บทที่ 3

การทดสอบระบบการรู้จำ

3.1. บทนำ

การทดสอบการรู้จำสายอักษรตัวพิมพ์ไทยนั้น ตัวอักษรที่ใช้ทดสอบอาจจะเป็นประโยค คำ หรือตัวอักษรที่เรียงต่อกัน ซึ่งอาจมีหลายประโยคหรือหลายคำ ข้อความเหล่านั้นจะผ่านเครื่อง Scanner เป็นข้อมูลภาพที่มีลักษณะเป็น Bitmap เพื่อนำไปให้โปรแกรมทำการวิเคราะห์ต่อไป ดังนั้น ส่วนประกอบทางด้านฮาร์ดแวร์ในการรู้จำประกอบด้วย

1. เครื่อง Scanner ใช้สำหรับ scan ภาพที่เป็นตัวอักษรเข้ามาเก็บไว้ในเครื่อง โดยจะเก็บเป็นไฟล์ข้อมูลแบบ BMP โดยเครื่อง Scanner ที่ใช้ในงานวิจัยครั้งนี้ เป็นเครื่อง Scanner ยี่ห้อ Epson รุ่น GT-6500 ความละเอียดในการ scan 300 DPI

2. เครื่อง Computer ในงานวิจัยนี้ใช้เครื่องคอมพิวเตอร์แบบ Notebook ยี่ห้อ Compaq รุ่น LITE ความเร็วของ CPU 33 MHz

สำหรับเอกสารหรือข้อความที่ใช้ทดสอบนั้น ได้มาจากการใช้โปรแกรม Microsoft Word for Windows 6.0 แล้วพิมพ์ออกทางเครื่องพิมพ์ LASER ยี่ห้อ Hewlette Package รุ่น III+ ความละเอียด 300 dpi นำเอกสารที่ได้ไปทำการอ่านกลับมาเก็บไว้เป็นไฟล์ข้อมูลภาพแบบ BMP ด้วยเครื่อง Scanner สำหรับรูปแบบของอักษรที่ใช้สำหรับการทดสอบในงานวิจัยครั้งนี้ใช้ Font แบบ EucrosiaUPC ขนาด 18 point

3.2. ตัวอักษรต้นแบบ

สำหรับตัวอักษรต้นแบบที่ใช้ในการวิจัยครั้งนี้ ใช้สำหรับการเปรียบเทียบกับข้อมูลภาพที่ต้องการนำมาเปรียบเทียบ ซึ่งมีอยู่ด้วยกัน 2 แบบคือ

3.2.1. ตัวอักษรต้นแบบสำหรับการแยกตัวอักษร เป็นตัวอักษรต้นแบบที่ใช้สำหรับการแยกข้อมูลตัวอักษรออกจากกัน ซึ่งข้อมูลที่ใช้เป็นตัวอักษรต้นแบบนั้นประกอบด้วยข้อมูลต่าง ๆ ดังนี้

3.2.1.1. ขนาดความกว้างของตัวอักษร

3.2.1.2. ขนาดความสูงของตัวอักษร

3.2.1.3. ข้อมูลที่เป็นรูปสัญญลักษณ์ของตัวอักษรต้นแบบ

3.2.2. ตัวอักษรต้นแบบสำหรับการรู้จำตัวอักษร เป็นตัวอักษรต้นแบบที่ใช้สำหรับการรู้จำตัวอักษร ซึ่งจะเก็บรายละเอียดต่าง ๆ ของตัวอักษรต้นแบบไว้ ซึ่งประกอบด้วย

3.2.2.1. เส้นฐานของตัวอักษร

3.2.2.2. เส้นขอบบนของตัวอักษร

3.2.2.3. โครงสร้างของตัวอักษร

3.2.2.4. รหัสแอสกีของตัวอักษรต้นแบบ

ซึ่งตัวอักษรที่ใช้สำหรับการเป็นตัวอักษรต้นแบบในการแยกตัวอักษรและสำหรับการรู้จำตัวอักษรนั้น ใช้ตัวอักษรแบบ EucrosiaUPC ขนาด 56 point พิมพ์บนโปรแกรม Microsoft Windows Paintbrush Version 3.1 แล้วจัดเก็บลงแฟ้มข้อมูลแบบ Bitmap File เนื่องจากตัวอักษรขนาดนี้สามารถเก็บรายละเอียดของขนาดของตัวอักษรที่ scan เข้ามาได้

