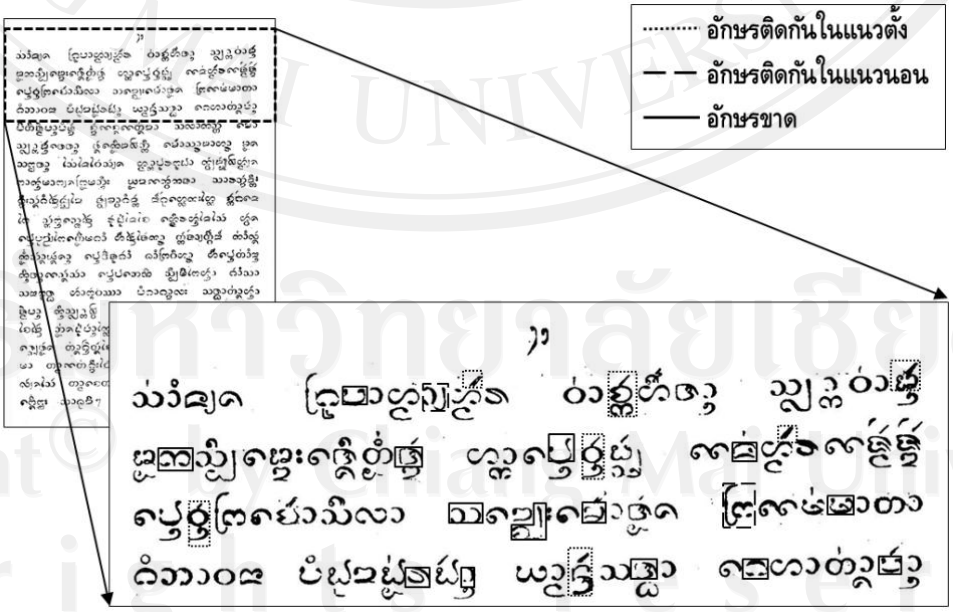
บทที่ 3
วิธีดำเนินการวิจัย

ในบทนี้กล่าวถึงวิธีดำเนินการวิจัย ซึ่งมีการดำเนินการทั้งหมด 3 ส่วน โดยส่วนแรกเป็นชุดข้อมูลที่ใช้ในการวิจัย ส่วนที่สองเป็นรายละเอียดกระบวนการรู้จำตัวพิมพ์อักษรธรรมล้านนา ซึ่งได้นำทฤษฎีและวิธีการที่กล่าวไว้ในบทที่ 2 มาใช้ในงานวิจัย และส่วนที่สามเป็นการรู้จำตัวพิมพ์อักษรธรรมล้านนาด้วยวิธีการอื่น โดยมีรายละเอียดดังต่อไปนี้

3.1 ชุดข้อมูลที่ใช้ในการวิจัย

ในงานวิจัยนี้ใช้ข้อมูลภาพจากโครงการ e-60 วรรณพิมพ์ล้านนา ซึ่งเป็นโครงการที่รวบรวมวรรณกรรมที่พิมพ์ด้วยอักษรธรรมล้านนาประมาณปี พ.ศ. 2450 ถึง พ.ศ. 2490 และนำมาจัดเก็บในรูปแบบดิจิทัล จำนวนทั้งหมด 60 เล่ม (จีรยุทธ ไชยจารุวณิช และคณะ, 2552)

ลักษณะของภาพเอกสาร เป็นภาพระดับสีเทาที่มีความกว้างอยู่ในช่วง 1,000 - 1,450 พิกเซล และมีความสูงอยู่ในช่วง 1,600 - 2,200 พิกเซล โดยภาพเอกสารถูกปรับปรุงคุณภาพเบื้องต้นแล้ว แต่ยังพบสัญญาณรบกวนปะปนอยู่ในเอกสารบ้าง ส่วนตัวอักษรถูกพิมพ์ขึ้นด้วยแม่พิมพ์ตะกั่วและเทคโนโลยีที่ยังมีข้อจำกัดในขณะนั้น ทำให้ลายเส้นของตัวอักษรบางตัวขาดหายไป หรือบางตัวอักษรมีลายเส้นหนาจนเกิดอักษรติดกัน ดังตัวอย่างในรูป 3.1 ซึ่งเป็นเอกสารในหมวดคร่าวซอ เรื่องตำนานแลสร้างวัดสวนดอกไม้ นครเชียงใหม่ พิมพ์โดยโรงพิมพ์อเมริกัน เมื่อปี พ.ศ. 2474 จำนวน 29 หน้า

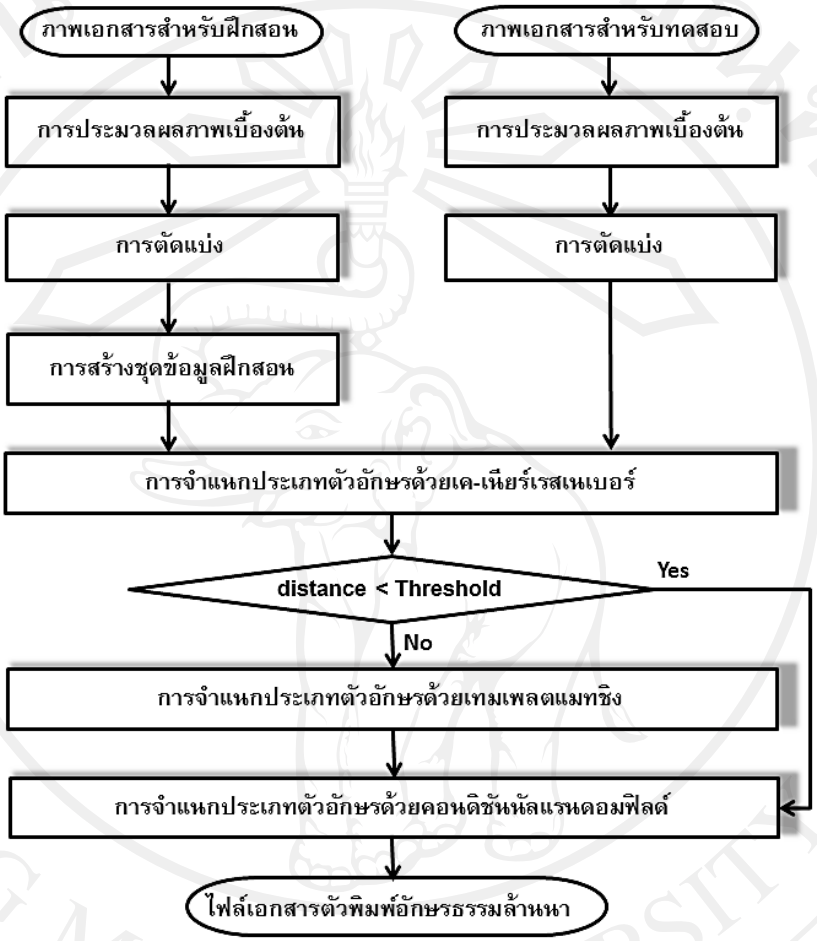


รูป 3.1 ตัวอย่างภาพเอกสารตัวพิมพ์อักษรธรรมล้านนา

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.2 กระบวนการรู้จำตัวพิมพ์อักษรธรรมล้านนา

กระบวนการรู้จำตัวพิมพ์อักษรธรรมล้านนาที่นำเสนอ ประกอบด้วย 6 ขั้นตอน แสดงดังรูป 3.2



รูป 3.2 กระบวนการรู้จำตัวพิมพ์อักษรธรรมล้านนา

ขั้นตอนแรก เริ่มจากการประมวลผลภาพเบื้องต้น เพื่อปรับปรุงคุณภาพและกำจัดส่วนที่ไม่เกี่ยวข้องในภาพ ต่อมาในขั้นตอนที่สอง การตัดแบ่ง เป็นการตัดแยกภาพให้เป็นบล็อกตัวอักษร เพื่อรู้จำทีละบล็อกตัวอักษร จากนั้นการสร้างชุดข้อมูลฝึกสอนในขั้นตอนที่สาม เพื่อให้ตัวจำแนกประเภทสามารถเรียนรู้จากกลุ่มตัวอย่างในชุดข้อมูลฝึกสอน และสามารถจำแนกข้อมูลทดสอบให้เป็นประเภทต่างๆ ได้ ขั้นตอนที่สี่ การจำแนกประเภทด้วยวิธีเค-เนียร์เรสเนเบอร์ เพื่อระบุว่าข้อมูลทดสอบแต่ละบล็อกเป็นตัวอักษรใด โดยผลลัพธ์ของขั้นตอนนี้ คือ สายลำดับของประเภทของบล็อกตัวอักษร แต่เนื่องจากลักษณะของเส้นตัวอักษรในภาพเอกสารไม่สมบูรณ์ ขาดจากกัน หรือติดกัน อีกทั้งประเภทของชุดข้อมูลฝึกสอนไม่ครอบคลุมข้อมูลทดสอบทั้งหมด หากรู้จำด้วยวิธีเค-เนียร์เรสเนเบอร์เพียงขั้นตอนเดียวอาจมีความผิดพลาดสูง จึงมีขั้นตอนที่ห้า การจำแนกประเภท

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ด้วยเทมเพลตแมทชิง และขั้นตอนที่หก การจำแนกประเภทด้วยคอนดิชันนัลแรนดอมฟิลด์ เพื่อจำแนกประเภทซ้ำอีกครั้ง กำหนดให้ค่าเทรชโฮลด์ คือค่าระยะห่างระหว่างบล็อกกับข้อมูลฝึกสอนที่ยอมรับได้ ถ้าจำแนกประเภทด้วยวิธีเค-เนียร์เรสเนเบอร์แล้วระยะห่างระหว่างบล็อกกับข้อมูลฝึกสอนน้อยกว่าค่าเทรชโฮลด์ แสดงว่าผลการจำแนกประเภทนั้นมีโอกาสผิดพลาดน้อย ดังนั้นอักษรธรรมล้านนาที่จำแนกได้จะถูกตรวจสอบความถูกต้องด้วยการจำแนกด้วยคอนดิชันนัลแรน-ดอมฟิลด์ โดยการพิจารณาตัวอักษรรอบข้าง แต่สำหรับกรณีที่จำแนกบล็อกแล้วระยะห่างระหว่างบล็อกกับข้อมูลฝึกสอนมากกว่าหรือเท่ากับค่าเทรชโฮลด์ แสดงว่าผลการจำแนกประเภทนั้นมีโอกาสผิดพลาดสูง ดังนั้นบล็อกจะถูกจำแนกซ้ำอีกครั้งด้วยวิธีเทมเพลตแมทชิงก่อน แล้วจึงตรวจสอบความถูกต้องของอักษรธรรมล้านนาที่จำแนกได้ โดยการจำแนกด้วยคอนดิชันนัลแรน-ดอมฟิลด์

3.2.1 การประมวลผลภาพเบื้องต้น

การประมวลผลภาพเบื้องต้นเป็นขั้นตอนการปรับปรุงคุณภาพของรูปภาพ เพื่อลดองค์ประกอบอื่นๆ ที่ไม่เกี่ยวข้อง ซึ่งช่วยลดความผิดพลาดที่อาจเกิดขึ้นในกระบวนการรู้จำได้ ในงานวิจัยนี้ได้แบ่งเป็น 3 ขั้นตอน คือ การแปลงเป็นภาพสองระดับ การกำจัดสัญญาณรบกวน และการปรับความเอียงของเอกสาร

