บทที่ 2
แนวคิด ทฤษฎีและผลงานที่ผ่านมา

การรู้จำอักษร (Character Recognition) เป็นส่วนหนึ่งของการรู้จำรูปแบบ(Pattern Recognition) ซึ่งในระบบรู้จำอักษรที่สมบูรณ์นั้น S. Tsujimoto et. al. (1992) เสนอว่าประกอบด้วยขั้นตอนหลัก ๆ 3 ขั้นตอนดังนี้

1. การวิเคราะห์เอกสาร (Document Analysis) จะทำการวิเคราะห์ว่ามีตาราง รูปภาพ บรรทัดของอักษร แบบของอักษร ขนาดของอักษร อยู่ที่ใดตำแหน่งใด ทำการแยก รูปภาพ ตาราง ออกแล้วนำเกาะของส่วนที่เป็นภาพอักษรซึ่งอาจเป็นอักษรเดี่ยวหรืออักษรที่ติดกันก็ได้ส่งให้ส่วนตัดแยกอักษรที่ติดกัน
2. การตัดแยกอักษรที่ติดกัน (Segmentation of Connected Characters) นำภาพอักษรมาตรวจว่าเป็นอักษรที่ติดหรือไม่แล้วตัดแยกจนเป็นตัวอักษรเดี่ยวส่งไปให้ส่วนรู้จำตัวเดี่ยว
3. การรู้จำตัวอักษรเดี่ยว (Recognition of Single Character) ทำการรู้จำภาพของอั กษรเดี่ยวนำผลที่ได้ส่งกลับให้ส่วนวิเคราะห์เอกสารเพื่อสร้างแฟ้มเอกสาร

การประยุกต์ใช้ทฤษฎีของการรู้จำรูปแบบนั้นสามารถให้ได้กับงานทุกส่วนยกตัวอย่างเช่น

- เรามีภาพอักษรอยู่ 4 บรรทัด ต้องการหาจุดตัดแบ่งระหว่างบรรทัด เราทำการนับจุดในแนวนอนแล้วนำไปเขียนกราฟ เราสามารถหาจุดตัดได้โดยหาส่วนที่เป็นหุบเขา การหาว่ามีหุบเขาอยู่บริเวณใดของกราฟนี้เป็นตัวอย่างของการรู้จำรูปแบบ
- เรามีภาพของอักษรที่ติดกันอยู่จำนวนหนึ่งให้เป็นชุดสำหรับสอน เราทำการนับจุดในแนวดิ่งแล้วนำไปเขียนกราฟ เราสอนให้ระบบรู้จำรูปแบบว่าถ้ากราฟเป็นแบบนี้แล้วให้ตัดที่จุดนี้ เมื่อเรามีภาพอักษรที่ติดกันใหม่เข้ามาสามารถหาจุดที่จะตัดโดยส่งให้ระบบรู้จำรูปแบบ
- เรามีภาพอักษร " ก - ฮ " เราสอนให้ระบบรู้จำรูปแบบว่าถ้าภาพ "ก" เข้ามาให้ตอบเป็นรหัส ASCII 163 เมื่อคราวหลังมีภาพอักษรคล้าย "ก" เข้ามาให้ตอบว่าเป็นรหัส ASCII 163

<u>ทฤษฎีในการรู้จำ</u>

ทฤษฎีในการรู้จำรูปแบบ Robert J. Schalkoff (1992) สรุปแนวทางไว้ดังนี้

1. Statistical Pattern Recognition (StatPR)
2. Syntactic Pattern Recognition (SyntPR)
3. Neural Pattern Recognition (NeurPR)

แล้วเปรียบเทียบแนวทางไว้ดังตารางที่ 2.1

หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 2.1 เปรียบเทียบแนวทาง StatPR SyntPR และ NeurPR

	StatPR	SyntPR	NeurPR
1. Pattern Generation (Storing) Basis	Probabilistic Models	Formal Grammars	Stable State or Weight Array
2. Pattern Classification (Recognition /Description) Basis	Estimation/Decision Theory	Parsing	Based on (Predictable) Properties of NN
3. Feature Organization	Feature Vector	Primitives and Observed Relations	Neural Input or Stored States
4. Typical Learning (Training) Approaches *Supervised:* *Unsupervised:*	 Density/distribution estimation (usually parametric) Clustering	 Forming grammars (heuristic or grammatical interference) Clustering	 Determining NN system parameters (e.g. ,weights) Clustering
5. Limitations	Difficulty in Expressing structural information	Difficulty in learning structure rules	Often little semantic information from network

หลักการของ StatPR คือการที่เราใช้ค่าทางสถิติของข้อมูลที่เรามี เช่น ค่าเฉลี่ย (mean) ค่าเบี่ยงเบนมาตรฐาน (Standard diviation) เพื่อใช้การแยกกลุ่ม ความยากของวิธีนี้อยู่ที่หา Feature Vector ที่เหมาะสม เช่นการหาความแตกต่างระหว่างลูกระเบิดกับส้ม ข้อมูลที่เรามี เช่น ขนาดความกว้างหรือสูง น้ำหนัก ความหนาแน่นปริมาตร เราจะเลือกข้อมูลใดใช้ในการตัดสินใจว่าของที่มีน้ำหนักเท่านี้ หนาแน่นเท่านี้เป็นระเบิดหรือส้ม ซึ่งจะต้องสอดคล้องกับการเลือก Estimation/Decision Theory ซึ่งมีหลายแบบ เช่น k-NN , Maximum Likehood และยังต้องเลือกว่าเป็นการเรียนรู้แบบ Supervised (มีผู้ดูแลในการแบ่งกลุ่ม) หรือ Unsupervised

หลักการของ NeurPR คือการที่เรานำข้อมูลมาใส่ให้ระบบโครงข่ายประสาทเทียมเรียนรู้ด้วยการปรับค่าตัวคูณภายในระบบหรือการทำให้ระบบเสถียร ความยากของการใช้งานระบบนี้เช่นเดียวกับ StatPR คือการเลือกข้อมูลที่จะสอนให้ระบบ จากนั้นคือเลือกวิธีต่อเชื่อมของหน่วยย่อยของโครงข่ายประสาทเทียม ซึ่งมีให้เลือกมากมาย เช่น แบบข้อมูลเข้าทางเดียวออกทางเดียว ไปโดยไม่มีการย้อนกลับ มีการควบคุมจากภายนอก และยังต้องเลือกว่าเป็นแบบแบ่งกลุ่มด้วยคนหรือให้ระบบแบ่งกลุ่มให้โดยอัตโนมัติ ข้อมูลที่เราจะสอนให้ระบบนั้นถ้าเลือกการต่อไม่เหมาะสมระบบจะเรียนรู้ได้ไม่ดีหรือไม่ได้เลย เช่นเราต้องการให้ระบบเราทำหน้าที่เหมือนตัวดำเนินการ xor (คือ มีข้อมูลเข้า 2 ตัวออก 1 ตัวโดย ถ้าต่างกันให้ผลเป็นจริง ถ้าเหมือนกันให้เป็นเท็จ) ถ้าใช้การเชื่อมต่อแบบ feed forward คือไปทางเดียวและไม่มี hidden layer จะไม่สามารถรู้จำได้เลย