3.3. ข้อมูลที่ใช้ในการทดสอบ

ข้อมูลที่ใช้ในการทดสอบเป็นข้อมูลที่ทดสองสร้างขึ้นเพื่อใช้ในการทดสอบ โดยพิมพ์ผ่าน โปรแกรม Microsoft Word for Windows 6.0 สำหรับรูปแบบของตัวอักษรที่ใช้ทดสอบในงานวิจัยนี้ใช้ Font แบบ EucrosiaUPC ขนาด 18 point ดังที่ได้กล่าวมาแล้ว การทดสอบจะแบ่งข้อความที่ใช้ในการทดสอบออกเป็นประเภทต่าง ๆ ดังนี้

3.3.1. ข้อความที่ประกอบด้วยตัวอักษรที่อยู่ในระดับกลาง ซึ่งตัวอักษรที่อยู่ในระดับนี้สำหรับ font แบบ EucrosiaUPC แบ่งได้ดังนี้

3.3.1.1. font ที่เป็นตัวอักษร ได้แก่ ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ฑ ฒ ณ ด ต ถ ท ธ น บ ผ พ ภ ม ย ร ล ว ศ ษ ส ห อ ฮ

3.3.1.2. font ที่เป็นสระ ได้แก่ ะ า แ เ ฯ

3.3.1.3. font ที่เป็นตัวเลขไทย ได้แก่ ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙

.33.1.4. font ที่เป็นตัวเลขอารบิค ได้แก่ 0 1 2 3 4 5 6 7 8 9

สำหรับตัวอย่างข้อความหรือกลุ่มคำที่ใช้ทดสอบได้แก่ “ธนาคารและธนาคาร”

3.3.2. ข้อความที่ประกอบด้วยตัวอักษรที่อยู่ในระดับกลางและระดับบน ซึ่งเป็นการผสมตัวอักษรกันระหว่างตัวอักษรในข้อ 3.3.1. และตัวอักษรในระดับบน ตัวอักษรในระดับบนนี้ประกอบด้วยตัวอักษรต่าง ๆ ดังนี้

สำหรับตัวอย่างข้อความหรือกลุ่มคำที่ใช้ทดสอบได้แก่ "อัชระการพิมพ์"

3.3.3. ข้อความที่ประกอบด้วยตัวอักษรที่อยู่ในระดับกลางและตัวอักษรที่อยู่ในระดับล่าง ผสม กัน ซึ่งตัวอักษรที่อยู่ในระดับล่างได้แก่ ุ ู ญ ฐ เป็นต้น

สำหรับตัวอย่างข้อความหรือกลุ่มคำที่ใช้ทดสอบได้แก่ "กรุงเทพมหานคร"

3.3.4. ข้อความที่ประกอบด้วยตัวอักษรที่อยู่ในระดับกลางและตัวอักษรที่เกินกว่าเส้นขอบบน ผสมกัน ซึ่งตัวอักษรที่เกินกว่าเส้นขอบบนได้แก่

4.3.4.1. ตัวพยัญชนะ ได้แก่ ป ฝ ฟ ฬ

4.3.4.2. สระ ได้แก่ ไ โ ใ

สำหรับตัวอย่างข้อความหรือกลุ่มคำที่ใช้ทดสอบได้แก่ "ปราบปราม"

3.3.5. ข้อความที่ประกอบด้วยตัวอักษรที่อยู่ในระดับกลางและตัวอักษรที่เกินกว่าเส้นขอบบน ผสมกัน ซึ่งตัวอักษรที่ต่ำกว่าเส้นขอบล่างได้แก่ ฎ ฏ ฤ ฦ

สำหรับตัวอย่างข้อความหรือกลุ่มคำที่ใช้ทดสอบได้แก่ "กฎหมาย"

3.3.6. ข้อความที่ประกอบด้วยตัวอักษรทุก ๆ แบบ โดยการรวมตัวอักษรที่อยู่ในข้อ 3.3.1 - 3.3.5 ไว้ด้วยกันทั้งหมด