(1) การแปลงเป็นภาพสองระดับ

ภาพเอกสารที่นำมาประมวลผลเป็นภาพสีเทา 256 ระดับ ค่าระดับสีเทาของแต่ละพิกเซลมีค่าเป็นไปได้อยู่ในช่วง 0 ถึง 255 ซึ่งสิ่งที่ต้องการพิจารณาในเอกสารนี้ คือ ตัวอักษรสีดำที่มีค่าเข้าใกล้ 0 ดังนั้นเพื่อให้แยกตัวอักษรจากสีพื้นหลังในภาพและหาขอบเขตของตัวอักษรได้ จึงต้องแปลงภาพระดับสีเทาให้เป็นภาพขาวดำสองระดับ ในงานวิจัยนี้ได้ใช้วิธีการของโอสึ เพื่อหาค่าเทรชโฮลด์ที่เหมาะสม แล้วแปลงภาพเป็นสองระดับ

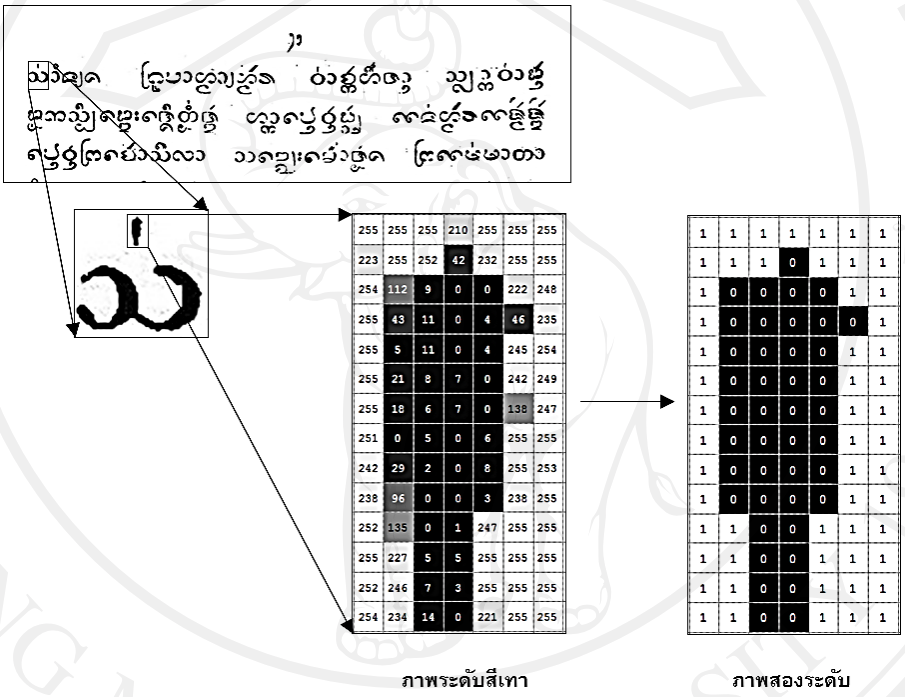
จากรูป 3.3 ภาพระดับสีเทาบริเวณตัวอักษรไม้หยัก (ˊ) มีค่าอยู่ในช่วง 0 ถึง 138 ส่วนพื้นหลังมีค่าเข้าใกล้ 255 เมื่อหาค่าเทรชโฮลด์ด้วยวิธีโอสึ ได้ค่าเท่ากับ 0.509 แล้วคูณกับระดับภาพ คือ 255 ดังนั้นค่าเทรชโฮลด์ของภาพตัวอย่างเท่ากับ 130 ดังนั้นในภาพที่มีค่ามากกว่าหรือเท่ากับค่าเทรชโฮลด์ จะเปลี่ยนเป็น 1 และค่าที่น้อยกว่าค่าเทรชโฮลด์ จะเปลี่ยนเป็น 0

เนื่องจากการดำเนินการกับภาพสองระดับนั้นดำเนินการกับวัตถุที่สนใจในภาพ ซึ่งมีค่าเท่ากับ 1 ซึ่งเป็นสีขาว ดังนั้นเมื่อผ่านขั้นตอนการแปลงภาพเป็นสองระดับแล้ว ต้องกลับสีของพิกเซลทุกจุดในภาพ ซึ่งอธิบายได้ดังสมการ (3.1)

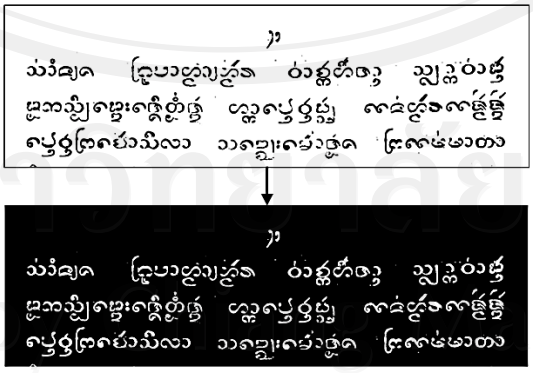
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

$$I_{bw}(x,y) = \begin{cases} 1 & if\ forground\ pixel \\ 0 & if\ background\ pixel \end{cases} \quad (3.1)$$

โดยภาพ I_{bw} ที่มีขนาด $I_{width} \times I_{height}$ เป็นภาพขาวดำที่ตำแหน่ง x และ y ใดๆ เมื่อ $x \in \{1,2,...,I_{width}\}$ และ $y \in \{1,2,...,I_{height}\}$ จุดที่มีสีดำที่เป็นส่วนของตัวอักษรจะถูกเปลี่ยนเป็นสีขาว หรือเปลี่ยนค่าของพิกเซลจาก 0 เป็น 1 และจุดที่มีสีขาวที่เป็นส่วนของพื้นหลังจะถูกเปลี่ยนเป็นสีดำ หรือเปลี่ยนค่าของพิกเซลจาก 1 เป็น 0 เพื่อใช้ในขั้นตอนต่อไป ดังรูป 3.4



รูป 3.3 ตัวอย่างการแปลงเป็นภาพสองระดับ



รูป 3.4 ตัวอย่างการกลับสีของภาพสองระดับ

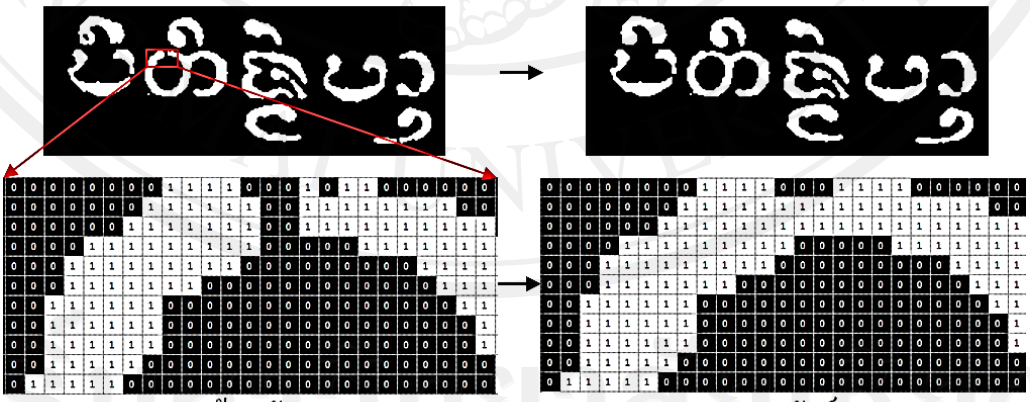
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

(2) การกำจัดสัญญาณรบกวน

เนื่องจากเอกสารต้นฉบับมีความเก่าแก่ จึงทำให้ภาพเอกสารมีสัญญาณรบกวนปะปนมาด้วย ซึ่งสัญญาณรบกวนในที่นี้หมายถึงสัญญาณรบกวน 2 ประเภท คือ สัญญาณรบกวนจุดสีขาวที่ปรากฏในภาพเอกสารในตำแหน่งที่ควรจะเป็นสีดำ และสัญญาณรบกวนจุดสีดำที่ปรากฏในภาพเอกสารในตำแหน่งที่ควรจะเป็นสีขาว ในขั้นตอนนี้ประกอบด้วย 2 ขั้นตอนย่อย คือ การสร้างความสมบูรณ์ให้กับตัวอักษรด้วยการดำเนินการเชิงสัณฐานวิทยา และการกำจัดสัญญาณรบกวนด้วยการระบุป้ายให้องค์ประกอบที่เชื่อมต่อ

1) การสร้างความสมบูรณ์ให้กับตัวอักษร

เนื่องจากตัวอักษรในภาพมีจุดสีดำปรากฏในเส้นตัวอักษรสีขาว ทำให้เกิดช่องว่างภายในเส้นตัวอักษร บางตัวอักษรมีช่องว่างมากจนเส้นอักษรเกือบขาดจากกัน งานวิจัยนี้จึงใช้การดำเนินการเชิงสัณฐานวิทยาเพื่อสร้างความสมบูรณ์ให้กับตัวอักษรโดยอาศัยหลักการทางคณิตศาสตร์เกี่ยวกับรูปแบบและโครงสร้าง ซึ่งทำงานโดยนำสตรักเจอร์อิลิเมนต์ ประกอบด้วยค่า 0 หรือ 1 ขนาดหนึ่งไปวางบนภาพทีแต่ละตำแหน่งบนภาพ โดยใช้ตัวดำเนินการระหว่างสตรักเจอร์-อิลิเมนต์กับภาพที่อยู่ใต้สตรักเจอร์อิลิเมนต์ แล้วนำผลลัพธ์มาปรับค่าที่ตำแหน่งนั้น ในงานวิจัยนี้ใช้การดำเนินการเชิงสัณฐานวิทยาแบบโคลสซิง และสตรักเจอร์อิลิเมนต์ที่มีความกว้าง 3 พิกเซล ยาว 3 พิกเซล จากการดำเนินการส่งผลให้ช่องว่างขนาดเล็กในตัวอักษรถูกเติมจนเต็ม เส้นเชื่อมต่อเป็นตัวอักษรเดียวกันเมื่อวัตถุอยู่ใกล้กันมาก และตัวอักษรมีขอบเรียบขึ้น โดยที่ขนาดไม่เปลี่ยนไป



รูป 3.5 ตัวอย่างการสร้างความสมบูรณ์ให้กับตัวอักษรด้วยการดำเนินการเชิงสัณฐานวิทยา

จากรูป 3.5 ตัวอักษรต้นฉบับในกรอบสี่เหลี่ยม คือ อักษร ๓ ที่ตำแหน่งใกล้รอยหยักมีเส้นอักษรขาดจากกัน เมื่อผ่านขั้นตอนการสร้างความสมบูรณ์ให้กับตัวอักษรด้วยการดำเนินการเชิงสัณฐานวิทยา ทำให้เส้นอักษรที่ขาดสามารถเชื่อมต่อกันได้ และช่องว่างในตัวอักษร