หลักการของ SyntPR คือการที่เราหา Grammar ที่สามารถสร้างสมาชิกในกลุ่มได้ เมื่อเรามีสมาชิกใหม่จะดูว่าอยู่ในกลุ่มนี้หรือไม่ เราก็นำไป parsing กับ Grammar ของกลุ่มนั้นๆ ถ้าเป็นสมาชิก ก็จะเข้ากับ Grammar นั้นได้ ยกตัวอย่างเช่นเรามีอักษร 5 ตัวคือ ก ถ บ ผ พ เราแบ่งเป็น 2 กลุ่มคือ ก ถ และ บ ผ พ เราตั้ง Grammar ของกลุ่มแรกเป็น มีเส้นขึ้นบน + เส้นไปทางขวา + เส้นลงล่าง ส่วนกลุ่มหลังเป็น มีเส้นลงล่าง + เส้นไปทางขวา + เส้นขึ้นบน เมื่อเรามี ภ เข้ามา ถามว่าตรงกับ Grammar ใด จะได้คำตอบว่าตรงกับกลุ่มแรก

ความยากของระบบนี้คือหา Grammar ที่เหมาะสม การหาส่วนประกอบที่แยกไม่ได้อีกแล้ว (primitive) ที่จะมาประกอบเป็น Grammar และการหาวิธีแจงส่วน (parsing) ที่สามารถครอบลุ่ม Grammar นั้นได้ ข้อดีของวิธีนี้คือสามารถใช้ความสัมพันธ์กันระหว่าง primitive ช่วยในการแบ่งแยกกลุ่มของข้อมูลที่เข้ามา ซึ่งตรงกับลักษณะของอักษรภาษาไทยที่มีการลากเส้นอักษรต่อกันจะเป็นรูปร่างขึ้นมา ซึ่ง primitive คือ เส้นตรง เส้นโค้ง หัว ฯลฯ โดย Grammar จะแสดงกฎการสร้างอักษรนั้นๆ ขึ้นมา

<u>ทฤษฎีที่เกี่ยวข้องกับลักษณะบ่งความต่าง</u>

"ลักษณะบ่งความต่างมีบทบาทสำคัญคือเป็นเครื่องมือในการแยกสมาชิกของระบบใดระบบหนึ่งออกจากกัน แนวคิดนี้สามารถใช้ได้กับระบบใดๆ ไม่ว่าจะเป็นสรรพสิ่งต่างๆ เช่นแยกหน่วยเสียงหนึ่งออกจากระบบหน่วยเสียง" (Brasington , R. 1994)

การที่จะนำลักษณะบ่งความต่างมาใช้ได้อย่างมีประสิทธิภาพนั้นจำเป็นต้องมีทฤษฎีเป็นพื้นฐาน ทฤษฎีที่เกี่ยวข้องโดยตรงคือ "ทฤษฎีปฏิลักษณ์" (Opposition Theory) (Ogden 1967)

สุดาพร (2529) ได้นำทฤษฎีปฏิลักษณ์ซึ่งความสัมพันธ์ในลักษณะตรงข้ามแบบสมบูรณ์ (Opposition by Cut) ของ Ogden (1967) และความตรงข้ามแบบสัมพัทธ์ (เทียบได้กับ Opposition by scale ของ Ogden ,1967) มาใช้ในการอธิบายความหมายของคำว่าขึ้นกับลงทั้งในการแง่ความหมายอ้างอิงและความหมายทางทัศนคติและความรู้สึก

Ogden , 1967 ได้แบ่งประเภทของปฏิลักษณ์เป็น 3 อย่างคือ โดยการตัด(by Cut) โดยปริมาณ(by Scale) และ โดยนิยาม(by Definition) โดยยกตัวอย่างดังนี้ โดยการตัด เช่น ซ้าย-ขวา เห็น-ไม่เห็น ขึ้น-ลง โดยปริมาณ เช่น ดำ-ขาว แข็ง-นุ่ม และโดยนิยาม เช่น คน-สัตว์

Trubetzkoy 1971 ได้กล่าวถึงการแบ่งประเภทของปฏิลักษณ์ไว้โดยใช้พื้นฐานต่างๆ กัน 3 แบบได้แก่

1. แบ่งโดยโดยพื้นฐานของความสัมพันธ์กันระหว่างปฏิลักษณ์ของระบบทั้งหมด ได้แก่ Multilateral และ Bilateral , Isolated และ Proportional

ปฏิลักษณ์ไม่เพียงบ่งความต่างระหว่างสมาชิกในกลุ่ม แต่แสดงถึงสมบัติที่เหมือนกันของสมาชิกในกลุ่มด้วย ปฏิลักษณ์ที่มีลักษณะ Bilateral คือจะต้องบ่งความต่างระหว่าง 2 สิ่งในกลุ่มโดยส่วนที่เหมือนกันของ 2 สิ่งนั้น ไม่มีคู่ใดในกลุ่มมีคู่ที่มีจุดเหมือนกันเหมือนกับคู่นี้อีกแล้ว เช่น อักษร E กับ F ในกลุ่มของอักษรลาติน อธิบายได้โดยจุดร่วมของ 2 อักษรนี้คือเส้นในแนวดิ่ง 1 เส้นและเส้นในแนวนอนชี้ไปทางขวา 2 เส้นโดยเส้นหนึ่งยื่นจากบส่วนบนของเส้นแนวตั้งและอีกเส้นยื่นจากส่วนกลางของเส้นแนวตั้ง ซึ่งลักษณะเช่นนี้ไม่เกิดกับอักษรลาตินอื่นใดอีก ส่วนปฏิลักษณ์ที่มีลักษณะ Multilateral คือ มีสมาชิกที่มีลักษณะร่วมมากกว่า 2 อักษรเช่น P และ R เหมือนกันที่เส้นแนวตั้งและเส้นโค้งที่ลากจากจุดบนของเส้นแนวตั้งถึงจุดกลางของเส้นแนวตั้งซึ่งมีอักษรที่มีลักษณะร่วมเหมือนกันคืออักษร B