สำหรับตัวอย่างข้อความหรือกลุ่มคำที่ใช้ทดสอบได้แก่

"เป็นมนุษย์สุดประเสริฐเลิศคุณค่า กว่าบรรดาฝูงสัตว์เดรัจฉาน

จงฝ่าฟันพัฒนาวิชาการ อย่าล้างผลาญฤาเข่นฆ่าบีฑาใคร

ไม่ถือโทษโกรธแช่งซัดฮึดฮัดด่า หัดอภัยเหมือนกีฬาอัชฌาสัย

ปฏิบัติประพฤติกฎกำหนดใจ พูดจาให้ จ๊ะ ๆ จ๋า ๆ น่าฟังเอยฯ"

ซึ่งเป็นสำนวนมาตรฐานสำหรับตรวจสอบอักขระตัวพิมพ์ภาษาไทยของสมาคมคอมพิวเตอร์แห่งประเทศไทยในพระบรมราชูปถัมภ์

3.4. วิธีการทดสอบ

3.4.1. การทดสอบการแยกข้อมูลภาพ เป็นการทดสอบในส่วนของการแยกข้อมูลภาพออกเป็นตัวอักษรแต่ละตัว ซึ่งตัวอักษรที่นำมาทดสอบจะมีทั้งตัวอักษรธรรมดาที่สามารถนำไปทำการรู้จำได้เลย และตัวอักษรที่ติดกัน รวมถึงตัวอักษรที่มีจุดภาพของตัวอักษรข้างเคียงด้วย โดยตัวอักษรที่ใช้ทดสอบนั้นได้มาจากการพิมพ์ออกทางเครื่องพิมพ์เลเซอร์ ด้วยโปรแกรม Microsoft Word for Windows จากนั้นทำการ scan เอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์เพื่อเก็บไว้เป็นไฟล์ข้อมูลแบบ BMP และทำการตกแต่งข้อมูลภาพที่ได้นั้นให้เกิดกรณีตัวอักษรติดกันหรือตัวอักษรที่มีจุดภาพของตัวอักษรข้างเคียงด้วยโปรแกรม Microsoft Paintbrush เพื่อทำการทดสอบในส่วนของการแยกข้อมูลภาพและการแยกตัวอักษรที่ติดกัน

3.4.2. การทดสอบระบบการรู้จำ การทดสอบระบบจะนำเอาข้อความหรือกลุ่มคำที่ได้กำหนดไว้มาพิมพ์ออกทางเครื่องพิมพ์เลเซอร์ ด้วยโปรแกรม Microsoft Word for Windows จากนั้นทำการ scan เอกสารที่พิมพ์จาก เครื่องพิมพ์เลเซอร์เพื่อเก็บไว้เป็นไฟล์ข้อมูลแบบ BMP ซึ่งในการ scan เอกสารนั้นจะ scan เอกสารในแบบ 2 ระดับ ซึ่งจะมีความเข้มของสีเพียง 2 ระดับ คือความเข้มของฉากและความเข้มของวัตถุเท่านั้น สำหรับในกรณีที่ต้องการ scan เอกสารในลักษณะอื่น เช่น scan เอกสารในลักษณะ 16 ระดับ ก็จะนำเอาแฟ้มของเอกสารที่ได้จากการ scan ไปทำให้เป็นภาพที่มีความเข้มเพียง 2 ระดับเท่านั้น โดยการกำหนดค่า Threshold ที่ใช้นั้นจะมีค่าเท่ากับระดับความเข้มที่ 50 % (ตัวอย่างการทำภาพให้เหลือความเข้มเพียง 2 ระดับดูได้ในภาคผนวก ฉ.) และนำเอาไฟล์ข้อมูลที่ได้มานั้นทำการทดสอบผ่านระบบที่ได้พัฒนาขึ้น ระบบจะทำการหาค่าต่าง ๆ โดยผลลัพธ์ที่ได้จะเป็นข้อความที่เก็บอยู่ในแฟ้มข้อมูล (สำหรับตัวอย่างประโยคและผลลัพธ์ที่ได้ในแต่ละขั้นตอนของการรู้จำสามารถดูได้ใน ภาคผนวก จ.) จากนั้นนำแฟ้มข้อมูลที่ได้นี้ไปทำการเปรียบเทียบกับแฟ้มข้อมูลต้นแบบที่เป็นผลลัพธ์ของแฟ้มข้อมูลแบบ BMP ที่ได้จัดทำขึ้นเพื่อใช้สำหรับการตรวจสอบผลลัพธ์ที่ได้จากการวิเคราะห์ตัวอักษร โดยภายหลังจากการที่ได้เปรียบเทียบแล้ว ทำให้ทราบได้ว่าข้อมูลที่เราป้อนให้กับระบบนั้นระบบสามารถวิเคราะห์ได้ถูกต้องมากน้อยเพียงไร รวมถึงปริมาณตัวอักษรในการวิเคราะห์ และปริมาณตัวอักษรที่ถูกต้อง ปริมาณตัวอักษรที่วิเคราะห์ไม่ได้ และปริมาณตัวอักษรที่ไม่สามารถรู้จำได้ ซึ่งตัวอักษรที่ใช้ในการทดสอบจะแบ่งออกเป็นทั้งหมด 5 กลุ่มแต่ละกลุ่มประกอบด้วย คำต่างๆ หรือ ประโยคต่าง ๆ รวมกัน 30 สายอักษร โดยเมื่อรวมตัวอักษร ทั้งหมดแล้วจะมีตัวอักษรทั้งหมด 1974 ตัว แบ่งออกได้ดังนี้