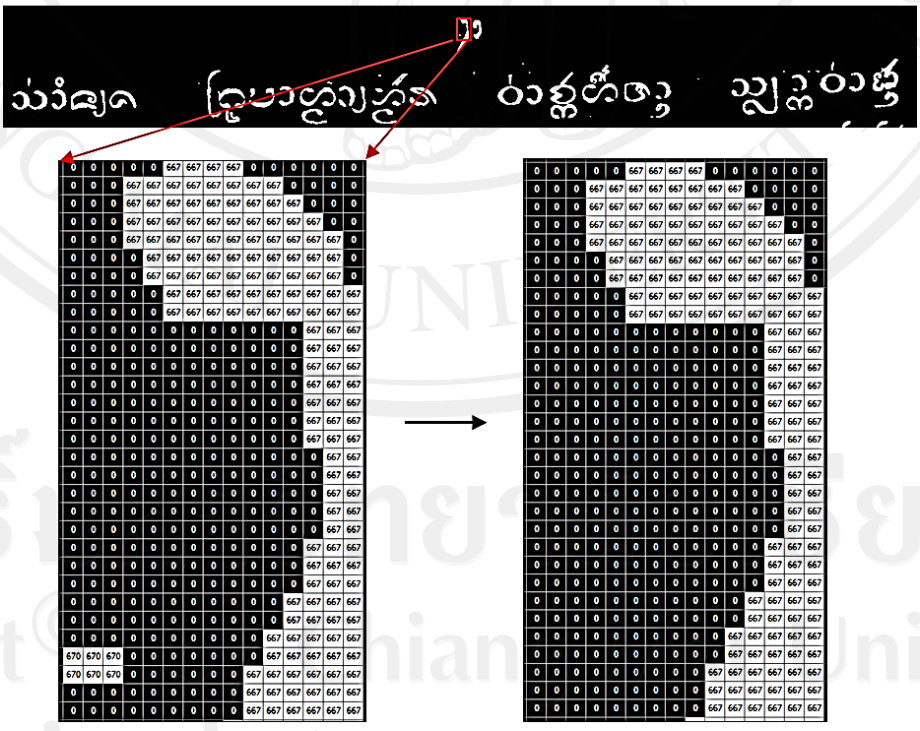
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ถูกเติมให้เต็ม นอกเหนือจากตัวอักษรในกรอบแล้ว อักษร ◌ั ที่เส้นขาดจากกัน เมื่อดำเนินการเชิงสัณฐานวิทยา เส้นถูกทำให้เชื่อมต่อกันเช่นเดียวกัน หากไม่มีการดำเนินการเชิงสัณฐานวิทยา ชิ้นส่วนเล็กๆ ที่ขาดอาจถูกกำจัดออกไปในขั้นตอนการกำจัดสัญญาณรบกวน เนื่องจากมีขนาดใกล้เคียงกับสัญญาณรบกวน

2) การกำจัดสัญญาณรบกวน

สัญญาณรบกวน คือ พิกเซลขนาดเล็กที่ไม่ต้องการพิจารณา ดังนั้นการกำจัดออกจากภาพ จึงใช้จำนวนของวัตถุเป็นตัวแยกแยะว่า วัตถุนั้นเป็นตัวอักษรหรือสัญญาณรบกวน ในงานวิจัยนี้ใช้วิธีการระบุป้ายให้องค์ประกอบที่เชื่อมต่อแบบ 8 ทิศทาง เพื่อหาขนาดของวัตถุแต่ละชิ้นในภาพ เมื่อระบุหมายเลขให้กับพิกเซลทั้งหมดแล้วนั้น หมายเลขอันดับสูงสุดหมายถึงจำนวนวัตถุทั้งหมดที่อยู่ภายในภาพ และจำนวนของพิกเซลที่ถูกกำหนดด้วยหมายเลขเดียวกัน หมายถึงจำนวนของวัตถุนั้น จากนั้นพิจารณาจำนวนพิกเซลที่ถูกกำหนดด้วยหมายเลขแต่ละตัว ถ้ามีค่าน้อยกว่า N พิกเซล วัตถุหมายเลขที่กำลังถูกพิจารณานั้นจะถูกกำจัดออก

จากรูป 3.6 ภาพต้นฉบับมีพิกเซลสัญญาณรบกวนที่ถูกระบุด้วยหมายเลข 670 ส่วนพิกเซลตัวอักษร ◌ู ถูกระบุเป็นหมายเลข 667 กำหนดค่า N เท่ากับ 30 สัญญาณรบกวนหมายเลข 670 มีจำนวน 6 พิกเซล ซึ่งน้อยกว่าค่า N ดังนั้นจะถูกกำจัดออกโดยการเปลี่ยนค่าให้เท่ากับ 0



ต้นฉบับ ผลลัพธ์

รูป 3.6 ตัวอย่างการกำจัดสัญญาณรบกวนด้วยการระบุป้ายให้องค์ประกอบที่เชื่อมต่อ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

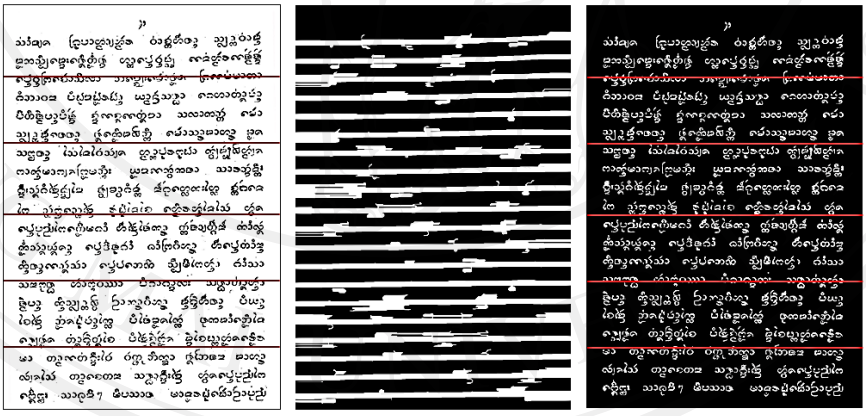
(3) การปรับความเอียงภาพ

ปัญหาเอกสารเอียงอาจเกิดจากขั้นตอนการสแกนเอกสาร หรือเอกสารต้นฉบับมีความเอียงอยู่แล้ว แม้ว่าภาพเอกสารมีความเอียงเพียง 1 องศา อาจส่งผลกระทบต่อการรู้จำตัวอักษร จึงต้องมีการปรับมุมเอียงเอกสารให้ตรง โดยแบ่งเป็น 2 ขั้นตอน คือ การหามุมเอียงของเอกสาร และการปรับมุมเอกสาร

1) การหามุมเอียงของเอกสาร

ในงานวิจัยนี้หามุมเอียงของเอกสาร โดยการปรับให้เรียบด้วยการนับความยาวเพื่อเชื่อมต่ออักษรที่อยู่ในบรรทัดเดียวกันให้รวมเป็นวัตถุชิ้นเดียวกัน ดังสมการ (2.6)

กำหนดให้ $I_{run-length}(x,y)$ คือภาพ $I_{bw}(x,y)$ ที่ถูกปรับให้เรียบด้วยการนับความยาว และ C เท่ากับ 50 เพื่อเป็นเงื่อนไขการปรับให้เรียบด้วยการนับความยาว จากนั้นใช้วิธีการระบุป้ายให้องค์ประกอบที่เชื่อมต่อ เพื่อหามุมเอียงของแต่ละบรรทัด แล้วหาตัวแทนของมุมเอียงสำหรับทั้งเอกสาร โดยหาได้จากค่ามัธยฐานมุมเอียงของทุกบรรทัดในหน้าเอกสาร ถ้ามุมเอียงที่ได้เป็นค่าบวก หมายถึง ภาพเอกสารเอียงไปในทิศทวนเข็มนาฬิกา ถ้ามุมเอียงเป็นค่าลบ หมายถึง ภาพเอกสารเอียงไปในทิศตามเข็มนาฬิกา



(ก) (ข) (ค)

รูป 3.7 ตัวอย่างการปรับให้เรียบด้วยการนับความยาวเพื่อหามุมเอียงของแต่ละบรรทัด

จากรูป 3.7 (ก) ภาพต้นฉบับเอียงไปในทิศทวนเข็มนาฬิกา เมื่อปรับให้เรียบด้วยการนับความยาว ดังรูป 3.7 (ข) และหาค่ามัธยฐานได้เท่ากับ 1.854 องศา

2) การปรับมุมเอกสาร

หลังจากได้มุมเอียงของภาพเอกสาร ภาพเอกสารจะถูกปรับความเอียงให้ตรงโดยการหมุนภาพในทิศทางตรงกันข้ามกับมุมที่หาได้

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

เมื่อมุมที่หาได้มีค่าเป็นลบ ต้องหมุนภาพในทิศทางทวนเข็มนาฬิกา ดังสมการ (3.2)

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3.2)$$

เมื่อมุมที่หาค่าได้มีค่าเป็นบวก ต้องหมุนภาพในทิศทางตามเข็มนาฬิกา ดังสมการ (3.3)

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3.3)$$

จากรูป 3.8 (ค) ภาพต้นฉบับถูกปรับมุมเอกสารให้ไปในทิศตามเข็มนาฬิกา 1.854 องศา ทำให้บรรทัดตัวอักษรในภาพเอกสารเอียงน้อยลงและง่ายต่อการตัดแบ่งบรรทัด ซึ่งเป็นขั้นตอนต่อไป

3.2.2 การตัดแบ่ง

หลังจากปรับปรุงคุณภาพของภาพให้เหมาะสมแล้ว ภาพจะถูกตัดแบ่งเพื่อสกัดให้ได้เฉพาะส่วนที่เป็นตัวอักษร โดยการตัดแบ่งประกอบด้วย 3 ขั้นตอน คือ การตัดแบ่งบรรทัด การตัดแบ่งบล็อกตัวอักษร และการหาจุดแบ่งระดับของบล็อกตัวอักษร มีรายละเอียดดังต่อไปนี้