ปฏิลักษณ์ที่มีความสำคัญไม่น้อยกว่า Multilateral และ Bilateral คือ Isolated และ Proportional โดยปฏิลักษณ์จะเรียกว่า Proportional เมื่อความสัมพันธ์ระหว่างสมาชิกของมันเหมือนกับความสัมพันธ์ระหว่างสมาชิกของปฏิลักษณ์อื่นๆ ในระบบเดียวกัน ตัวอย่างการปฏิลักษณ์ในคู่เสียง p-b ในภาษาเยอรมันเป็น Proportional เพราะความสัมพันธ์ระหว่าง p และ b นั้นเหมือนกันกับความสัมพันธ์ระหว่าง t และ d หรือระหว่าง k และ g ถ้าไม่มีคู่ที่มีความสัมพันธ์เหมือนกันเรียกปฏิลักษณ์นั้นว่า Isolated ยกตัวอย่างอักษรภาษาไทยเช่นปฏิลักษณ์ของอักษร ด-ต เป็น Proportional เพราะสัมพันธ์กันเหมือนกับ ค-ท แต่ความสัมพันธ์ระหว่าง ฏ-ฎ เป็น Isolated เป็นต้น

ลักษณะความเป็นปฏิลักษณ์ของสมาชิกในระบบของภาษานั้น อาจมีได้ต่างๆ กัน บางภาษามีความสัมพันธ์แบบ multilateral มากกว่า bilateral หรือมีความสัมพันธ์แบบ isolated มากกว่า proportional เช่น ระบบเสียงภาษาเยอรมัน แต่ภาษารัสเซียและภาษาพม่ามีความสัมพันธ์แบบ proportional มากกว่าแบบ isolated อย่างชัดเจนมาก

2. แบ่งโดยความสัมพันธ์ระหว่างสมาชิกของปฏิลักษณ์ โดยแบ่งเป็น Privative , Gradual และ Equipollent
   - Privative แสดงเป็นแบบ "มี" หรือ "ไม่มี" เช่น ออกเสียง หรือ ไม่ออกเสียง เป็นต้น
   - Gradual แสดงเป็นปริมาณ เช่น เสียงวรรณยุกต์
   - Equipollent เป็นการแบ่งสมาชิก 2 ตัวด้วยเกณฑ์หนึ่งๆ เช่น เสียงในภาษาเยอรมัน p - t และ f - k หรืออักษรไทยเช่น ข - บ ใช้ความกว้างเป็นเกณฑ์ ซึ่งจะพบมากในระบบทั่ว ๆ ไป ซึ่งสามารถเทียบได้กับการแบ่งของ Ogden (1967) คือ Private เหมือนแบบ by Cut แบบ Gradual เหมือนแบบ by Scale และแบบ Equipollent เหมือนแบบ by Definition

3. แบ่งโดยการขยายของแรงในการแยกแยะ แบ่งเป็น Constant และ Neutralizable

ตัวอย่าง Constant เช่น ในภาษา Danish æ และ e เกิดขึ้นในทุกตำแหน่ง พวกมันเป็นปฏิลักษณ์แบบ Constant ซึ่งสมาชิกเป็น Phoneme อิสระ สำหรับแบบ Neutralizable ตัวอย่างในภาษาฝรั่งเศส โดย e และ ɛ เกิดขึ้นเพียงแค่ตอนท้ายของพยางค์เปิดเป็นสมาชิกของปฏิลักษณ์ "les"/ "lait" / "allez" / "allait" สำหรับตำแหน่งอื่นๆ การเกิดของ e และ ɛ สามารถทำนายได้ คือ ɛ เกิดในพยางค์ปิด ส่วน e ในพยางค์เปิด สระ 2 ตัวนี้จะต้องแยกใช้เป็น Phoneme 2 ตัวในกรณีของพยางค์สุดท้ายเปิด ส่วนในกรณีที่อยู่ตำแหน่งอื่นๆ ถือเป็น 1 Phoneme ได้ สำหรับตัวอย่างเสียงในภาษาไทยที่เห็นได้ชัด เช่น เสียงของอักษรแม่ "กบ" เช่น อักษร บ ป พ ฟ ภ เมื่ออยู่ในตำแหน่งอักษรนำจะออกเสียงต่างกัน ซึ่งสามารถแยกด้วยลักษณะบ่งความต่างทางเสียง แต่เมื่ออยู่ในตำแหน่งตัวสะกดจะออกเสียงเหมือนกันทั้งหมดไม่ต้องแยกด้วยลักษณะบ่งความต่างทางเสียง

ทฤษฎีปฏิลักษณ์นี้สามารถนำมาใช้สร้างระบบรู้จำอักษรได้โดยต้องทำการเลือกลักษณะบ่งความต่างให้เหมาะสมขึ้นกับระบบที่ใช้ ในบทความนี้จะใช้เฉพาะลักษณะบ่งความต่างของอักษรไทยเพื่อทำการรู้จำอักษรไทย

สำหรับตัวอย่างงานวิจัยทางด้านทฤษฎีปฏิลักษณ์ซึ่งเสนอการใช้ลักษณะบ่งความต่าง Luksaneeyanawin, Sudapom (1993) เสนอหลักการใช้ลักษณะบ่งความต่างของภาษาไทย (Thai Distinctive Feature) ในการแบ่งแยกอักษรไทยออกจากกันโดยมีลักษณะเป็นลำดับชั้นโดยลักษณะบ่งความต่างของภาษาไทย ลักษณะบ่งความต่างของพยัญชนะมี 12 ลักษณะ แยกสระด้วยลักษณะบ่งความต่าง 10 ลักษณะ และแยกวรรณยุกต์ด้วยลักษณะบ่งความต่าง 3 ลักษณะดังแสดงในรูปที่ 2.1 และ 2.2 สำหรับลักษณะบ่งความต่างที่ใช้ในการแยกพยัญชนะ คือ Box , Top Connect , Loop , Medial Loop , Right End , Top End , Body Loop , Serrated , Left Body Loop , Clockwise Head , Neck , Tail ตัวอย่างเช่น ส มี Box , Top Connect , Loop , Right End และไม่มี Medial Loop , Body Loop ส่วนลักษณะอันอื่นไม่จำเป็นต้องตรวจสอบ หรือตัวอย่างเช่น จ มี Top Connect , Loop , Medial Loop และไม่มี Box , Right End ส่วนลักษณะอันอื่นไม่จำเป็นต้องตรวจสอบ พบว่าลักษณะบ่งความต่างบางอย่างหายากมาก และลำดับความสำคัญของลักษณะบ่งความต่างแต่ละอย่างไม่เท่ากัน เช่น การหาว่ามี medial Loop หรือไม่ หรืออักษรบางตัวเมื่อถูกสัญญาณรบกวนอาจได้ลักษณะบางตัวซึ่งไม่ตรงกับกฎที่ตั้งไว้ เช่น จ อาจตรวจพบ Right End ทำให้รู้จำผิด ดังนั้นลำดับของการหาลักษณะบ่งความต่างจึงสำคัญมากสำหรับวิธีแยกแยะแบบนี้

