<u>การค้นคืน</u>

Edit("rug-swg-Ma-Nun" , "rug-swg-Ma-Nun")

= Edit("rug","rug") + Edit("swg","swg") + Edit("Ma","Ma") + Edit("Nun","Nun")

= 0 + 0 + 0 + 0 = 0

จากตัวอย่างพบว่า รหัสคำจะถูกแยกออกมาเป็นส่วนๆ ในหน่วยของรหัสพยางค์ แล้วทำการเปรียบเทียบแบบประมาณทีละพยางค์ แล้วหาค่าผลรวมของค่าความแตกต่างๆ ในแต่ละพยางค์ ซึ่งจากตัวอย่างด้านบนพบว่า ค่าความแตกต่างของเสียงอ่านของทั้งสองคำมีค่าผลรวมของความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า "roongsangmanoon" จากคำว่า "รุ่งแสงมนูญ" ได้

4.3 สรุป

ในบทนี้ได้กล่าวถึงการค้นคืนข้ามภาษา โดยได้อธิบายถึงวิธีการคำนวณหาค่าความต่างของรหัสคำ ด้วยการเปรียบเทียบแบบประมาณด้วยเทคนิคระยะแก้ไขสั้นสุด จากนั้นได้อธิบายถึงเกณฑ์การเปรียบเทียบรหัสคำ เพื่อใช้ในการตัดสินว่ารหัสคำทั้งสองรหัสที่เปรียบเทียบกันนั้นมีเสียงอ่านที่ตรงกันหรือไม่ และได้กล่าวถึงวิธีการที่จะใช้ในการทดลองทั้ง 4 กรณีของงานวิจัยนี้