(1) การตัดแบ่งบรรทัด

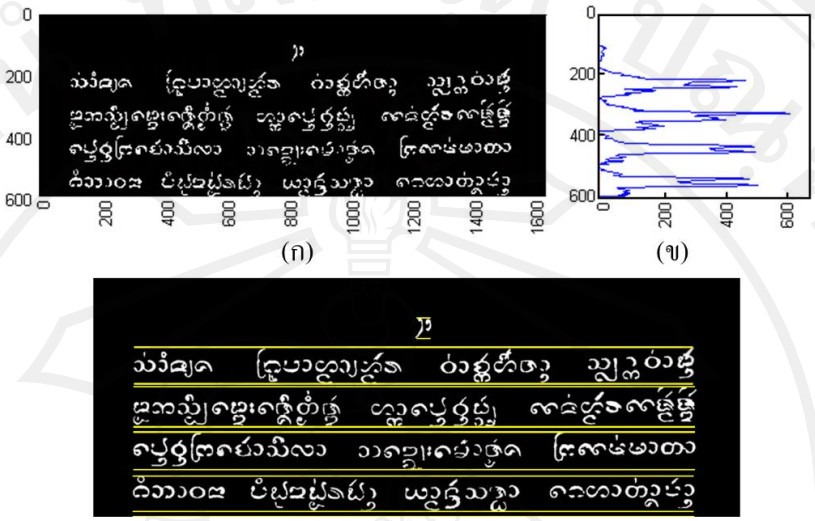
การตัดแบ่งบรรทัดในภาพเอกสารใช้วิธีโพรเจกชันโพรไฟล์ในแนวนอน (Horizontal projection profiles) ซึ่งเป็นการหาผลรวมของพิกเซลที่เป็นตัวอักษรหรือพิกเซลที่มีค่าเท่ากับ 1 ในแต่ละแถวของภาพเอกสาร โดยฮิสโตแกรมของโพรเจกชันโพรไฟล์ในแนวนอนจะเกิดส่วนที่มีค่าน้อยมากจนเข้าใกล้ค่า 0 สลับกับส่วนที่มียอดสูง ซึ่งส่วนที่มีค่าเข้าใกล้ค่า 0 คือบริเวณช่องว่างระหว่างบรรทัด และส่วนที่มียอดสูง คือส่วนของบรรทัดของตัวอักษร การสร้างกราฟฮิสโตแกรมของโพรเจกชันโพรไฟล์ในแนวนอนทำได้ดังสมการ (3.4)

$$H(y) = \sum_{i=1}^{I_{width}} I_{bw}(i,y) \quad (3.4)$$

โดยกำหนดให้ $H(y)$ เป็นโพรเจกชันโพรไฟล์ในแนวนอนแต่ละแถว ซึ่งการโพรเจกชันโพรไฟล์ในแนวนอน ทำให้ทราบตำแหน่งขอบเขตความสูงแต่ละบรรทัด และจำนวนบรรทัดในเอกสาร จากรูป 3.8 (ก) คือภาพต้นฉบับ รูป 3.8 (ข) คือฮิสโตแกรมของโพรเจกชันโพรไฟล์ใน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

แนวนอน โดยแกน x หมายถึง จำนวนพิกเซล และแกน y หมายถึง ตำแหน่งแถวของภาพเอกสาร และผลลัพธ์การตัดแบ่งบรรทัดแสดงดังรูป 3.8 (ค)



(ก) (ข)

(ค)

รูป 3.8 การตัดแบ่งบรรทัดด้วยโพรเจกชันโพรไฟล์ในแนวนอน

(2) การตัดแบ่งบล็อกตัวอักษร

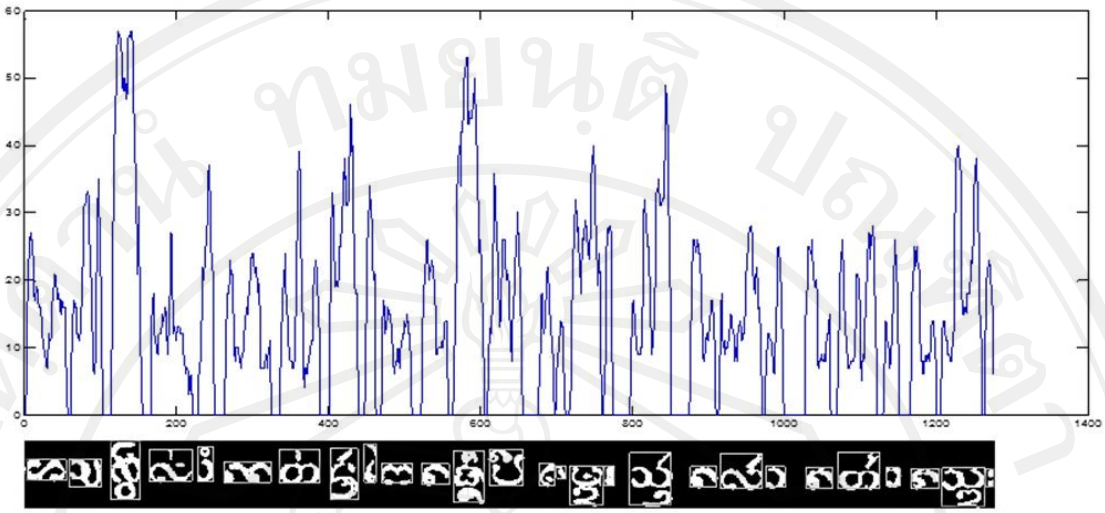
หลังจากการตัดแบ่งภาพเอกสารให้เป็นบรรทัดแล้ว ต่อมาเป็นการตัดแยกแต่ละบรรทัดให้เป็นตัวอักษรเพื่อเข้าสู่การจำแนก แต่ในงานวิจัยนี้จะไม่ตัดแยกให้เป็นอักษรเดี่ยว เนื่องจากการตัดแยกภาพอาศัยหลักการเชื่อมต่อกันของพิกเซลและช่องว่างรอบตัวอักษร ส่งผลให้ตัวอักษรที่มีลายเส้นขาดอาจถูกตัดแยกออกเป็นหลายส่วน ซึ่งนำไปสู่การรู้จำที่ผิดพลาด ในขั้นตอนนี้จึงเป็นการตัดแยกตัวอักษรให้เป็นบล็อก โดยการโพรเจกชันโพรไฟล์ในแนวตั้ง ซึ่งตำแหน่งที่มีค่าโพรเจกชันโพรไฟล์ในแนวตั้งเท่ากับ 0 จะเป็นตำแหน่งที่ถูกตัดแบ่ง ผลของการตัดแบ่งอาจได้อักษรเพียงตัวเดียว หรือเป็นกลุ่มของตัวอักษร ขอบเขตที่ตัดแบ่งได้ในขั้นตอนนี้เรียกว่าบล็อกตัวอักษร กราฟฮิสโตแกรมของโพรเจกชันโพรไฟล์ในแนวตั้งทำได้ดังสมการ (3.5)

$$VL_n(x) = \sum_{i=1}^{Ln_{height}} L_n(x,i) \tag{3.5}$$

ให้ VL_n เป็นฮิสโตแกรมของโพรเจกชันโพรไฟล์ในแนวตั้งแถวที่ x ของภาพเอกสารบรรทัดที่ n

จากรูป 3.9 เป็นตัวอย่างการตัดแบ่งบล็อกด้วยโพรเจกชันโพรไฟล์ในแนวตั้ง โดยส่วนที่เป็นมีค่าเป็น 0 ในแกน x คือตำแหน่งที่จะถูกตัดแบ่งเป็นบล็อก และกรอบสี่เหลี่ยมที่ล้อมรอบตัวอักษร คือผลลัพธ์การตัดแบ่งบล็อก ซึ่งพบได้ทั้งตัวอักษรเดี่ยวและกลุ่มตัวอักษร

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



รูป 3.9 การตัดแบ่งบล็อกตัวอักษรด้วยโพรเจกชันโพรไฟล์ในแนวตั้ง

(3) การหาจุดแบ่งระดับของตัวอักษร

การหาจุดแบ่งระดับของตัวอักษร คือ การหาขอบเขตบนและขอบเขตล่างของตัวอักษรระดับกลางแต่ละบรรทัด เพื่อแบ่งระดับของตัวอักษรแต่ละตัวที่อยู่ในบล็อกว่าอยู่ในระดับกลาง บน หรือล่าง ซึ่งมีขั้นตอนดังต่อไปนี้

1) ใช้วิธีโพรเจกชันโพรไฟล์ในแนวนอนสำหรับภาพเอกสารแต่ละบรรทัด ซึ่งหาได้จากสมการ (3.6) โดยกำหนดให้ HL_n เป็นฮิสโตแกรมของโพรเจกชันโพรไฟล์ในแนวนอนของภาพเอกสารแถวที่ y ของบรรทัดที่ n

$$HL_n(y) = \sum_{i=1}^{L_{n_{width}}} L_n(i,y) \tag{3.6}$$

2) หาค่าส่วนเบี่ยงเบนมาตรฐาน (σ) ของ HL_n ดังสมการ (3.7) โดยที่ $\overline{HL_n}$ เป็นค่าเฉลี่ยของ HL_n

$$\sigma = \sqrt{\frac{1}{L_{n_{height}}} \sum_{y=1}^{L_{n_{height}}} (HL_n(y) - \overline{HL_n})^2} \tag{3.7}$$

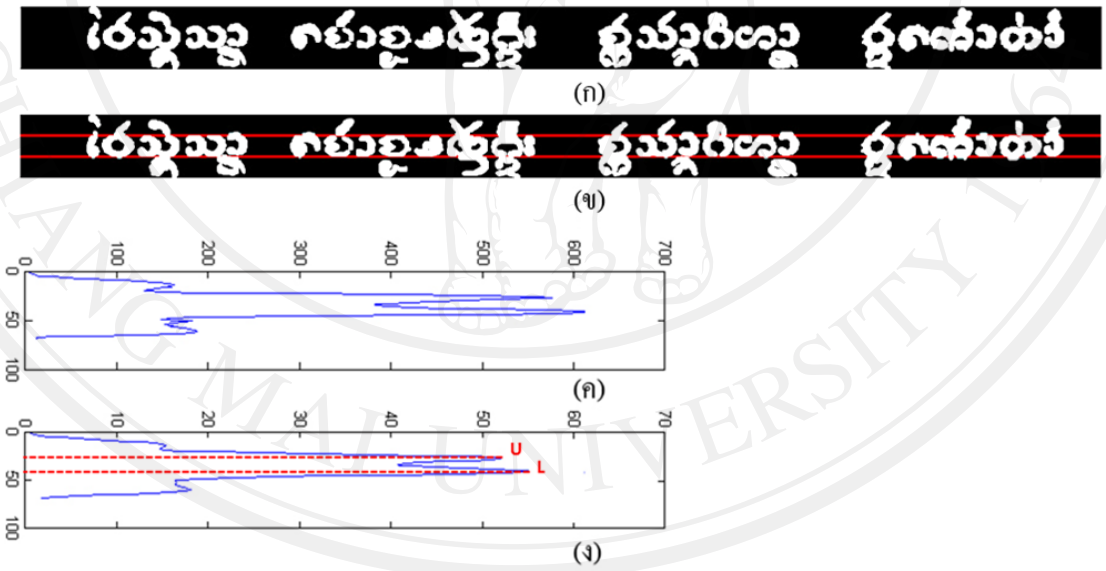
3) คำนวณหาค่าเกาส์เซียนของ HL_n เพื่อปรับฮิสโตแกรมของ HL_n ให้เรียบดังสมการ 3.8

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

$$g(HL_n) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} \cdot e^{\frac{HL_n^2}{2\sigma^2}} \quad (3.8)$$

4) คำนวณหาตำแหน่งของค่าสูงสุดเฉพาะที่ของภาพ L_n จำนวน 2 ค่า ซึ่งเป็นตำแหน่งขอบเขตบนและขอบเขตล่างของตัวอักษรที่อยู่ในระดับกลาง