Distinctive Features of Thai Tone Characters

	่	้	๊	๋
Loop	-	+	+	-
2 Strokes	-	(-)	(-)	+
Serrated Body	(-)	-	+	(-)

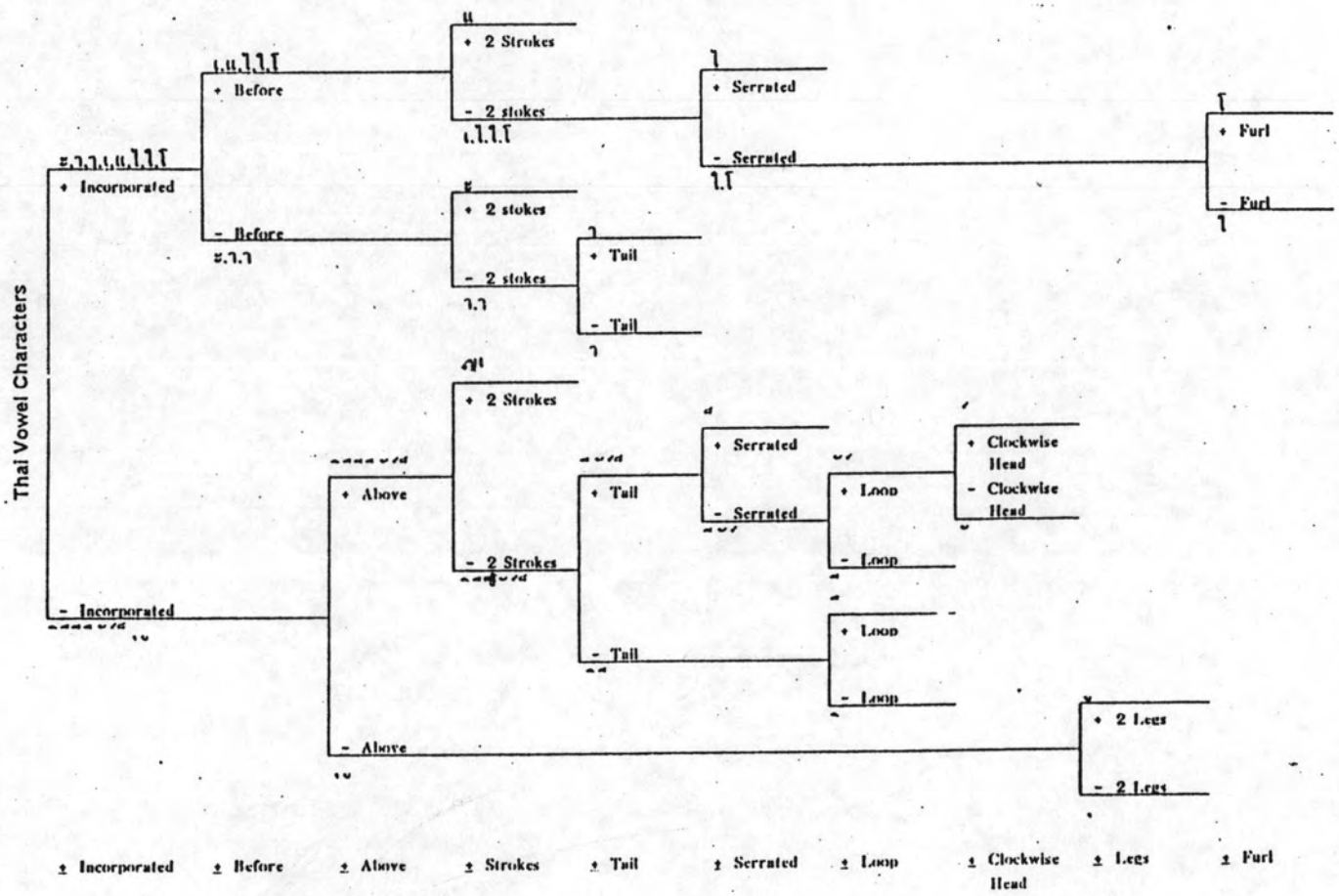
Distinctive Features of Thai Vowel Characters

	ะ	า	ำ	เ	แ	ไ	ใ	โ	็	์	๎	๏	ๆ	ฯ	ิ	ี
Incorporated	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
Before .	-	-	-	+	+	+	+	+	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Above	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	+	+	+	+	+	+	+	-
2 Strokes	+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	(-)
Tail	(+)	-	+	-	(-)	(+)	(+)	(+)	-	-	+	(+)	+	+	+	(-)
Serrated	(-)	(-)	(-)	(-)	(-)	+	-	-	(-)	(-)	-	(-)	-	-	+	(-)
Loop	(+)	(-)	(-)	(+)	(+)	(+)	(+)	(+)	-	+	-	(-)	+	+	(+)	(+)
Clockwise Head	(-)	(-)	(-)	(+)	(+)	(+)	(+)	(+)	(-)	(-)	(-)	(-)	-	+	(+)	(+)
2 Legs	(-)	(-)	(-)	(-)	(+)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	-
Furl	(-)	(-)	(-)	(-)	(-)	(-)	-	+	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)

Distinctive Features of Thai Consanant Characters

	ม	ฆ	น	ป	บ	ษ	พ	ฟ	ห	ฝ	ผ	ย	ท	ฑ	ธ	ก	ฉ	ฤ	ภ	ฦ	ฏ	ฎ	ญ	ณ	ฌ	ล	ส	ค	ค	ศ	ด	ต	ฒ	อ	ช	ฉ	ช	ข	ซ	ซ	ง	จ	ฐ	ว	ร	ธ
Box	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-
Top Connect	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+
Loop	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	(+)	(+)	(+)	(+)	(+)	+	+	+	+	-
Medial Loop	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	(-)	(-)	(-)	(-)	(-)	+	+	-	-	(-
Right End	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	+	-	+	+	+	+	-	-	+	-	+	(+
Top End	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	(-)	-	-	-	-	-	-	+	+	+	(-)	+	-	-	+	-	-	+	(-)	+	(-)	(+)	(+)	(+)	(+)	(+)	(-)	(-)	(-)	(+)	(+
Body Loop	+	+	+	-	-	+	-	-	+	-	-	-	-	-	+	(-)	-	-	-	-	+	+	+	+	+	(-)	-	(-)	(-)	-	(-)	(-)	+	-	+	+	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-
Serrated	-	-	-	-	-	-	+	+	+	+	+	+	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	-	+	(-)	(-)	(-)	(-)	(-)	+	-	(-)	-	+	+	(-)	-	(-)	(-)	(-)	(-)	(-)	(-)	(+)	(-)	(-)	(-
Left Body Loop	+	+	-	(-)	(-)	-	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(+)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	+	+	-	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-
Clockwise Head	(+)	(+)	(+)	(+)	(+)	(+)	+	+	(+)	-	-	-	(+)	(+)	(+)	(-)	+	+	-	-	(-)	(-)	(+)	(+)	(+)	(+)	(+)	-	-	(-)	+	+	(+)	(-)	(-)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(-)	(-)	(-
Neck	-	+	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	-	+	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	-	+	-	+	(-)	(-)	(-)	(-)	(-)	(-
Tail	(-)	(-)	-	+	-	+	-	+	(+)	+	-	-	(-)	(-)	(-)	(-)	-	+	-	+	(+)	(+)	+	-	(-)	(-)	(+)	(-)	(-)	(+)	(-)	(-)	(-)	(+)	(-)	-	-	+	+	(-)	(-)	(+)	(-)	(-)	(-