3.4.1. ตัวพยัญชนะ 1239 ตัว

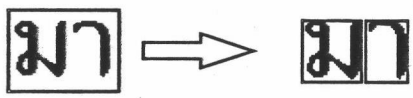
3.4.2. สระและวรรณยุกต์ 695 ตัว

3.4.3. ตัวเลขไทยและตัวเลขอารบิค 40 ตัว

3.5. ผลการทดสอบ

3.5.1. ผลการทดสอบในส่วนของการแยกข้อมูลภาพ จากการทดสอบโดยป้อนตัวอักษรเข้าไปเพื่อตรวสอบการแยกข้อมูลภาพ โดยมีตัวอย่างภาพอยู่ 3 ลักษณะ ผลการทดสอบเป็นดังนี้

3.5.1.1. สำหรับตัวอักษรธรรมดา การแยกตัวอักษรสามารถกระทำได้ดังแสดงในรูปที่ 3.1 โดยผลลัพธ์ที่ได้จะเป็นตำแหน่งของตัวอักษรแต่ละตัว

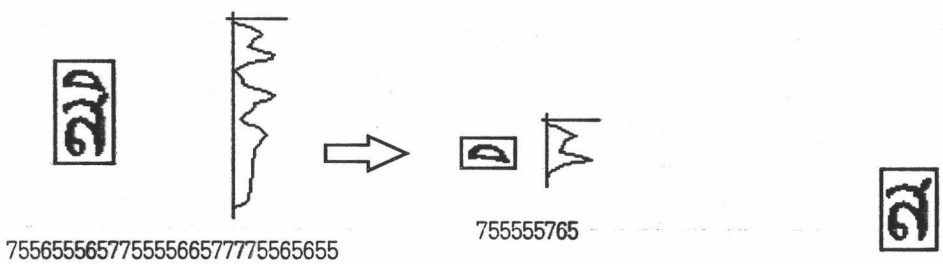


x1	y1	x2	y2
1	1	26	32
27	1	52	32

รูปที่ 3.1 แสดงผลลัพธ์ที่ได้จากการแยกตัวอักษร

3.5.1.2. สำหรับตัวอักษรที่ติดกัน การตรวจสอบตัวอักษรที่ติดกัน บางครั้งจะได้ตัวอักษรธรรมดาที่มีความกว้างมาก ๆ เช่น ฒ ณ เนื่องจากการกำหนดค่าความกว้างของตัวอักษรที่ติดกันทางแนวนอนน้อยกว่าตัวอักษรเหล่านั้น แต่เมื่อนำไปทำการแยกตัวอักษรแล้ว การแยกตัวอักษรสามารถตรวจได้ว่าเป็นตัวอักษรธรรมดา ไม่ต้องมีการแยกออกเป็นตัวอักษร 2 ตัว แต่สำหรับตัวอักษรที่ต้องแยกออกเป็น 2 ตัวนั้น บางครั้งไม่สามารถแยกออกจากกันได้ เนื่องจากการเลื่อมล้ำกันของตัวอักษร 2 ตัวซึ่งเกิดขึ้นกับตัวอักษรที่ติดกันทางแนวตั้ง เช่น ป ฟ เป็นต้น โดยผลลัพธ์ที่ได้ในลักษณะต่าง ๆ เป็นดังนี้