ตัวอย่างการหาจุดแบ่งระดับของตัวอักษรแสดงดังรูป 3.10 โดยรูป 3.10 (ก) เป็นภาพบรรทัดของตัวอักษรธรรมล้านนาที่ต้องการหาจุดแบ่งระดับ เมื่อใช้วิธีโพรเจกชันโพรไฟล์ในแนวนอนของภาพบรรทัดได้ฮิสโตแกรมดังรูป 3.10 (ค) หลังจากนั้นปรับฮิสโตแกรมให้เรียบขึ้นด้วยค่าเกาส์เซียนได้ผลลัพธ์ดังรูป 3.10 (ง) ซึ่งค่าสูงสุดเฉพาะที่ที่มีค่ามากที่สุด 2 ค่า ในที่นี้ขอบเขตระดับบนคือ U มีค่าเท่ากับ 25 และขอบเขตล่าง คือ L มีค่าเท่ากับ 47 และสามารถกำหนดจุดเพื่อแบ่งระดับตัวอักษรได้ดังรูป 3.10 (ข)



รูป 3.10 การหาจุดแบ่งระดับของตัวอักษร

3.2.3 การสร้างชุดข้อมูลฝึกสอน

ในขั้นตอนการจำแนกประเภทตัวอักษร ตัวจำแนกประเภทจะเรียนรู้จากกลุ่มตัวอย่างในชุดข้อมูลฝึกสอน เพื่อให้จำแนกข้อมูลทดสอบเป็นประเภทต่างๆ ได้ ดังนั้นก่อนการจำแนกชุดข้อมูลทดสอบ จำเป็นต้องมีการสร้างชุดข้อมูลฝึกสอนที่ประกอบด้วยตัวอย่างของตัวอักษรแต่ละประเภท ในงานวิจัยนี้สร้างชุดข้อมูลฝึกสอนด้วยการจัดกลุ่มแบบเคมีน โดยแบ่งกลุ่มข้อมูลจากความคล้าย โดยไม่มีการกำหนดประเภทข้อมูลไว้ก่อน ซึ่งช่วยให้แยกหมวดหมู่ของตัวอักษรได้ใน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ระดับหนึ่ง จากนั้นผู้วิจัยจึงแยกประเภทให้ละเอียดและตรวจสอบความถูกต้องอีกครั้ง พร้อมกำหนดประเภทให้ข้อมูลแต่ละกลุ่ม

หลังจากการสร้างชุดข้อมูลฝึกสอน ได้แบ่งหมวดอักษรธรรมล้านนาตามหน้าที่และตำแหน่ง ซึ่งแบ่งเป็น 12 หมวด เพื่อสะดวกแก่การจำแนกประเภท มีรายละเอียดดังนี้

(1) พยัญชนะระดับกลาง (Consonant middle level : CM)

เป็นพยัญชนะที่อยู่ในระดับกลาง ซึ่งสามารถเทียบเคียงได้กับอักษรไทย ก-ฮ มีรูปที่พบ 32 ประเภท

(2) พยัญชนะซ้อนไม่เปลี่ยนรูป (Consonant lower level : CL)

เป็นพยัญชนะที่อยู่ใต้พยัญชนะต้น ทำหน้าที่เป็นตัวสะกด ซึ่งมีรูปเหมือนพยัญชนะระดับกลาง แต่มีขนาดย่อมกว่า มีรูปที่พบ 19 ประเภท

(3) พยัญชนะซ้อนเปลี่ยนรูป (Transformation consonant : TC)

เป็นพยัญชนะที่อยู่ใต้พยัญชนะตัวต้น ทำหน้าที่เหมือนพยัญชนะซ้อนไม่เปลี่ยนรูป แต่มีรูปที่เปลี่ยนไปไม่เหมือนพยัญชนะระดับกลาง หรือมีหางที่ยาวขึ้นไปเท่ากับพยัญชนะต้น มีรูปที่พบ 11 ประเภท

(4) อักษรพิเศษระดับกลาง (Special consonant : SC)

เป็นอักษรพิเศษที่อยู่ในระดับกลาง มีรูปที่พบ 7 ประเภท

(5) สระและอักษรพิเศษระดับบน (Upper vowel and special character : UV)

เป็นสระหรืออักษรพิเศษที่วางอยู่บนพยัญชนะระดับกลาง มีรูปที่พบ 9 ประเภท

(6) สระระดับกลาง (Middle vowel : MV)

เป็นสระที่วางระดับเดียวกับพยัญชนะระดับกลาง ซึ่งมีทั้งประเภทที่วางอยู่ด้านหน้าหรือด้านหลังพยัญชนะ มีรูปที่พบ 5 ประเภท

(7) สระระดับล่าง (Lower vowel : LV)

เป็นสระที่อยู่ใต้พยัญชนะ ซึ่งอาจอยู่ด้านใต้พยัญชนะต้น กรณีไม่มีพยัญชนะซ้อนหรือกรณีที่มีพยัญชนะซ้อนจะอยู่ใต้พยัญชนะซ้อน มีรูปที่พบ 2 ประเภท

(8) วรรณยุกต์ (Tone : To)

เป็นวรรณยุกต์กำกับเสียง ซึ่งวางอยู่บนพยัญชนะ มีรูปที่พบ 2 ประเภท

(9) ตัวเลขโหรา (Lanna Number : NL)

เป็นเลขแบบที่นิยมใช้ในการเขียนเอกสารทุกชนิด มีรูปที่พบ 10 ประเภท

(10) กลุ่มอักษรหลายระดับ (Group Character : GC)

เป็นกลุ่มอักษรที่ประกอบด้วยตัวอักษรตั้งแต่ 2 ตัว ซึ่งอาจเป็นได้ทั้งพยัญชนะ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ระดับกลาง หรืออักษรพิเศษระดับกลางที่อยู่ร่วมกับพยัญชนะซ้อน สระ หรือวรรณยุกต์ มีรูปที่พบ 1052 ประเภท

(11) กลุ่มอักษรระดับบน (Group Upper Character : GU)

เป็นกลุ่มอักษรที่ประกอบด้วยตัวอักษรตั้งแต่ 2 ตัว ที่มีเพียงสระระดับบนหรือวรรณยุกต์ มีรูปที่พบ 11 ประเภท

(12) กลุ่มอักษรระดับล่าง (Group Lower Character : GL)

เป็นกลุ่มอักษรที่ประกอบด้วยตัวอักษรตั้งแต่ 2 ตัว ที่มีเพียงสระระดับล่าง พยัญชนะซ้อนไม่เปลี่ยนรูป หรือพยัญชนะซ้อนเปลี่ยนรูป มีรูปที่พบ 10 ประเภท

ชุดข้อมูลฝึกสอนตัวพิมพ์อักษรธรรมล้านนาที่จัดกลุ่มได้ ยกเว้นตัวอักษรในหมวด GC GU และ GL เนื่องจากมีประเภทจำนวนมาก แสดงดังตาราง 3.1

ตาราง 3.1 ชุดข้อมูลฝึกสอนตัวพิมพ์อักษรธรรมล้านนา

ประเภท	อักษรธรรม	อัตราส่วนความกว้างต่อความสูงเฉลี่ย	จำนวนที่พบ	ประเภท	อักษรธรรม	อัตราส่วนความกว้างต่อความสูงเฉลี่ย	จำนวนที่พบ
CM01		1.88	1164	CM02		1.32	196
CM04		1.34	477	CM05		0.70	85
CM06		1.76	10	CM07		1.40	1794
CM08		1.34	574	CM10		1.35	418
CM11		0.71	73	CM13		1.00	156
CM14		1.37	878	CM15		1.35	30
CM18		1.80	82	CM19		1.75	693
CM20		1.37	173	CM21		1.36	538
CM22		1.38	349	CM23		1.37	466
CM24		1.30	1175	CM25		0.91	313
CM26		1.34	202	CM27		0.70	31

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตาราง 3.1 ชุดข้อมูลฝึกสอนตัวพิมพ์อักษรธรรมล้านนา (ต่อ)

ประเภท	อักษรธรรม	อัตราส่วนความกว้างต่อความสูงเฉลี่ย	จำนวนที่พบ	ประเภท	อักษรธรรม	อัตราส่วนความกว้างต่อความสูงเฉลี่ย	จำนวนที่พบ
CM28		1.38	1211	CM29		0.71	34
CM30		1.75	214	CM31		1.40	1725
CM32		1.76	307	CM33		1.36	2010
CM34		1.85	589	CM35		1.33	721
CM38		1.83	1339	CM39		1.84	611
CM40		0.74	1	CM41		1.34	802
CL01		1.76	182	CL02		1.31	126
CL04		1.31	224	CL07		1.44	779
CL08		1.30	232	CL14		1.22	38
CL15		1.34	11	CL18		1.77	18
CL19		1.62	126	CL21		1.58	23
CL22		1.40	12	CL33		1.38	349
CL34		1.79	80	CL35		1.36	400
CL36		1.42	22	CL38		1.78	26
CL39		1.82	9	LV01		1.02	395
LV02		1.09	284	MV01		0.50	636
MV02		0.86	3969	MV05		1.21	5888
MV06		2.21	1169	MV07		0.48	224
MV09		0.33	3348	Nl01		0.49	39
Nl02		0.35	67	Nl03		0.39	59

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตาราง 3.1 ชุดข้อมูลฝึกสอนตัวพิมพ์อักษรธรรมล้านนา (ต่อ)

ประเภท	อักษรธรรม	อัตราส่วนความกว้างต่อความสูงเฉลี่ย	จำนวนที่พบ	ประเภท	อักษรธรรม	อัตราส่วนความกว้างต่อความสูงเฉลี่ย	จำนวนที่พบ
Nl04		0.48	36	Nl05		0.47	37
Nl06		0.44	19	Nl07		0.58	97
Nl08		0.93	5	Nl09		0.49	19
Nl10		0.83	5	SC03		0.34	516
SC07		1.36	322	SC08		0.92	148
SC09		0.69	112	SC11		2.23	39
SC12		0.75	32	SC13		0.49	202
TC02		1.57	947	TC03		0.96	320
TC05		1.55	28	TC06		1.55	222
TC07		0.46	547	TC08		0.26	13
TC09		0.96	108	TC11		0.94	34
TC12		0.94	16	TC13		1.72	1195
TC14		0.90	30	To01		0.45	841
UV01		1.19	536	UV02		1.50	230
UV04		1.35	9	UV05		1.66	7
UV06		1.47	924	UV08		0.98	573
UV09		1.02	5	UV10		0.90	78
UV11		1.02	1978				

จากตาราง 3.1 พบว่า ตัวอักษรเดี่ยวส่วนมากมีความกว้างมากกว่าความสูง และจำนวน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ที่พบขึ้นอยู่กับความถี่ของการปรากฏในหน้าเอกสาร แต่มีบางตัวอักษรที่ความถี่ของการปรากฏในหน้าเอกสารมาก แต่ตัวอักษรมักติดกับตัวอักษรอื่น หรือมีลายเส้นขาด จึงทำให้มีจำนวนที่พบในชุดข้อมูลฝึกสอนน้อย

3.2.4 การจำแนกประเภทตัวอักษรด้วยเค-เนียร์เรสเนเบอร์

ในขั้นตอนนี้เป็นการจำแนกประเภทบล็อกตัวอักษร โดยใช้เค-เนียร์เรสเนเบอร์ ซึ่งมีรายละเอียดดังต่อไปนี้