รูปที่ 2.1 แสดงลักษณะเด่นของอักษรไทย



Distinctive Features of Thai Vowel Characters

รูปที่ 2.2 แสดงลำดับการตรวจสอบลักษณะบ่งความต่างของสระ

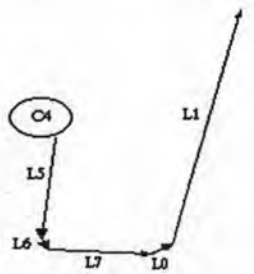
งานวิจัยด้านการรู้จำอักษรเดี่ยว

ในช่วงแรกของ OCR ราวทศวรรต 1950 จะเป็นในลักษณะ template/mask matching ซึ่งสะท้อนถึงเทคโนโลยีสมัยนั้นซึ่งใช้ photodetecter จับแสงที่ผ่านหน้ากาก ซึ่งต่อมาจึงมีการใช้การหา moment , Fourier Series , Correlation/inner product ฯลฯ (S. Mori et. al. 1992) ต่อมามีแนวคิดด้าน Structural Analysis โดย K. S. Fu เป็นผู้นำ ซึ่งเป็นการมองว่าการประกอบกันเป็นรูปร่างหนึ่งของส่วนประกอบย่อยๆ นั้นจะมีกฎอยู่คล้ายการประกอบกันของศัพท์ในภาษาเป็นประโยค กฎนั้นเทียบได้กับไวยากรณ์ทางภาษานั่นเอง ซึ่งการประกอบกันนั้นสามารถแสดงในรูปสายของตัวอักษร(String) หรือต้นไม้

ในช่วงทศวรรตที่แล้วมีการนำโครงข่ายประสาทเทียม(Artificial Neural Network (ANN))มาใช้ในงาน OCR อย่างมาก ซึ่ง ANN ทำงานคล้าย Pattern mapping มีโครงสร้างเลียนแบบสมองมนุษย์ มีความสามารถในการจดจำว่ารูปแบบข้อมูลเข้าแบบใดจะให้ข้อมูลออกอย่างไรและสามารถจดจำแม้ว่าข้อมูลเข้านั้นจะไม่เหมือนกับข้อมูลที่ให้เรียนรู้ มีผู้เสนอรูปแบบมากมายเช่น Hopfield , Back Propagation Learning , Neocognitron ฯลฯ ซึ่งสามารถทำกระบวนการของ OCR ได้หลายกระบวนการตั้งแต่ noise reduction , feature extraction , Classification เช่นในงานของ กิตติพงษ์ เจนวิถีสุข(2539) ใช้การดึง feature โดยใช้ฟรีแมนเวคเตอร์แทนจุดภาพตัวอักษร แล้วนำ feature นั้นมาผ่านโครงข่ายประสาทเทียมแบบ Back Propagation Learning เพื่อรู้จำ

สำหรับอักษรภาษาไทยนั้น การวิเคราะห์ตัวอักษรเดี่ยวมีการศึกษาวิจัยต่อเนื่องดังต่อไปนี้

งานวิจัยของ สนธยา เมรินทร์ (2537) ทำการรู้จำอักษรเดี่ยวภาษาไทยใช้ SyntPR โดยใช้ Grammar 1 ตัวต่ออักษร 1 ตัว ดังนั้นจะมี Grammar เท่ากับจำนวนอักษร สำหรับ primitive ที่นำมาสร้างเป็น Grammar นั้นหาได้โดยทำภาพอักษรให้บางแล้วแทนการต่อเชื่อมจุดด้วยรหัสทิศทาง 8 ทิศ (เรียก freeman vector) และวงกลม เช่น ป สามารถแทนได้ดังรูปที่ 2.3



รูปที่ 2.3 แสดงตัวอย่างการแทนจุดภาพอักษรด้วยรหัสทิศทาง

กฎการเรียง primitive เป็นรูป "ป" ถือเป็น 1 Grammar ดังนั้นสมาชิกในกลุ่มจะมีแค่ 1 เท่านั้น ส่วนการ parsing นั้นใช้ Tree-to-Tree Distance มีหลักการดังนี้ คือ เมื่อเรามีอักษรเข้ามาตัวหนึ่ง ทำการแทนอักษรตัวนั้นด้วย primitive เราจะได้ลักษณะการต่อเชื่อมแบบหนึ่ง นำไปเทียบกับ Grammar ของทุกตัวทำการลบหรือแทรก primitive จนอักษรตัวนั้นเหมือน Grammar ถ้าลบหรือแทรกมากครั้งแสดงว่าไกลจาก Grammar นั้นมาก ให้ตอบผลเป็นตัวที่ลบหรือแทรกน้อยที่สุด

ข้อจำกัดที่เห็นคือมี Grammar เท่ากับจำนวนอักษรซึ่งช้า รวมทั้งการหา primitive ทำได้ยากเมื่อตัวอักษรเล็กลงจุดภาพจะติดกันเช่นหัว อ. ติดกับ หาง อ. หรือหาหัวของอักษรไม่พบเนื่องจากหัวอักษรเล็กจึงทึบหาช่องภายในไม่เจอ ทำให้ Grammar ที่ได้ไม่ตรงกับต้นแบบซึ่งทำไว้ตอนอักษรตัวใหญ่ จำเป็นต้องมี Grammar สำหรับอักษรตัวเล็กที่ อ. มีหัวกับหางติดกันเพิ่มขึ้นมาอีก ข้อดีคือสามารถเพิ่มตัวอักษรใหม่ๆ ได้ตลอดโดยเพิ่มจำนวน Grammar ขึ้น สำหรับผลการทดลองได้ความถูกต้อง 98.5 % และงานที่ต่อเนื่องกันของ เดชา รัตนาธาร (2538) ใช้วิธีของ สนธยา เมรินทร์ (2537) โดยเพิ่มการตัดสินใจด้วย Fuzzy Logic และปรับปรุงในส่วนทำให้บาง ได้ความถูกต้องถึง 99.64% แต่จำกัดตัวอักษรเล็กที่สุดที่ขนาด 18 point

เห็นว่างานของทั้ง 2 ให้ความถูกต้องสูง เนื่องจากได้นำลักษณะของตัวอักษรมาใช้ นั่นคือใช้ความต่อเนื่องลักษณะการลากเส้นมาใช้ประยุกต์เข้ากับ SyntPR แต่ยังมีข้อจำกัดอยู่บ้างในการหา primitive สำหรับอักษรที่ตัวเล็ก และการที่เลือกใช้ Grammar จำนวนมากทำให้การ parsing ทำได้ยาก

โปรแกรมของ สนธยา เมรินทร์ และ เดชา รัตนาธาร นั้นเมื่อใช้งานในขนาดตัวอักษร 12 ถึง 16 (เป็นขนาดของสิ่งพิมพ์ทั่วไป) พบปัญหาที่ส่วนการดึง feature ของตัวอักษร คือ การหาหัว การติดกันภายในตัวอักษร เช่น อักษร อ อ่าง ส่วนหางกับหัวอักษรติดกัน และส่วนใหญ่ช่องว่างในหัวตัวอักษรจะไม่มีทำให้หาหัวตัวอักษรด้วยอัลกอริทึมแบบเดิมไม่ได้ ทำให้โครงสร้างต้นไม้ที่ได้ไม่ตรงกับต้นแบบที่เก็บไว้ทำให้ความถูกต้องลดไปมาก การเพิ่มต้นแบบเป็นแนวทางหนึ่งที่จะเพิ่มความถูกต้องได้แต่ต้องแลกกับเวลาที่สูญเสียไป

ในขั้นตอนการหาระยะห่างกับต้นแบบ พบว่าผู้วิจัยส่วนใหญ่ให้ความสำคัญกับหัวของตัวอักษรมาก เช่น ในงานของ สุรพันธ์ เอื้อไพบูลย์ (2531) รู้จำลายมือเขียนโดยพิจารณาหาส่วนหัวตัวอักษรแล้วใช้เป็นเกณฑ์ในการแบ่งกลุ่มตัวอักษร หรือ สุเจตน์ จันทรังษ์ , ศุภชัย นำเกียรติสกุล และ สุริยา วิทยาประดิษฐ์ (2536) แบ่งกลุ่มด้วยลักษณะของหัวอักษร ระดับหัว ระดับจุดเริ่มต้น ฯลฯ

ข้อด้อยที่เห็นคือการพิจารณาลักษณะบ่งความต่างที่สำคัญเช่น "ส่วนหัวตัวอักษร" ของภาพอักษรทั้งภาพทำให้มีการคำนวณเกินความจำเป็น หรืออย่างในกรณีของ สนธยา เมรินทร์ (2537)นั้น การเทียบหาระยะห่างต้องทำกับ template ของตัวอักษรทั้งหมด แนวคิดที่เสนอจึงเป็นการแบ่งกลุ่มของตัวอักษรโดยใช้ลักษณะโดยรวมดังนั้นเราสามารถใช้ template ตัวแทนของกลุ่มก็เพียงพอ จากนั้นค่อยพิจารณาลักษณะเฉพาะ(เช่นกลุ่มของ ค ฅ ด ต ศ เห็นว่าลักษณะคล้ายกันโดยมีลักษณะพิเศษ คือ ค และ ฅ มีหัวออกด้านนอกและอยู่ตรงกลางการหาหัวของตัวอักษรก็เพียงพิจารณาส่วนกลางของภาพก็พอ) ดังนั้นเราสามารถมุ่งความสนใจไปที่การแยกลักษณะเฉพาะกลุ่ม ไม่ต้องเสียเวลาเกินความจำเป็น (ตัวอย่างเช่นการหาหัวบริเวณตรงกลางของ "บ")

C. Kimpan et al (1993) ทำงานรู้จำอักษรไทยตัวเขียนโดยใช้ตำแหน่งของหัวและจุดปลายของภาพอักษรแบ่งกลุ่มแล้วแยกแยะในกลุ่มด้วยเทคนิคพิเศษโดยทดสอบกับอักษร 42 ตัวเขียนโดยคน 50 คนๆ ละ 1 ชุดได้ผล

การรู้จำถูกต้อง 98.2% จะเห็นว่าข้อจำกัดอยู่ที่หัวของอักษรเช่นกัน และการหาปลายยังขึ้นกับกระบวนการทำอักษรให้บางอีกด้วย

การแบ่งกลุ่มตามโครงสร้างหลักนี้สามารถใช้ทฤษฎีของ SyntPR ได้ ซึ่งจำเป็นต้องใช้ Grammar ไม่มาก และไม่ซับซ้อนมาก จะมีแค่เส้นแนวดิ่งและเส้นแนวนอนเท่านั้น แล้วยังได้ใช้ความสัมพันธ์ระหว่างเส้นต่างๆ ที่ประกอบเป็นตัวอักษรอีกด้วย

ในวิทยานิพนธ์นี้จะเสนอการแบ่งกลุ่มอักษรออกจากกันด้วยโครงสร้างหลักก่อนโดยใช้ SyntPR ร่วมกับระดับของอักษรและแยกอักษรในกลุ่มออกจากกันด้วยลักษณะบ่งความต่างของภาษาไทย

<u>งานวิจัยด้านการแยกอักษรที่ติดกัน (ตัวอักษรคนละตัวแต่มีจุดภาพติดกัน)</u>

ปัญหาที่เกิดขึ้นที่ส่วนประกอบของระบบส่วนแรกคือการสแกนเอกสาร มาจากข้อจำกัดด้านความละเอียดของสแกนเนอร์ที่ใช้ (อาจเนื่องจากไม่ต้องการเสียเวลา ประหยัดความจุของสื่อที่เก็บหรือราคาของสแกนเนอร์ที่ประหยัด) หรือปัญหาของตัวเอกสารเลอะ ขนาดตัวอักษรเล็ก หรือความด้อยคุณภาพของกระดาษ ทำให้ภาพอักษรที่ได้จากการสแกนติดกันทำให้ต้องหาวิธีแยกซึ่งเรียกว่า Segmentation (H. Fujisawa et al. 1992)

งานในด้านการแยกอักษรที่ติดกัน(ตัวอักษรคนละตัวแต่มีจุดภาพติดกัน) Tsujimoto et. al (1992) กล่าวว่าเป็นเรื่องยากที่จะออกแบบกระบวนการแยกตัวอักษรที่มีประสิทธิภาพเสร็จในตัว ปัญหาหลักคือการจำแนกว่าเป็นอักษรที่ติดกันหรือไม่ และจะแยกตรงไหน วิธีง่ายๆ ที่พอจะใช้ได้คือไล่อักษรตั้งแต่ต้นบรรทัด หาจุดที่น่าจะแบ่ง ตัดแล้วส่งไปแยกแยะว่าเป็นตัวอะไร แต่กลับง่ายกว่าถ้าจะรวมขั้นตอนตัดแยกแยกตัวอักษรที่ติด แยกแยะและการตรวจสอบความถูกต้องทางการสะกดเข้าไว้ด้วยกัน โดยทำงานในลักษณะลองตัด แยกแยะแล้วตรวจสอบจนได้ที่ดีที่สุด

งานวิจัยของไทยเช่นในงานของ สมศักดิ์(2539) ทำการรู้จำอักษรภาษาไทยเป็นบรรทัดโดยเน้นที่การตัดแยกอักษรที่ติดกัน มีการกำหนดขนาดตัวอักษรตายตัว วิธีการทำงานของโปรแกรมคือการดึงภาพอักษรออกจากบรรทัดโดยดูช่องว่างในแนวนอน ดังนั้นจึงตรวจสอบว่าเป็นอักษรที่ติดหรือไม่โดยใช้จำนวนความกว้างจุดภาพสำหรับการตัดนั้นจะนำอักษรที่สงสัยว่าติดมาเทียบกับฐานข้อมูลที่สร้างเอาไว้ เมื่อตรงกับแบบที่ติดแบบใดแล้วจึงตัดตามแบบนั้น ฐานข้อมูลที่ว่านี้คือการทำ horizontal projection และ vertical projection สำหรับตัวติดทุกประเภท

ข้อที่จำกัดคือลักษณะอักษรที่ทำเป็นขนาดตายตัวซึ่งถ้าจะใช้ในทางปฏิบัติต้องใช้อักษรหลายขนาด จะต้องทำฐานข้อมูลสำหรับอักษรขนาดอื่นด้วย อาจแก้ได้ด้วยการปรับขนาดอักษร และยังพบว่าอักษรที่สแกนนั้นมีอัตราการติดต่ำมาก ฐานข้อมูลที่ทำจริงๆ จึงมีขนาดไม่ใหญ่มาก โดยถ้ามีอักษรติดกันมากกว่านี้ ขนาดฐานข้อมูลจะใหญ่ตามไปด้วย อีกทั้งการเปรียบเทียบ อักษรที่เข้ามากับฐานข้อมูลนั้นใช้การเปรียบเทียบแบบจุดต่อจุด ทำให้

น่าสงสัยว่าถ้ามีแบบอักษรที่ติดมากๆ จะใช้ได้ผลหรือไม่ เช่น "ฉ" มีลักษณะ horizontal projection และ vertical projection เหมือนกับ "เม" จะทราบได้อย่างไรว่าต้องตัดหรือไม่

Su Liang. et al (1993) ทำงานแยกอักษรภาษาอังกฤษที่ติดกันโดยใช้ discrimination function ในการแยกตัวอักษรโดยใช้ข้อมูลจากการรวมจุดในแนวตั้งและจุดรอบตัวอักษร (profile of image) ซึ่งจะให้ข้อมูลของจุดร่วมจุดแยกของอักษร ร่วมกับการใช้การตรวจตัวสะกด โดยทดสอบกับเอกสารจริง 12 ฉบับ("NEWS LINE" an U . of Windsor publication) 8 ฉบับมีอักษรติดประมาณ 40 % ได้ความถูกต้อง 99.85 % อีก 4 ฉบับมีอักษรติดประมาณ 60 % ได้ความถูกต้อง 99.4% ข้อดีของวิธีนี้คือใช้ได้ดีกับอักษรภาษาอังกฤษซึ่งเป็นอักษรระดับเดียวส่วนใหญ่ แต่เป็นข้อจำกัดเช่นกันคือไม่สามารถสอนระบบเพิ่มเติมเมื่อมีอักษรติดแบบใหม่ๆ เข้ามาได้

Thomas A. Bayer et al.(1993) ใช้ Time Delay Neural Network ในการจดจำรูปแบบการติดกันของตัวอักษร โดยระบบจะเรียนรู้จากตัวติดกันตั้งแต่ 2-10 ตัวอักษร ใช้การเรียนแบบ Supervised มีข้อมูลเข้าเป็น local feature ของจุดในแต่ละแนวคอลัมน์ ให้ข้อมูลออกเป็น 1 ณ คอลัมน์ที่จะตัดและ 0 ที่อื่นๆ ดังนั้นเมื่อมีข้อมูลจริงเข้ามาค่าข้อมูลออกจะเป็น probability ของแต่ละคอลัมน์ local feature ดังกล่าวคือจำนวนจุดดำในแนวดิ่ง จำนวนจุดขาว(ดำ)นับจากบนก่อนถึงจุดดำ(ขาว) จำนวนการเปลี่ยนจุดดำเป็นขาว จำนวนจุดดำ(ขาว)ข้างเคียง ความสูงเฉลี่ยของศัพท์ และ aspect ratio ข้อดีของวิธีนี้คือได้ระบบที่ตัดแยกอักษรที่เกิดขึ้นจริงได้ ซึ่งสามารถนำมาเรียนรู้กับอักษรติดของไทยได้ ข้อจำกัดคือต้องหาข้อมูลที่จะนำมาสอน Neural Network ให้เหมาะสม

เนื่องจากอักษรภาษาอังกฤษเป็นอักษรระดับเดียว แต่อักษรไทยมี 3 ระดับ (คือ บน กลางและล่าง) ดังนั้นตัวอย่างจากภาษาอังกฤษจึงใช้ได้กับเฉพาะการตัดแยกอักษรภาษาไทยที่ติดกันในระดับกลางเท่านั้น สำหรับที่ติดกันเป็นต่างระดับนั้นต้องพัฒนาขึ้นมาเอง ซึ่งจะอธิบายละเอียดต่อไปในบทที่ 3 โดยเราสามารถแบ่งประเภทที่ติดและตรวจสอบว่าเป็นแบบใดได้ง่ายโดยใช้ลักษณะบ่งความต่างของภาษาไทย และสามารถหาจุดที่จะตัดแยกได้ง่ายด้วย

<u>งานวิจัยด้านกระบวนการวิเคราะห์เอกสาร</u>

งานวิจัยในต่างประเทศจำนวนมากที่ได้นำเสนอเกี่ยวกับระบบทั้งหมดของงานรู้จำเอกสารเช่นในงานของ S. Tsujimoto et. al. (1992) ได้นำเสนอรูปแบบ ระบบอ่านเอกสารภาษาอังกฤษซึ่งประกอบด้วย 3 ส่วนหลักคือ

- ***วิเคราะห์เอกสาร*** หาส่วนประกอบที่เป็น บรรทัดของอักษร ตาราง รูปภาพ โดยใช้การตีกรอบแต่ละตัวอักษรก่อนแล้วรวมเป็นบรรทัดและย่อหน้าซึ่งสามารถแยกตารางและรูปภาพออกไปได้ด้วยขนาดกรอบ
- ***เข้าใจเอกสาร*** คือการดูว่าส่วนประกอบเหล่านั้นมีความสัมพันธ์กันอย่างไร อันไหนมาก่อนหลัง
- ***แยกตัวติดและรู้จำตัวอักษร*** ซึ่งใช้การนับจุดในแนวดิ่งร่วมกับตัวดำเนินการ AND หาจุดที่น่าจะตัดร่วมกับการรู้จำและคำศัพท์ ผลที่ได้ให้ความถูกต้อง 99.5% โดยมีตัวติด 28% (ความถูกต้องของการรู้จำตัวอักษรเดี่ยว 99.6 % ใช้สแกนเนอร์ละเอียด 300 dpi เอกสารมาจากนิตยสาร หนังสือพิมพ์ คู่มือ จดหมาย ฯลฯ)

ส่วนงานในด้านการวิเคราะห์เอกสารสำหรับวิทยานิพนธ์นี้จะเริ่มจากการอ่านแฟ้มภาพเอกสารที่ได้จากการสแกน โดยจะมีการกำจัดสัญญาณรบกวน การแก้ความเอียงซึ่งเกิดจากการวางหน้ากระดาษ การหาเส้นแบ่งคอลัมน์และการแบ่งเส้นบรรทัด รวมถึงการสร้างแฟ้มเก็บข้อมูลที่รู้จำได้

การแก้ความเอียงนั้น ในขอบเขตการวิจัยได้กำหนดองศาไว้เนื่องจากการทดลองกับผู้ใช้ที่ไม่ได้รับการฝึกการใช้สแกนเนอร์ 10 คนจำนวน 50 หน้าพบว่ามีความเอียงตั้งแต่ +5 ถึง -5 องศา จึงกำหนดช่วงปลอดภัยเป็น +7.5 ถึง -7.5 องศา

<u>ตัวอักษรภาษาไทย</u>

ตัวอักษรภาษาไทยจะประกอบด้วย พยัญชนะ สระ และวรรณยุกต์ ซึ่งสามารถแยกออกได้เป็นดังนี้

- ตัวพยัญชนะ มีทั้งหมด 44 ตัวอักษร ได้แก่
  ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ
  ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
- ตัวสระ ิ ี ึ ื ุ ู ็ ั ์ เ โ ใ ไ ำ า ฤ ฦ
- ตัววรรณยุกต์ มีทั้งหมด 4 ตัว ่ ้ ๊ ๋
- เครื่องหมายต่างๆ ได้แก่ ๆ ฯ
- ตัวเลขไทย มีทั้งหมด 10 ตัว ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐

เพื่อความสะดวกในการจัดการจึงแบ่งประเภทของตัวอักษรที่มีระดับต่างๆ เป็น 6 ลักษณะตามระดับของอักษรที่ปรากฏในบรรทัดดังตารางที่ 2.2

ตารางที่ 2.2 แสดงการแบ่งลักษณะตัวอักษร

ลักษณะ	เกณฑ์การแบ่ง	สมาชิกในแต่ละลักษณะ
1	อักษรที่อยู่ระดับบน	ิ ี ึ ื ็ ั ์ ่ ้ ๊ ๋ ํ
2	อักษรที่อยู่ระดับกลางและบน	ป ฝ ฟ ฬ ไ ใ โ
3	อักษรที่อยู่ระดับกลาง	ฯ ะ เ า 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 ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐
4	อักษรที่อยู่ระดับกลางและล่าง	ฦ ฤ ๆ ฯ ญ ฎ ฏ ฐ
5	อักษรที่อยู่ระดับล่าง	ุ ู
6	วรรณยุกต์ที่อยู่พร้อมกับสระบน	อี่ , อิ้ , อิ๊ , อิ๋ , อิ์

* <u>หมายเหตุ</u> บางอักษร (เช่น “ฯ”) เมื่อต่างแบบอักษรกันจะมีลักษณะต่างกันคือแบบอักษรหนึ่งเป็นลักษณะที่ 2 แต่อีกแบบอักษรเป็นลักษณะที่ 3

i 17413795

ทฤษฎีของการรู้จำแบบ SyntPR

K.S.Fu (1974) ได้เสนอไว้ดังนี้

$G = (V_N , V_T , P , S)$

G : Grammar , V_N : Non-Terminal, V_T : Terminal ,

P : Production rule , S : Start

เราสร้าง Grammar ด้วย Production rule ซึ่งกำหนดความสัมพันธ์ของ S V_N และ V_T

*ตัวอย่าง*เช่น ต้องการสร้าง Grammar ที่สร้างสายอักษร abbbb....bbbc

V_T มีสมาชิกคือ a , b , c

เราให้ V_N มีสมาชิกคือ X , Y

เราสร้าง Production Rule ดังนี้

$S \rightarrow aX$

$X \rightarrow Yc$

$Y \rightarrow bY$

$Y \rightarrow b$

สามารถแบ่งแบบของ SyntPR หลักๆ ได้ 4 แบบคือ

Type 0 (unrestricted) เป็นแบบที่ไม่มีข้อจำกัดในการสร้าง Grammar

Type 1 (Context - Sensitive) Grammar มีกฎในการสร้างดังนี้

$\xi_1 A \xi_2 \rightarrow \xi_1 B \xi_2$

$A \in V_N$, $B \in V_T$ โดยหลักคือ "A can be replaced by B in the ξ_1 ξ_2 context "

นั่นคือสายอักษรจะยาวออกไปได้เรื่อยไม่ว่าจากทางไหนและจุดไหนของสายอักษรโดยมี ξ_1 ξ_2 เป็นเงื่อนไขในการโต

Type 2 (Context - Free) Grammar มีกฎในการสร้างดังนี้

$A \rightarrow B$

$A \in V_N$, $B \in V_T$ โดย A สามารถแทนได้ด้วย B ได้ตลอด

นั่นคือสายอักษรจะยาวออกไปได้เรื่อยไม่ว่าจากทางไหนและจุดไหนของสายอักษร

Type 3 (Finite - State or Regular) Grammar มีกฎการสร้างดังนี้

$A \rightarrow aB$ หรือ $A \rightarrow b$

$A,B \in V_N$ และ $a , b \in V_T$

แบบนี้สายอักษรถูกกำหนดให้ยาวออกไปได้ทางเดียว

การแจงส่วน (parsing)ของ Type 3 เหมาะสมกับงานของเราเพราะการแทนโครงสร้างหลักด้วย Grammar นั้น ตัวโครงสร้างหลักไม่มีความซับซ้อนมาก จึงใช้ Type 3 ในการสร้างได้