บทที่ 1

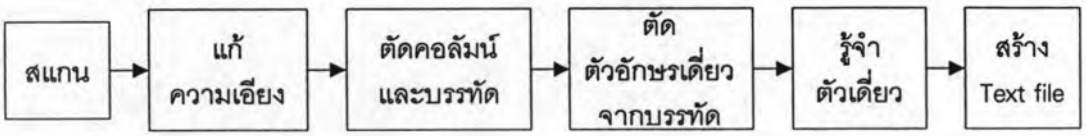
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ทรัพยากรบุคคลของประเทศชาติจะได้รับการพัฒนาก็ต่อเมื่อได้รับความรู้จากสื่อต่างๆ ทั้งจากในและนอกห้องเรียน การรับสารที่สำคัญที่สุดอย่างหนึ่งคือการอ่าน ปัญหาเกิดขึ้นมาว่าจะทำอย่างไรให้คนกลุ่มที่ไม่สามารถอ่านออกหรือมองไม่เห็นได้รับข่าวสารได้โดยง่ายโดยที่ไม่ต้องให้ใครมาอ่านให้ฟังหรือต้องทำอักษรเบรลล์ (Braille)ให้ การทำให้คอมพิวเตอร์ทำหน้าที่นี้แทนเป็นทางออกทางหนึ่ง การพัฒนาโปรแกรมที่มีระบบรู้จำเอกสารภาษาไทยทำงานร่วมกับระบบสังเคราะห์เสียงจึงเป็นสิ่งจำเป็น

ข้อดีของการเก็บเอกสารในรูปสื่ออิเล็กทรอนิกส์คือความสะดวกในการจัดเก็บ ความคล่องตัวในการค้นหา แก้ไข และปรับปรุง ตลอดจนอายุที่ยาวนานของสื่อที่ใช้เก็บ อย่างไรก็ดีการจัดเก็บเอกสารจากการสแกนและเก็บเป็นรูปภาพแบบบิทแมปเป็นการสิ้นเปลืองหน่วยความจำและส่งผลเสียต่อกระบวนการจัดการต่างๆที่จะตามมา ตัวอย่างเช่นในการเก็บเอกสาร 1 หน้าขนาด A4 (8x11นิ้ว) โดยมี 30 บรรทัด 80 คอลัมน์ โดยสแกนด้วยความละเอียด 200 จุดต่อนิ้ว โดยไม่มีการบีบอัดต้องใช้ความจุถึง 3 Mbytes (8x200x11x200 = 3M) แต่ถ้าเปลี่ยนเอกสารให้เป็นรหัส ASCII โดยใช้ระบบรู้จำตัวอักษรแล้วจะใช้ความจุเพียง 2 Kbytes (30 x 80 = 2400) และยังสามารถแก้ด้วยโปรแกรมคอมพิวเตอร์ประมวลผลคำ(Word Processor) ได้อย่างสะดวกอีกด้วย

กระบวนการหลักโดยทั่วไปของระบบรู้จำเอกสารแสดงได้ดังรูปที่ 1.1



รูปที่ 1.1 กระบวนการหลักของระบบรู้จำเอกสารภาษาไทย

ระบบรู้จำอักษร (Character Recognition System) ประกอบด้วยขั้นตอนหลัก 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) ทำการรู้จำภาพของอักษรเดี่ยวนำผลที่ได้ส่งกลับให้ส่วนวิเคราะห์เอกสารเพื่อสร้างแฟ้มเอกสาร

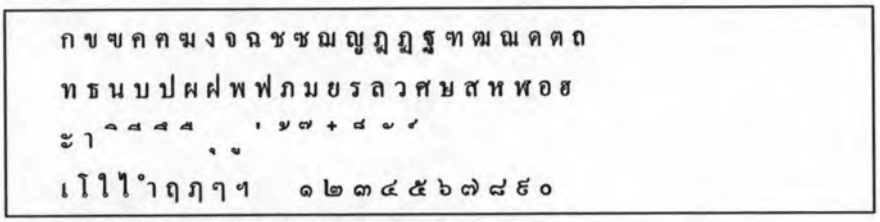
สำหรับภาษาไทยเดิมงานวิจัยที่ผ่านมาส่วนใหญ่มุ่งเป้าหมายไปที่การวิเคราะห์ตัวอักษรเดี่ยว (Single Character) แต่เมื่อจะนำไปใช้งานจริงพบว่าระบบที่สามารถรู้จำอักษรเดี่ยวได้สมบูรณ์นั้นไม่เพียงพอที่จะใช้ในการแปลงเอกสารที่มีความซับซ้อน(เช่น การเรียงอักษรแบบหลายคอลัมน์ ตาราง รูปภาพ อักษรขนาดต่างๆ กัน) หรือไม่สมบูรณ์ (เนื่องจากเหตุต่างๆ เช่น หมึกเลอะ ตัวอักษรติดกันเนื่องจากความละเอียดของสแกนเนอร์ไม่พอ หรือมีขีดเส้นใต้ ตัวอักษรขาด) ดังนั้นเป้าหมายในวิทยานิพนธ์นี้จึงครอบคลุมตั้งแต่สแกนเอกสารเป็นหน้าไปจนถึงได้ข้อมูลออกมาเป็นแฟ้มข้อมูลแบบ Text

<u>วัตถุประสงค์</u>

1. เพื่อพัฒนาระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของตัวอักษรไทย
2. ศึกษาและพัฒนาโปรแกรมคอมพิวเตอร์ส่วนวิธีวิเคราะห์หาส่วนประกอบของเอกสาร
3. ศึกษาลักษณะบ่งความต่างเฉพาะของตัวอักษรไทยในกรณีที่เป็นตัวอักษรเดี่ยวและตัวอักษรที่ติดกันที่ใช้ในกระบวนการรู้จำ
4. พัฒนาโปรแกรมคอมพิวเตอร์โดยใช้ลักษณะบ่งความต่างของตัวอักษรไทยในการแยกตัวอักษรไทยที่ติดกันและการรู้จำอักษร

<u>เป้าหมายและขอบเขตของวิทยานิพนธ์</u>

1. ได้โปรแกรมที่สามารถรู้จำเอกสารที่มีลักษณะดังนี้
   - รับข้อมูลเป็นแฟ้มภาพเอกสารที่สแกนด้วยความละเอียด 300 dpi (เก็บในรูปแบบ *.PCX) และส่งข้อมูลออกเป็น Text file
   - ทำงานบนระบบปฏิบัติการ Ms. Windows 3.1x พัฒนาโดย Ms. Visual C++ 1.52
   - ลักษณะบรรทัดเป็นเส้นขนานและเป็นแนวระดับโดยสามารถเอียงได้ +7.5 ถึง -7.5 องศา
   - เส้นฐานตัวอักษรเรียงกันเป็นเส้นตรง
   - ในบรรทัดมีขนาดตัวอักษรขนาดเดียว
   - ไม่มีภาพประกอบหรือตาราง ไม่มีขีดเส้นใต้ ไม่มีตัวแบบ Bold และ Italic
   - แบบของตัวอักษร 2 แบบคือ CordiaUPC และ AngsanaUPC ขนาดตัวอักษรอยู่ระหว่าง 14-24 point
   - ตัวอักษรที่มีได้ดังนี้



2. อัตราความถูกต้องในการรู้จำไม่น้อยกว่าร้อยละ 90
3. ทำงานโดยอัตโนมัติทุกขั้นตอนและผู้ใช้ไม่ต้องได้รับการฝึกการใช้งาน

<u>ขั้นตอนและวิธีการดำเนินงาน</u>

1. ศึกษาผลงานในอดีตเกี่ยวกับระบบรู้จำเอกสารภาษาต่างๆ ใน 3 ด้านคือ การวิเคราะห์เอกสาร วิธีการแยกตัวอักษรที่ติดกัน และการรู้จำตัวอักษรเดี่ยว
2. สรุปปัญหาที่ต้องปรับปรุง พร้อมทั้งกำหนดขอบเขตการทำงานในแต่ละด้าน
3. เขียนโปรแกรมส่วนวิเคราะห์เอกสาร
4. เขียนโปรแกรมรู้จำตัวอักษรเดี่ยว แยกตัวอักษรที่ติดกัน พร้อมสร้างฐานข้อมูลและทดสอบความแม่นยำ แก้ไขไปพร้อมๆ กัน
5. นำโปรแกรมทั้ง 3 ส่วนมารวมกัน และทดสอบปรับปรุงระบบโดยรวมให้ได้ความแม่นยำตามเป้าหมาย
6. ทดสอบกับตัวอักษรแบบอื่น (คือในชุด UPC ของ Microsoft Windows ได้แก่ BrowalliaUPC , DilleniaUPC , EucrosiaUPC , FreeciaUPC , IrisUPC , JasmineUPC , KodchiangUPC , LilyUPC ขนาดตัวอักษร 14-24 point) ปรับปรุงครั้งสุดท้ายเพื่อเพิ่มความคล่องตัวในการใช้งานและเพิ่มความแม่นยำสำหรับตัวอักษรแบบอื่น
7. ประเมินผล สรุป และเขียนวิทยานิพนธ์

<u>ประโยชน์ที่คาดว่าจะได้รับ</u>

1. ได้ระบบรู้จำตัวอักษรภาษาไทยที่สามารถทำงานได้จริงกับอักษรตัวพิมพ์ไทย
2. ได้เรียนรู้ถึงวิธีการและเทคนิคต่างๆที่จะนำมาใช้ในระบบรู้จำตัวอักษร
3. เป็นแนวทางในการพัฒนางานทางด้านนี้ต่อไป