(1) ประเภทอักษรธรรมล้านนาสำหรับการจำแนกด้วยเค-เนียร์เรสเนเบอร์

เนื่องจากในขั้นตอนนี้เป็นการแยกประเภทบล็อกตัวอักษร ซึ่งบล็อกต้องประกอบด้วยอักษรที่อยู่ในระดับกลางอย่างน้อย 1 ตัวอักษร ดังนั้นในชุดข้อมูลฝึกสอนจึงไม่พบตัวอักษรบางประเภท เช่น สระระดับบน วรรณยุกต์ สระระดับล่าง พยัญชนะซ้อน เป็นต้น โดยชุดข้อมูลฝึกสอนสำหรับการจำแนกด้วยเค-เนียร์เรสเนเบอร์มีทั้งหมด 6 หมวด รวม 1,108 ประเภท จำนวน 20,917 ตัวอย่าง แสดงดังตาราง 3.2

ตาราง 3.2 ประเภทอักษรธรรมล้านนาสำหรับการจำแนกด้วยเค-เนียร์เรสเนเบอร์

หมวด	CM	SC	MV	GC	NL	TC
จำนวน	32	7	6	1,052	10	1

(2) การสกัดคุณลักษณะสำหรับการจำแนกประเภทโดยเค-เนียร์เรสเนเบอร์

คุณลักษณะที่ใช้สำหรับการจำแนก คือ บิตเมตริกซ์ของบล็อกตัวอักษร โดยภาพบล็อกตัวอักษรที่ปรากฏในเอกสารมีขนาดที่หลากหลาย จึงต้องเปลี่ยนขนาดบล็อกตัวอักษรให้มีขนาดเท่ากัน โดยในที่นี้เปลี่ยนให้ทุกบล็อกมีความกว้าง 44 พิกเซล และสูง 44 พิกเซล เมื่อแต่ละบล็อกอยู่ในรูปเวกเตอร์คุณลักษณะจะมีความยาวเท่ากับ 1,936

(3) การจำแนกประเภทโดยเค-เนียร์เรสเนเบอร์

เค-เนียร์เรสเนเบอร์เป็นกระบวนวิธีที่ใช้ในการจำแนกประเภท โดยใช้หลักการพิจารณาหาค่าความแตกต่างระหว่างข้อมูลทดสอบกับชุดข้อมูลฝึกสอน ซึ่งกำหนดให้ $D = \{(x_1, c_1), (x_2, c_2), \ldots, (x_n, c_n)\}$ เป็นชุดข้อมูลฝึกสอน โดยที่ x_i คือเวคเตอร์คุณลักษณะของข้อมูล c_i คือประเภทของเวคเตอร์ x_i และ n คือ จำนวนข้อมูลฝึกสอน เมื่อ x' เป็นเวคเตอร์ของข้อมูลทดสอบที่ยังไม่ระบุประเภท สามารถทำนายหาประเภทของ x' ได้ดังขั้นตอนต่อไปนี้

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

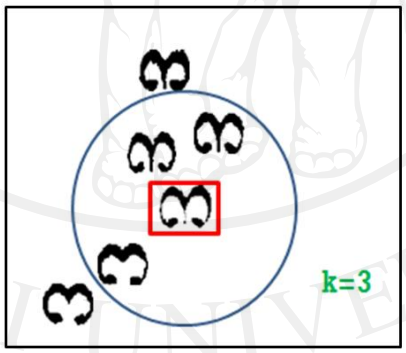
1) คำนวณหาระยะทางแบบยูคลิเดียนระหว่าง x' กับ x_i ทุกๆ ตัวในเซต D ดังสมการ (3.9)

$$dis(x_i, x') = \sqrt{\sum_{j=1}^{f} \left(x_{ij} - x'\right)^2} \quad (3.9)$$

2) หาเวคเตอร์ที่มีระยะทางน้อยที่สุด k อันดับแรก

3) ตรวจสอบว่าในเวคเตอร์ k อันดับดังกล่าวเป็นข้อมูลประเภทใดมากที่สุด ให้เวคเตอร์ x' จัดอยู่ในประเภทนั้น

ตัวอย่างการจำแนกประเภทด้วยเค-เนียร์เรสเนเบอร์ แสดงดังรูป 3.11 โดยตัวอักษรในกรอบสี่เหลี่ยม คือตัวอักษรที่ต้องการจำแนกประเภท เมื่อกำหนดให้ค่า k มีค่าเท่ากับ 3 จะพิจารณาเวคเตอร์ที่มีระยะทางน้อยที่สุด 3 อันดับแรก จากตัวอย่างนี้ผลของการจำแนกทำนายว่าอักษรนี้มีชนิดเป็น ๓ ไม่ใช่ ๓ เนื่องจากมีจำนวนข้อมูลฝึกสอนประเภท ๓ จำนวน 2 ตัว จากทั้งหมดที่พิจารณา 3 ตัว ซึ่งมากกว่าประเภท ๓ ที่มีเพียง 1 ตัว



รูป 3.11 ตัวอย่างการจำแนกประเภทด้วยเค-เนียร์เรสเนเบอร์

(4) การจัดเรียงลำดับตัวอักษรหลังจากการจำแนกประเภทด้วยเค-เนียร์เรสเนเบอร์

ผลลัพธ์ของการจำแนกด้วยเค-เนียร์เรสเนเบอร์ คือประเภทของบล็อกตัวอักษรที่จำแนกได้ และ $dis(x_i, x')$ ซึ่งเป็นระยะทางที่ใกล้ที่สุดระหว่างบล็อกตัวอักษรกับข้อมูลฝึกสอน x_i เนื่องจากลักษณะของเส้นตัวอักษรในภาพเอกสารไม่สมบูรณ์ ขาดจากกันหรือติดกัน อีกทั้งประเภทของชุดข้อมูลฝึกสอนไม่ครอบคลุมข้อมูลทดสอบทั้งหมด หากรู้จำด้วยวิธีเค-เนียร์เรสเนเบอร์เพียงขั้นตอนเดียว อาจมีความผิดพลาดสูง จึงมีการกำหนดให้ค่าเทรชโฮลด์ของระยะทางระหว่างบล็อกกับข้อมูลฝึกสอนคือ dis_{th} ถ้า $dis(x_i, x') \leq dis_{th}$ แสดงว่าผลการจำแนกประเภทนั้นมีโอกาส

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ผิดพลาดน้อย อักษรธรรมล้านนาที่จำแนกได้จะถูกตรวจสอบความถูกต้องด้วยคอนดิ-ชันนัลแรน-ดอมฟิลด์ แต่ถ้า $dis(x_i, x') > dis_{th}$ แสดงว่าผลการจำแนกประเภทนั้นมีโอกาสผิดพลาดสูง บล็อกอักษรนั้นจะถูกจำแนกซ้ำอีกครั้งด้วยเทมเพลตแมทชิง แล้วจึงตรวจสอบความถูกต้องของอักษรธรรมล้านนาที่จำแนกได้ด้วยคอนดิชันนัลแรนดอมฟิลด์

สำหรับทุกบล็อกที่มี $dis(x_i, x') \leq dis_{th}$ จะถูกจัดเรียงให้เป็นสายลำดับของตัวอักษรเดี่ยวต่อกันตามลำดับบรรทัด และตามลำดับบล็อกจากซ้ายไปขวา เพื่อจำแนกซ้ำด้วยคอนดิชันนัลแรนดอมฟิลด์ โดยแต่ละบล็อกจะเรียงตามรูปแบบการพิมพ์อักษรธรรมล้านนา ดังรูป 2.2 คือเรียงจาก ระดับที่ 1 ระดับที่ 2 ระดับที่ 3 ระดับที่ 4 และระดับที่ 5 สำหรับบล็อกที่เป็นกลุ่มตัวอักษรจะต้องมีการแปลงให้อยู่ในประเภทของตัวอักษรเดี่ยวโดยใช้ตารางสำหรับอ้างอิงระหว่างประเภทของบล็อกกับคุณลักษณะดังตาราง 3.3 จะทำให้ทราบอักษรเดี่ยวที่อยู่ภายในบล็อก พร้อมทั้งคุณลักษณะของอักษรนั้น 2 คุณลักษณะ คือ ระดับ และประเภทของตัวอักษรเดี่ยว เช่น ประเภท GC0036 คือ บล็อกกลุ่มอักษร ᨠ ประกอบด้วย ᩅ และ ᩁ อยู่ในระดับที่ 1 และ ◌ั อยู่ในระดับที่ 2 ตัวอย่างตารางอ้างอิงแสดงดังตาราง 3.3 เมื่อจัดเรียงลำดับตัวอักษรแล้วได้ผลลัพธ์เป็นสายลำดับประเภทของอักษรธรรมล้านนา คือ { MV09, CM14, To02} และได้ผลลัพธ์เป็นสายลำดับอักษรธรรมล้านนา คือ {ᩅ, ᩁ, ◌ั }

ตาราง 3.3 การอ้างอิงระหว่างประเภทของบล็อกและตัวอักษรเดี่ยวภายในบล็อก

ประเภท	ลำดับที่	ระดับที่ 1	ระดับที่ 2	ระดับที่ 3	ระดับที่ 4	ระดับที่ 5
GC0036	1	MV09	-	-	-	-
GC0036	2	CM14	To02	-	-	-

3.2.5 การจำแนกประเภทตัวอักษรด้วยเทมเพลตแมทชิง

บล็อกตัวอักษรจะถูกจำแนกซ้ำอีกครั้งด้วยเทมเพลตแมทชิง เมื่อระยะทางระหว่างบล็อกกับข้อมูลฝึกสอนมากเกินกว่าค่าที่ยอมรับได้ ซึ่งมีรายละเอียดดังต่อไปนี้

(1) ประเภทอักษรธรรมล้านนาสำหรับการจำแนกด้วยเทมเพลตแมทชิง

การจำแนกประเภทตัวอักษรด้วยเทมเพลตแมทชิง เรียกข้อมูลฝึกสอนแต่ละภาพว่าเทมเพลต และแบ่งชุดข้อมูลฝึกสอนเป็น 3 ระดับ คือ ระดับบน ระดับกลาง และระดับล่าง โดยระดับบนประกอบด้วย 3 หมวด จำนวน 21 ประเภท รวม 312 ตัวอย่าง ระดับกลางประกอบด้วย 7 หมวด จำนวน 66 ประเภท รวม 894 ตัวอย่าง และระดับล่างประกอบด้วย 4 หมวด จำนวน 29 ประเภท รวม 690 ตัวอย่าง ประเภทตัวอักษรแต่ละระดับแสดงดังตาราง 3.4

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตาราง 3.4 ประเภทอักษรธรรมล้านนาสำหรับการจำแนกด้วยเทมเพลตแมทชิง

ระดับบน	**หมวด**	GU	To	UV				**รวม**
	จำนวน	11	1	9				21
ระดับกลาง	**หมวด**	CM	GC	GL	MV	NL	SC	TC	**รวม**
	จำนวน	32	2	3	5	10	7	7	66
ระดับล่าง	**หมวด**	CL	GL	LV	TC				**รวม**
	จำนวน	19	4	2	4				29