3.5.1.2.1. กรณีตัวอักษรติดกันทางแนวนอน แสดงดังรูปที่ 3.2



(ก) (ข) (ค)

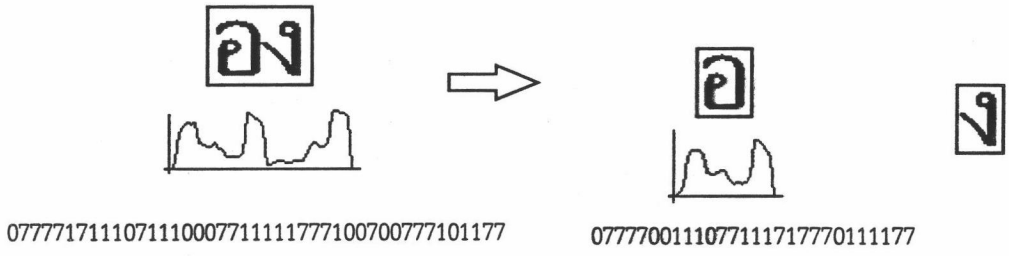
รูปที่ 3.2 การแยกตัวอักษรทางแนวตั้ง

(ก) ตัวอักษรที่ต้องการนำมาแยกและรูปสัญญลักษณ์ที่ได้

(ข) ตัวอักษรตัวแรกที่ได้ มีรูปสัญญลักษณ์ต่างจากตัวอักษรต้นแบบน้อยที่สุด

(ค) ตัวอักษรที่ได้หลังจากแยกตัวอักษรตัวแรกแล้ว

3.5.1.2.2. กรณีตัวอักษรติดกันทางแนวตั้ง แสดงดังรูปที่ 3.3



(ก) (ข) (ค)

รูปที่ 3.3 การแยกตัวอักษรทางแนวนอน

(ก) ตัวอักษรที่ต้องการนำมาแยกและรูปสัญญลักษณ์ที่ได้

(ข) ตัวอักษรตัวแรกที่ได้ มีรูปสัญญลักษณ์ต่างจากตัวอักษรต้นแบบน้อยที่สุด

(ค) ตัวอักษรที่ได้หลังจากแยกตัวอักษรตัวแรกแล้ว

3.5.1.2.3. กรณีตัวอักษรติดกันที่แยกแล้วทำให้เกิดการผิดพลาด แสดงดังรูปที่ 3.4.

(ก) (ข) (ค)

รูปที่ 3.4 การแยกตัวอักษรติดกันที่ผิดพลาด

(ก) ตัวอักษรที่ติดกัน (ข) ตัวอักษรตัวแรกที่ได้ (ค) ตัวอักษรตัวที่สองที่ได้

3.5.1.2.4. กรณีความกว้างของตัวอักษรมากกว่าค่าที่ตั้งไว้ แต่เป็นตัวอักษรปกติ แสดงดังรูปที่ 3.5

(ก) (ข)

รูปที่ 3.5 ตัวอักษรที่มีขนาดมากกว่าความกว้างของตัวอักษรปกติที่กำหนด

(ก) ตัวอักษรที่ต้องการนำไปแยก (ข) ตัวอักษรที่ได้

ลักษณะของข้อมูล	จำนวนกลุ่มคำ	จำนวนตัวอักษรทั้งหมด	จำนวนตัวอักษรที่รู้จำ	จำนวนตัวอักษรที่ผิด		% ความถูกต้อง
				รู้จำผิด	รู้จำไม่ได้	
แบบที่ 3.3.1	30	347	318	21	8	91.64
แบบที่ 3.3.2	30	447	420	17	10	93.96
แบบที่ 3.3.3	30	225	209	4	12	92.88
แบบที่ 3.3.4 แบบที่ 3.3.5	30	212	189	13	10	89.15
แบบที่ 4.3.6	30	743	694	32	17	93.40
รวม	150	1974	1830	87	57	92.70

ตารางที่ 3.1 ผลลัพธ์ของการทดสอบระบบรู้จำสายอักษรตัวพิมพ์ไทย

3.5.2. ผลการทดสอบระบบการรู้จำ จากการทดสอบโปรแกรมโดยป้อนกลุ่มคำหรือข้อความต่าง ๆ ให้กับโปรแกรมที่ได้พัฒนาขึ้น ซึ่งจะแบ่งแยกข้อมูลเป็นประเภทต่าง ๆ ดังที่ได้กล่าวมาแล้วนั้น ได้ผลการทดสอบออกมาดังตารางที่ 3.1 ซึ่งจากข้อความที่ใช้ในการทดสอบนั้นมีตัวอักษรทั้งหมด 1974 ตัวอักษร มีความถูกต้องในการรู้จำ 1830 ตัวอักษร มีการรู้จำผิดพลาด 87 ตัวอักษร ไม่สามารถรู้จำได้ 57 ตัวอักษร โดยเมื่อคิดอัตราส่วนของตัวอักษรที่ไม่สามารถรู้จำได้ต่อตัวอักษรทั้งหมด พบว่าประโยคหรือกลุ่มคำในแบบที่ 3.3.4 และแบบที่ 3.3.5 จะมีการรู้จำผิดพลาดและไม่สามารถรู้จำได้มากที่สุด

สำหรับตัวอย่างคำและประโยคที่ใช้ในการทดสอบระบบการรู้จำและผลการทดสอบนั้นอยู่ในภาคผนวก ง.

จากการทดสอบตัวอักษรที่ต้องการรู้จำซึ่งได้แบ่งไว้เป็น 5 ประเภทนั้น ผลปรากฏดังนี้

ตัวอักษรที่รู้จำผิดของตัวอักษรในประเภทที่ 1 ได้แก่ ค ต บ ย ว ศ ส เ ๅ ๗ และตัวอักษรที่ไม่สามารถรู้จำได้ ได้แก่ ข ท น ร ว

ตัวอักษรที่รู้จำผิดของตัวอักษรในประเภทที่ 2 ได้แก่ ก ษ ส ◌ิ ◌ั ◌์ ◌่ และตัวอักษรที่ไม่สามารถรู้จำได้ ได้แก่ ศ ส ◌์

ตัวอักษรที่รู้จำผิดของตัวอักษรในประเภทที่ 3 ได้แก่ ค ษ ส า และตัวอักษรที่ไม่สามารถรู้จำได้ ได้แก่ ญ ฏ ณ น า ุ ั

ตัวอักษรที่รู้จำผิดของตัวอักษรในประเภทที่ 4 และ 5 ได้แก่ ค ย ษ ห ฤ า ั ิ และตัวอักษรที่ไม่สามารถรู้จำได้ ได้แก่ ก ป ฏ ฤ ไ

ตัวอักษรที่รู้จำผิดของตัวอักษรในประเภทที่ 6 ได้แก่ ค น ษ ห ิ ้ และตัวอักษรที่ไม่สามารถรู้จำได้ ได้แก่ ฆ ณ ฟ ป ใ ั ็ ้ ุ ์ ๆ

จากผลการทดลองปรากฏว่าตัวอักษรที่รู้จำผิดส่วนใหญ่จะเป็นตัวอักษร ในระดับบนที่มีลักษณะใกล้เคียงกันและตัวอักษรในระดับกลางที่มีลักษณะใกล้เคียงกัน ทั้งนี้เนื่องจากเมื่อผ่านการ scan ภาพเข้ามาแล้วรายละเอียดของตัวอักษรบางส่วนอาจผิดเพี้ยนไปบ้างหรือหายไป ทำให้กลายเป็นตัวอักษรตัวอื่นที่มีลักษณะใกล้เคียงกันได้ ซึ่งตัวอักษรที่รู้จำผิดส่วนมากได้แก่ ็ ็ ิ ส ล ค ศ ษ น ส่วนตัวอักษรที่ไม่สามารถรู้จำได้นั้น ส่วนมากเกิดขึ้นกับตัวอักษรที่ไม่สามารถแยกออกจาก กันได้ในกรณีที่ตัวอักษรคู่นั้นติดกัน เนื่องจากขั้นตอนการแยกตัวอักษรไม่สามารถแยกออกมาได้ เพราะว่า ตัวอักษรที่ติดกันนั้นเป็นตัวอักษรที่เกินไปในระดับบน กับตัวอักษรในระดับบน เช่น ป + ้ = ป้ , ฟ + ้ = ฟ้, ป + ็ = ป็ เป็นต้น และเกิดจากการรู้จำไม่สามารถวิเคราะห์ได้ว่าเป็นตัวอักษรอะไร