(2) การสกัดคุณลักษณะสำหรับการจำแนกประเภทด้วยเทมเพลตแมทชิง

เนื่องจากเทมเพลตแมทชิงใช้การวัดความคล้ายระหว่างพิกเซลแต่ละจุดของเทมเพลตกับบล็อกตัวอักษร ดังนั้นในงานวิจัยนี้จึงใช้บิตเมตริกซ์ของเทมเพลตและบล็อกที่ไม่ถูกปรับให้เป็นขนาดมาตรฐาน เพื่อเป็นเมตริกซ์คุณลักษณะของข้อมูลฝึกสอนและข้อมูลทดสอบ

(3) การจำแนกประเภทด้วยเทมเพลตแมทชิง

การจำแนกประเภทด้วยเทมเพลตแมทชิง เป็นการนำภาพเทมเพลตไปค้นหาในภาพบล็อกตัวอักษร ด้วยการวัดความคล้ายทุกๆ ตำแหน่งจากซ้ายไปขวาและจากบนลงล่าง ทำให้ใช้เวลาในการคำนวณมาก ในงานวิจัยนี้จึงแบ่งระดับบล็อกตัวอักษรออกเป็น 3 ระดับ ได้แก่ ระดับบน ระดับกลาง และระดับล่าง เพื่อความรวดเร็ว และลดความผิดพลาดในการเปรียบเทียบความคล้ายระหว่างภาพ

การจำแนกประเภทด้วยเทมเพลตแมทชิงสำหรับบล็อกกลุ่มตัวอักษร ในงานวิจัยนี้ใช้ตัววัดแบบสัมประสิทธิ์สหสัมพันธ์ ดังสมการ (2.13) ซึ่งมีขั้นตอนดังนี้

1) เริ่มจากการจำแนกตัวอักษรในระดับกลางก่อน โดยนำเทมเพลตของตัวอักษรเดี่ยวทั้งหมดในชุดข้อมูลฝึกฝนระดับกลาง วัดความคล้ายทุกๆตำแหน่งบนภาพบล็อกตัวอักษรจากซ้ายไปขวา และจากบนลงล่างทั่วทั้งภาพ เพื่อค้นหาอักษรเดี่ยวระดับกลางที่ปรากฏอยู่ภายในบล็อกนั้น ประเภทของเทมเพลตที่มีค่าสัมประสิทธิ์สหสัมพันธ์มากที่สุด จะถูกกำหนดให้เป็นประเภทของตัวอักษรเดี่ยวในบล็อกนั้น ซึ่งการวัดแบบสัมประสิทธิ์สหสัมพันธ์นี้อาจมีกรณีที่เทมเพลตของตัวอักษรขนาดเล็กวัดความคล้ายกับตัวอักษรตัวอื่นภายในบล็อกแล้วมีค่าสัมประสิทธิ์สหสัมพันธ์มากกว่าประเภทของอักษรที่แท้จริง จึงมีการกำหนดเทรชโฮลด์สำหับค่าสัมประสิทธิ์สหสัมพันธ์ที่น้อยที่สุดที่ยอบรับได้ คือ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

0.5 และพิจารณาให้ความสำคัญกับพื้นที่ของเทมเพลตที่มีขนาดใหญ่มากกว่าขนาดเล็ก

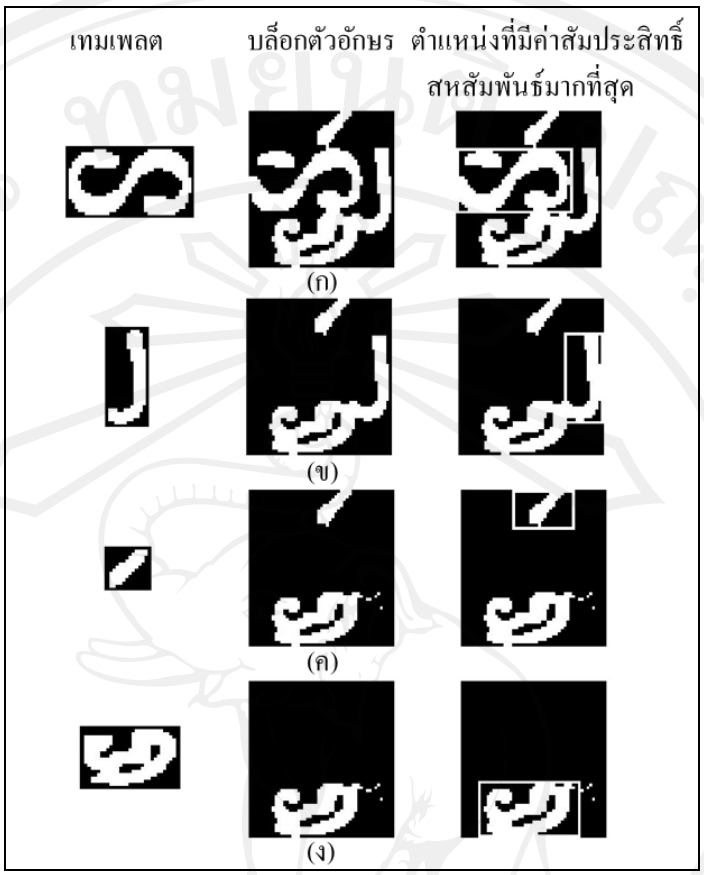
2) จากนั้นสกัดตัวอักษรเดี่ยวที่ถูกจำแนกแล้วในบล็อกนั้นออก และวนซ้ำเพื่อใช้เทมเพลตในชุดข้อมูลฝึกฝนระดับกลางหาค่าสัมประสิทธิ์สหสัมพันธ์กับตัวอักษรเดี่ยวตัวอื่น จนกว่าในส่วนระดับกลางของบล็อกไม่มีอักษรใดปรากฏอยู่ โดยทราบจุดแบ่งระดับของบล็อกตัวอักษร ได้จากการหาจุดแบ่งระดับของตัวอักษร ดังหัวข้อ (3) ในขั้นตอนการตัดแบ่ง

3) จากนั้นจำแนกอักษรในระดับบน เหมือนขั้นตอนที่ 1) และ 2) แต่ใช้เทมเพลตของตัวอักษรเดี่ยวทั้งหมดในชุดข้อมูลฝึกฝนระดับบน วัดความคล้ายทุกๆ ตำแหน่งบนภาพบล็อกตัวอักษรจากซ้ายไปขวา และจากขอบบนสุดจนถึงจุดแบ่งขอบเขตระดับบนเท่านั้น

4) จำแนกอักษรในระดับล่าง เหมือนขั้นตอนที่ 1) และ 2) แต่ใช้เทมเพลตของตัวอักษรเดี่ยวทั้งหมดในชุดข้อมูลฝึกฝนระดับล่าง วัดความคล้ายทุกๆ ตำแหน่งบนภาพบล็อกตัวอักษรจากซ้ายไปขวา และจากจุดแบ่งขอบเขตระดับล่างจนถึงขอบล่างสุดของภาพเท่านั้น

ซึ่งตัวอย่างการจำแนกบล็อกตัวอักษรด้วยเทมเพลตแมทชิง แสดงดังรูป 3.12 โดยในรูป 3.12 (ก) ข้อมูลฝึกฝนที่หาค่าสัมประสิทธิ์สหสัมพันธ์แล้วมีค่ามากที่สุด คือ 'ข' อยู่ในประเภท CM39 จากนั้นอักษร 'ข' ถูกสกัดออกจากบล็อก ซึ่งในบล็อกยังปรากฏอักษรอื่นในส่วนระดับกลาง จึงวนซ้ำเพื่อหาตัวอักษรอื่น ซึ่งในรูป 3.12 (ข) ข้อมูลฝึกฝนที่หาค่าสัมประสิทธิ์สหสัมพันธ์แล้วมีค่ามากที่สุด คือ 'ง' อยู่ในประเภท TC07 จากนั้นอักษร 'ง' ถูกสกัดออกจากบล็อก ซึ่งไม่ปรากฏอักษรใดอีกในส่วนระดับกลาง จึงจำแนกอักษรในระดับบนต่อ ในรูป 3.12 (ค) ข้อมูลฝึกฝนที่หาค่าสัมประสิทธิ์สหสัมพันธ์แล้วมีค่ามากที่สุด คือ ' ◌้ ' อยู่ในประเภท UV11 จากนั้นอักษร ' ◌้ ' ถูกสกัดออกจากบล็อก ซึ่งไม่ปรากฏอักษรใดอีกในส่วนระดับบน จึงจำแนกอักษรในระดับล่าง ในรูป 3.12 (ง) ข้อมูลฝึกฝนที่หาค่าสัมประสิทธิ์สหสัมพันธ์แล้วมีค่ามากที่สุด คือ ' ◌ฺ ' อยู่ในประเภท GD04 จากนั้นอักษร ' ◌ฺ ' ถูกสกัดออกจากบล็อก ซึ่งไม่ปรากฏอักษรใดอีกในส่วนระดับล่าง จึงสิ้นสุดการจำแนกบล็อก ผลลัพธ์ที่ได้ คือ อักษรระดับกลาง 2ประเภท คือ CM39 และ TC07 อักษรระดับบน 1 ประเภท คือ UV11 และอักษรระดับล่าง 1 ประเภท คือ GD04

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



รูป 3.12 ตัวอย่างการจำแนกบล็อกตัวอักษรด้วยเทมเพลตแมทชิง

(4) การจัดเรียงลำดับตัวอักษรหลังจากการจำแนกประเภทด้วยเทมเพลตแมทชิง

จากขั้นตอนที่ (3) ในแต่ละบล็อกจำแนกได้ผลลัพธ์ คือ สายลำดับของประเภทตัวอักษรเดี่ยวระดับกลาง ระดับบน และตามด้วยระดับล่าง ซึ่งประเภทตัวอักษรแต่ละระดับยังไม่ได้ถูกจัดเรียงให้ถูกต้องให้เป็นสายลำดับของตัวอักษรเดี่ยวต่อกันตามลำดับบรรทัด และตามลำดับบล็อกจากซ้ายไปขวา เพื่อจำแนกซ้ำด้วยคอนดิชันนัลแรนดอมฟิลด์ โดยแต่ละบล็อกจะเรียงตามรูปแบบการพิมพ์อักษรธรรมล้านนา ซึ่งมีขั้นตอนการจัดเรียงลำดับตัวอักษรของแต่ละบล็อกดังนี้

1) จัดเรียงลำดับอักษรระดับกลางก่อน โดยพิจารณาจากตำแหน่งขอบซ้ายสุดของตัวอักษรระดับกลางในภาพบล็อกนั้น เรียงตัวอักษรจากตัวอักษรที่มีตำแหน่งขอบซ้ายสุดที่มีค่าน้อยที่สุดไปมากที่สุด
2) จัดเรียงลำดับอักษรระดับบนของอักษรระดับกลางทีละตัว โดยพิจารณาจากตำแหน่งขวาสุด และซ้ายสุดของภาพอักษรระดับบนในภาพบล็อกนั้น ถ้าอักษรระดับบนมีตำแหน่งขอบขวาสุดน้อยกว่าตำแหน่งขอบขวาสุดของอักษรระดับกลางและตำแหน่งขอบซ้ายสุดมากกว่าตำแหน่งขอบซ้ายสุดของอักษร

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ระดับกลาง ให้อักษรระดับบนนั้นเป็นอักษรระดับบนของอักษรระดับกลางที่กำลังพิจารณา หากอักษรระดับกลางที่พิจารณามีอักษรระดับบน 2 ตัว ให้เรียงลำดับอักษรระดับบน โดยพิจารณาจากตำแหน่งขอบบนสุดของอักษรระดับบน ถ้าตำแหน่งขอบบนสุดมีค่าน้อยกว่า ให้เป็นอักษรระดับที่ 2 ถ้าตำแหน่งขอบบนสุดมีค่ามากกว่า ให้เป็นอักษรระดับที่ 3

3) จัดเรียงลำดับอักษรระดับล่างของอักษรระดับกลางทีละตัว เหมือนขั้นตอนที่ 2) แต่ในกรณีที่อักษรระดับกลางที่พิจารณามีอักษรระดับล่าง 2 ตัว ให้เรียงลำดับอักษรระดับล่าง โดยพิจารณาจากตำแหน่งขอบล่างสุดของอักษรระดับล่าง ถ้าตำแหน่งขอบล่างสุดมีค่าน้อยกว่า ให้เป็นอักษรระดับที่ 4 ถ้าตำแหน่งขอบล่างสุดมีค่ามากกว่า ให้เป็นอักษรระดับที่ 5

จากรูป 3.12 เมื่อจัดเรียงลำดับตัวอักษรแล้วได้ผลลัพธ์เป็นสายลำดับของประเภทตัวอักษร {‘CM39’, ‘UV11’ , ‘GD04’ , ‘TC07’} และเป็นสายลำดับอักษรธรรมล้านนา คือ {‘ᨷ’ , ‘ᩴ’ , ‘ᩣ’ , ‘᩠’}

3.2.6 การจำแนกประเภทตัวอักษรด้วยคอนดิชันนัลแรนดอมฟิลด์

การจำแนกประเภทตัวอักษรด้วยคอนดิชันนัลแรนดอมฟิลด์เป็นการแปลงปัญหาการรู้จำเป็นปัญหาการติดป้ายกำกับให้ลำดับข้อมูล โดยมีรายละเอียดดังต่อไปนี้

(1) ประเภทอักษรธรรมล้านนาสำหรับการจำแนกด้วยคอนดิชันนัลแรนดอมฟิลด์

เนื่องจากการจำแนกด้วยคอนดิชันนัลแรนดอมฟิลด์ได้รับข้อมูลนำเข้าจากการจำแนกด้วยเค-เนียร์เรสเนเบอร์และเทมเพลตแมทชิง จึงมีประเภทอักษรธรรมล้านนาทั้งหมด 12 หมวด 116 ประเภท แสดงดังตาราง 3.5

ตาราง 3.5 ประเภทอักษรธรรมล้านนาสำหรับการจำแนกด้วยคอนดิชันนัลแรนดอมฟิลด์

หมวด	CM	CL	MV	UV	LV	To	TC	SC	GC	GU	GL	NL
จำนวน	32	19	5	9	2	1	11	7	2	11	7	10

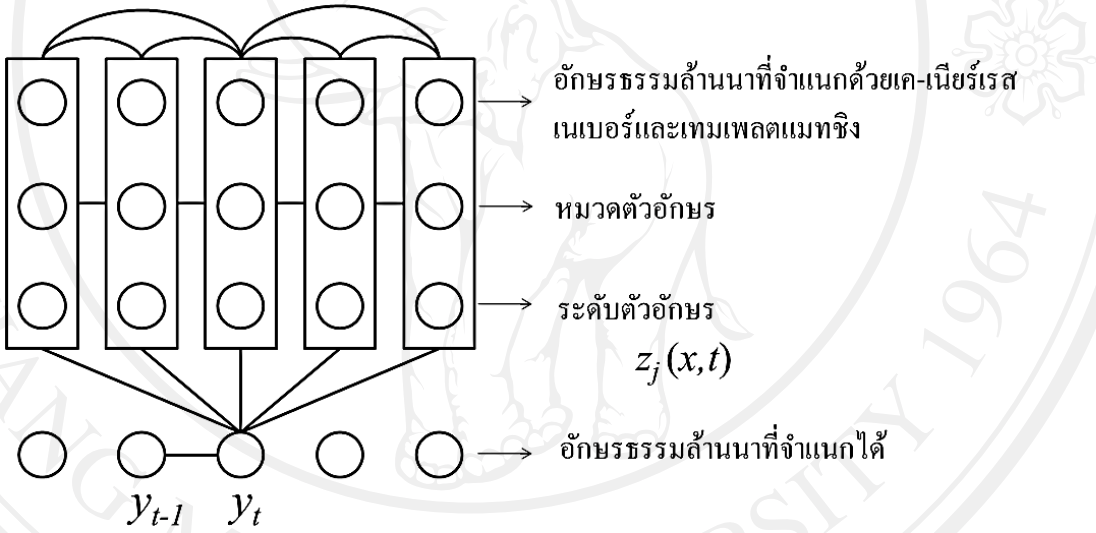
(2) การสกัดคุณลักษณะสำหรับการจำแนกประเภทด้วยคอนดิชันนัลแรนดอมฟิลด์

ในงานวิจัยนี้ใช้ 2 คุณลักษณะ คือ หมวดของตัวอักษร และระดับของตัวอักษร ที่เป็นผลลัพธ์จากขั้นตอนการจำแนกด้วยเค-เนียร์เรสเนเบอร์ และเทมเพลตแมทชิง

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

(3) การจำแนกประเภทด้วยคอนดิชันนัลแรนดอมฟิลด์

ในขั้นตอนการจำแนกประเภทคอนดิชันนัลแรนดอมฟิลด์นี้ ใช้ตัวแบบคอนดิชันนัลแรนดอมฟิลด์แบบห่วงโซ่ตรง ตามสมการแจกแจงความน่าจะเป็นแบบมีเงื่อนไข ดังสมการ 2.16 โดยกำหนดให้ x เป็นสายลำดับอักษรธรรมล้านนาที่จำแนกได้จากเค-เนียร์เรสเนเบอร์และเทมเพลตแมทชิง และ y เป็นสายลำดับผลลัพธ์ที่ถูกติดป้ายเพื่อระบุประเภทอักษร โดยตัวแบบคอนดิชันนัลแรนดอมฟิลด์จะทำนายประเภทของอักษรโดยการหาความน่าจะเป็นที่มากที่สุดจากตัวแบบฝึกฝน ตัวแบบเชิงรูปภาพของคอนดิชันนัลแรนดอมฟิลด์สำหรับจากจำแนกประเภทอักษรธรรมล้านนาแสดงดังรูป 3.13 จากตัวอย่างพิจารณาในช่วง 2 ตัวอักษร ทั้งด้านหน้าและด้านหลังตัวอักษรที่ต้องการจำแนก ซึ่งใช้รูปแบบการพิจารณานี้ทั้งในชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ



รูป 3.13 ตัวแบบของคอนดิชันนัลแรนดอมฟิลด์สำหรับจากจำแนกประเภทอักษรธรรมล้านนา

โปรแกรมที่ใช้ในการจำแนกประเภทตัวอักษรด้วยคอนดิชันนัลแรนดอมฟิลด์ คือ ซีอาร์เอฟพลัสพลัส ซึ่งเป็นโปรแกรมที่ได้รับการยอมรับว่ามีประสิทธิภาพ และถูกนำไปใช้ในงานวิจัยที่เกี่ยวข้องกับการติดป้ายสายลำดับข้อมูล ตัวอย่างผลลัพธ์การจำแนกด้วยซีอาร์เอฟพลัสพลัส แสดงดังรูป 3.14 โดยหลักที่หนึ่ง คือสายลำดับของอักษรธรรมล้านนาที่จำแนกได้จากเค-เนียร์เรสเนเบอร์และเทมเพลตแมทชิงซึ่งต้องการจำแนกซ้ำ หลักที่สอง คือระดับของตัวอักษร หลักที่สาม คือหมวดของตัวอักษร หลักที่สี่ คือสายลำดับอักษรธรรมล้านนาที่ถูกต้อง และหลักที่ห้า คือผลลัพธ์จากการจำแนกด้วยโปรแกรมซีอาร์เอฟพลัสพลัส

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

อักษรที่จะจำแนกซ้ำ	ระดับของตัวอักษร	หมวดตัวอักษร	อักษรธรรมล้านนาที่ถูกต้อง	ผลลัพธ์จากการจำแนกด้วยโปรแกรมซีอาร์เอฟพลัสพลัส
ᨻ	1	CM	ᨻ	ᨻ
ᩮ	2	UV	ᩮ	ᩮ
ᨦ	1	CM	ᨦ	ᨦ
ᩣ	4	LV	ᩣ	ᩣ
ᨶ	1	CM	ᨶ	ᨶ
ᨯ	1	CM	ᨯ	ᨯ
ᩁ	2	UV	ᩁ	ᩁ
ᨿ	1	CM	ᨿ	ᨿ
ᩴ	2	UV	ᩴ	ᩴ
ᩉ	1	MV	ᩉ	ᩉ
ᩮ	2	UV	ᩮ	ᩮ
ᨠ	1	CM	ᨠ	ᨠ
ᩴ	2	UV	ᩴ	ᩴ
ᩈ	4	TC	ᩈ	ᩈ

รูป 3.14 ตัวอย่างผลลัพธ์การจำแนกด้วยโปรแกรมซีอาร์เอฟพลัสพลัส

3.3 การรู้จำตัวพิมพ์อักษรธรรมล้านนาด้วยซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนเป็นกระบวนวิธีจำแนกประเภทที่ถูกใช้งานอย่างแพร่หลายและใช้ในการรู้จำตัวอักษรในหลายภาษา ซึ่งจัดเป็นเทคนิคที่ใช้ในการแก้ปัญหาทางด้านการรู้จำรูปแบบข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด ข้อมูลที่นำมาใช้ในการสอนให้ระบบเรียนรู้ ส่วนใหญ่มักเป็นข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยการนำเคอร์เนลฟังก์ชันมาใช้

ในงานวิจัยนี้ใช้ไลบรารี LIBSVM สำหรับโปรแกรมแมทแล็บ เพื่อจำแนกตัวพิมพ์อักษรธรรมล้านนา โดยเริ่มจากการสร้างตัวแบบจากชุดข้อมูลฝึกสอน จากนั้นนำไฟล์แบบจำลองดังกล่าวไปใช้ในขั้นตอนการทดสอบการจำแนกตัวอักษรต่อไป ซึ่งชุดข้อมูลฝึกสอน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

และเวกเตอร์คุณลักษณะเป็นแบบเดียวกับการจำแนกด้วยเค-เนียร์เรสเนเบอร์ และเปรียบเทียบผลการทดลองกับกระบวนการรู้จำตัวพิมพ์อักษรธรรมล้านนาที่ได้นำเสนอไปข้างต้น ซึ่งกระบวนการรู้จำตัวพิมพ์อักษรธรรมล้านนาที่กล่าวไปแล้วนั้นจะถูกทดสอบ และนำเสนอผลการวิจัยในบทที่ 4 ต่อไป

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved