การแก้ไขข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทย ด้วยเจเนติกอัลกอริทึม

ERROR CORRECTION USING GENETIC ALGORITHM FOR THAI OCR

กริช สมกันธา
KRICH SOMGUNTAR

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2546

ISBN 974-324-300-3

การแก้ไขข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทย ด้วยเจเนติกอัลกอริทึม

ERROR CORRECTION USING GENETIC ALGORITHM FOR THAI OCR

กริช สมกันธา
KRICH SOMGUNTAR

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2546

ISBN 974-324-300-3

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การแก้ข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทยด้วยเจเนติกอัลกอริทึม
ERROR CORRECTION USING GENETIC ALGORITHM FOR
THAI OCR

ชื่อนักศึกษา นายกริช สมกันธา

รหัสประจำตัว 43061624

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา วิศวกรรมคอมพิวเตอร์

อาจารย์ผู้ควบคุมวิทยานิพนธ์ รศ.ดร.บุญธีร์ เครือตราชู

อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม อาจารย์กฤตวัน ศิริบูรณ์

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.เอื้อน	ปิ่นเงิน	
ดร.วิศิษฏ์	หิรัญกิตติ	
ดร.อรัญญา	วลัยรัชต์	
รศ.ดร.บุญวัฒน์	อัตชู	
รศ.ดร.บุญธีร์	เครือตราชู	

วัน/เดือน/ปี ที่สอบ 25 กุมภาพันธ์ 2546 เวลา 14.00-16.00 น.

สถานที่สอบ ณ อาคาร 12 ชั้น ชั้น 4 (ห้อง E12-404)

บัณฑิตวิทยาลัยรับรองแล้ว

(รศ.ดร.บุญวัฒน์ อัตชู)

คณบดีบัณฑิตวิทยาลัย

วันที่ 25 **เดือน** มีนาคม **พ.ศ.** 2546

หัวข้อวิทยานิพนธ์	การแก้ไขข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทยด้วยเจเนติกอัลกอริทึม
นักศึกษา	นายกริช สมกันธา
รหัสประจำตัว	43061624
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2546
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ. ดร. บุญธีร์ เครือตราชู
อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม	อ. กฤตวัน ศิริบูรณ์

บทคัดย่อ

เนื้อหาของวิทยานิพนธ์เล่มนี้นำเสนอถึงวิธีการแก้คำผิดใน OCR ภาษาไทยโดยใช้เจเนติกอัลกอริทึม เจเนติกอัลกอริทึมสามารถปรับปรุงผลลัพธ์ของ OCR โดยเริ่มจากการหาคำที่เป็นไปได้ในประโยคและสร้าง Word Graph จากกระบวนการ Token Passing Algorithm โดย Word Graph จะเป็นโครงสร้างที่บอกว่าประโยคสามารถประกอบด้วยคำใดได้บ้างจากนั้นจะหาประโยคที่ถูกต้องโดยใช้ เจเนติกอัลกอริทึม โดยมี Fitness Function เป็นค่าความน่าจะเป็นที่คำนวณจาก Language Model และ ค่าความน่าจะเป็นของคำที่ได้จาก OCR ในกรณีประโยคที่ยาวในการหาประโยคที่ถูกต้องจาก Word Graph จะใช้เวลามาก เนื่องจากประโยคที่ยาวจะสามารถเกิดประโยคที่สามารถเป็นไปได้เป็น ล้านๆ ประโยค ดังนั้นงานวิจัยนี้จึงนำเสนอวิธีการเจเนติกอัลกอริทึมมาช่วยในการปรับปรุงการหาประโยคที่ถูกต้องให้เร็วขึ้น

Thesis Title	Error Correction Using Genetic Algorithm for Thai OCR
Student	Mr. Krich Somguntar
Student ID.	43061624
Degree	Master of Engineering
Programme	Computer Engineering
Year	2003
Thesis Advisor	Assoc. Prof. Dr. Boontee Kruatrachue
Thesis Co-Advisor	Kritawan Siriboon

ABSTRACT

This thesis proposes Thai OCR error correction using genetic algorithm. The correction process start with word graph construction from token passing algorithm, then a graph is search for a corrected sentence with the highest perplexity (using language model, bi-gram and tri-gram) and word probability from OCR. For a long sentence, a search space is huge and consume a lot of time. This thesis propose genetic algorithm method to reduce searching time for possible correct sentence.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จได้ด้วยความอนุเคราะห์จาก รศ.ดร.บุญธีร์ เครือตราชู อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ อ.กฤตวัน ศิริบูรณ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ได้กรุณาให้คำแนะนำ ให้ความช่วยเหลือให้กำลังใจ และช่วยตรวจสอบ แก้ไขเครื่องมือที่ใช้ในการวิจัย ตลอดจนการปรับปรุงข้อบกพร่องต่างๆ จนวิทยานิพนธ์นี้สำเร็จได้อย่างสมบูรณ์ ผู้วิจัยรู้สึกซาบซึ้งในความกรุณา และขอขอบพระคุณเป็นอย่างสูง

ขอขอบคุณอาจารย์ทุกท่าน ที่ได้ประสิทธิ์ประสาทความรู้ ตลอดจนข้อคิดต่างๆ อันก่อให้เกิดประโยชน์ต่อการศึกษาค้นคว้า และเป็นแนวทางในการจัดทำวิทยานิพนธ์จนประสบความสำเร็จ

ขอขอบพระคุณ คุณพ่อ และคุณแม่ ผู้เป็นที่เคารพรักยิ่ง รวมทั้ง พี่สาว ที่ได้ให้ความรัก ให้กำลังใจ ให้การสนับสนุน และช่วยเหลือทุกด้านตลอดมา

ขอขอบคุณมูลนิธิเพื่อการศึกษาคอมพิวเตอร์และการสื่อสาร(C&C) ที่ได้ให้ความกรุณาให้ทุนอุดหนุนในการทำการศึกษาวิจัยในระดับปริญญาโท

ขอขอบคุณรุ่นพี่ เพื่อนๆ และบุคคลที่ผู้วิจัยไม่ได้กล่าวไว้ในที่นี้ ที่ให้การสนับสนุน ตลอดจนให้ความช่วยเหลือในด้านต่างๆ และเป็นกำลังใจให้ผู้วิจัยมาโดยตลอด

ขอขอบคุณห้องปฏิบัติการวิจัย Information Science สำนักวิจัยการสื่อสารและเทคโนโลยีสารสนเทศ ที่ให้การสนับสนุนในการทำวิจัย

คุณค่า และประโยชน์ใดๆ ที่เป็นผลมาจากวิทยานิพนธ์นี้ ผู้วิจัยขอมอบแด่ คุณพ่อ คุณแม่ พี่สาวและ ครู-อาจารย์ ทุกท่าน ด้วยความเคารพยิ่ง

กริช สมกันธา

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

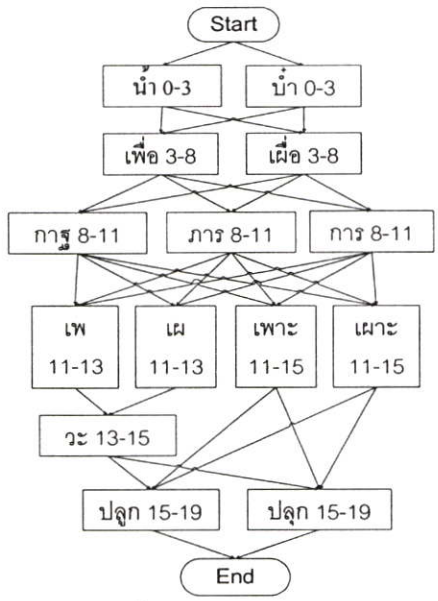
ในปัจจุบันการเพิ่มความถูกต้องของผลลัพธ์ที่ได้จาก OCR มีความสำคัญเนื่องจากการรู้จำตัวอักษรภาษาไทยนั้นมีข้อผิดพลาดอยู่มาก จึงมีหลายงานวิจัยได้มีการพัฒนาศักยภาพของการรู้จำตัวอักษรภาษาไทยแต่ความถูกต้องที่ได้นั้นก็ยังถูกต้องไม่สมบูรณ์ ดังนั้นจึงมีแนวคิดในการแก้คำผิดใน OCR ภาษาไทยโดยให้ผลลัพธ์ของการรู้จำอยู่ในกลุ่มของตัวอักษรที่เป็นไปได้ทั้งหมด จากนั้นสร้างคำจากผลลัพธ์ของการรู้จำในกลุ่มตัวอักษรที่เป็นไปได้ทั้งหมดโดยวิธีโทเกนพาสซิ่ง (Token passing) แล้วทำการคัดเลือกประโยคที่ถูกต้องที่สุดที่เป็นไปได้ เนื่องจากการสร้างคำจากผลลัพธ์ของการรู้จำในกลุ่มตัวอักษรที่เป็นไปได้ทั้งหมดนั้นจะทำให้เกิดคำจำนวนมากที่เป็นไปได้ ส่งผลทำให้เกิดประโยคที่เป็นไปได้เป็นไปได้จำนวนมาก ทำให้ยุ่งยากในการคัดเลือกประโยคที่ถูกต้อง ดังนั้นทางออกในการแก้ปัญหานั้นคือ นำเอาทฤษฎีเจเนติกอัลกอริทึม(Genetic Algorithm)มาใช้ในการ คัดเลือกประโยคที่ดีและเป็นไปได้ที่สุด ซึ่งเป็นกระบวนการจำลองรูปแบบทางชีววิทยา ของวิวัฒนาการทางธรรมชาติถ่ายทอดลักษณะต่างๆ ทางพันธุ์กรรมในการให้กำเนิดประชากรรุ่นใหม่ หรือขยายเผ่าพันธุ์ในรุ่นลูกรุ่นหลานต่อไป โดยทฤษฎีเจเนติกอัลกอริทึมมีความสามารถที่จะเรียนรู้ ใช้เหตุผล พัฒนาและปรับปรุงข้อบกพร่องของตนให้ดีขึ้น ทำให้ได้ผลลัพธ์ที่ดีที่สุดได้

การคัดเลือกประโยคที่ถูกต้องที่สุดนั้นมีหลักการเหมือนกับคนเราสามารถที่จะแก้ไขคำผิดได้เนื่องจากคนเรามีความรู้ด้านภาษาที่สะสมกันมาหลายๆ ปี ดังนั้นเราจะทำให้คอมพิวเตอร์สามารถรู้จักภาษาคนเราได้จึงสร้างแบบจำลองภาษา(Language Model) ซึ่งใช้มุมมองทางสถิติของภาษาและประยุกต์เอาหลักการทางสถิติเพื่อนำมาใช้ในการสร้างแบบจำลองทางภาษา

ในงานวิจัยนี้จึงเป็นการพัฒนาศักยภาพของ OCR ภาษาไทยอีกทั้งยังเป็นการพัฒนากระบวนการเรียนรู้ของคอมพิวเตอร์ในทางปัญญาประดิษฐ์เพื่อให้มีความสามารถในการเรียนรู้แยกแยะได้มากขึ้น เพื่อพัฒนาไปสู่เทคโนโลยีที่คอมพิวเตอร์สามารถทำงานแทนมนุษย์ได้อย่างมีประสิทธิภาพ ดังนั้นจากการศึกษางานวิจัยเกี่ยวกับการตรวจแก้คำผิดใน OCR ภาษาไทยเรามีหลักการที่สำคัญในการพัฒนา 3 หลักการดังนี้

1.1.1 โทเกนพาสซิ่งอัลกอริทึม (Token Passing)

ขบวนการโทเกนพาสซิ่งอัลกอริทึมเป็นกระบวนการสร้างคำที่เป็นไปได้โดยจะกระทำการตรวจสอบคำสะกดโดยจะนำตัวอักษรของคำเพิ่มเข้ามาในโทเกนทีละตัวแล้วนำไปเปรียบเทียบกับคำในพจนานุกรม เมื่อพบว่าเป็นคำหรือมีโอกาสที่จะเกิดเป็นคำ โทเกนนั้นก็จะยังเก็บเอาไว้ และถ้าโทเกนใดตรวจไม่พบคำหรือไม่มีโอกาสที่จะเกิดเป็นคำแล้วโทเกนนั้นก็จะถูกทิ้งไป สำหรับในงานวิจัยนี้ผลลัพธ์จาก OCR จะเป็นกลุ่มของตัวอักษรออกมา 5 ลำดับตัวอักษรซึ่งมีค่าความน่าจะเป็นของแต่ละตัวสูงกว่า 70% ขึ้นไป ด้วยเหตุนี้กรรมวิธีในการตรวจสอบคำสะกด จึงใช้เวลามากขึ้นเนื่องจากแต่ละตัวอักษร ถูกแทนที่ด้วยชุดของตัวอักษร ที่มีลักษณะใกล้เคียงกันทำให้เกิดคำที่เป็นไปได้หลายคำดังรูปที่ 1.1



รูปที่ 1.1 Word Graph ที่ได้จากโทเกนพาสซิ่งอัลกอริทึม

ในกรณีที่ตัวอักษรมีจำนวนหลายตัวอักษรซึ่งจะทำให้เกิดคำที่เป็นไปได้จำนวนมากและเมื่อนำมาสร้างเป็นประโยคก็จะทำให้เกิดประโยคที่เป็นไปได้จำนวนมาก ทำให้เวลาในการค้นหาประโยคที่ถูกต้องใช้เวลานาน ดังนั้นจากงานวิจัยที่ผ่านมาได้มีการนำเอาวิธีการ Prunning มาใช้เพื่อตัดบางคำที่มีค่าความน่าจะเป็นต่ำออกไป ซึ่งสามารถจะช่วยลดจำนวนประโยคที่เป็นไปได้จำนวนมาก แต่ในบางครั้งวิธีการนี้ก็ไม่ใช่วิธีการที่ดีนักเนื่องจากว่าในการตัดคำบางคำออกไปอาจจะเป็นคำที่ถูกต้อง ทำให้สูญเสียคำทีดีไปได้ วิธีการที่ดีที่สุดคือพยามหาวิธีการค้นหาที่ดีที่สุดที่สามารถหาประโยคที่ถูกต้องที่เป็นไปได้สูงสุดในทุกคำที่ได้มากจากวิธีการโทเกนพาสซิ่งอัลกอริทึม

1.1.2 **เจเนติกอัลกอริทึม** (Genetic Algorithm)

จากขบวนการโทเกนพาสซิ่งอัลกอริทึมเป็นกระบวนการสร้างคำที่เป็นไปได้โดยจะกระทำจากผลลัพธ์ของการรู้จำในกลุ่มตัวอักษรที่เป็นไปได้ทั้งหมดนั้นจะทำให้เกิดคำจำนวนมากที่เป็นไปได้ส่งผลทำให้เกิดประโยคที่เป็นไปได้เป็นไปได้จำนวนมาก ทำให้ยุ่งยากในการคัดเลือกประโยคที่ถูกต้อง ดังนั้นทางออกในการแก้ปัญหาในงานวิจัยนี้คือ นำเอาทฤษฎีเจเนติกอัลกอริทึมมาช่วยในการค้นหาประโยคที่ถูกต้องที่เป็นไปได้สูงสุดแทนวิธีการ Prunning

โดยทั่วไปการนำเอาทฤษฎีเจเนติกไปใช้ในงานด้านต่างๆ จะต้องมีการปรับปรุงรูปแบบของปัญหาในการนำเสนอเจเนติกอัลกอริทึมในลักษณะที่เหมาะสม เพราะเจเนติกอัลกอริทึมเป็นวิธีการค้นหาคำตอบโดยอาศัยวิธีการเลียนแบบการคัดเลือกทางธรรมชาติ และธรรมชาติทางพันธุกรรม โดยกระบวนการทางพันธุศาสตร์ คือการกำเนิดโครโมโซมใหม่โดยการผสมพันธุ์เพื่อถ่ายทอดยีนส์จากการครอสโอเวอร์(Crossover) หรือจากการมิวเตชั่น(Mutation) อันเป็นองค์ประกอบโครงสร้างของปัญหาที่ให้คำตอบที่ต้องการ ซึ่งอาศัยหลักการสุ่ม เพื่อปรับปรุงความสามารถในการค้นหาคำตอบที่ดียิ่งขึ้น การค้นหาคำตอบจากรุ่นหนึ่งไปรุ่นถัดไปตามวิวัฒนาการทางธรรมชาตินั้น คำตอบในรุ่นใหม่ที่เกิดขึ้นจากการสร้างความสัมพันธ์ของโครงสร้างต่างๆ ที่ประกอบด้วยค่าตัวแปรที่เหมาะสมดีในรุ่นก่อนๆ ดังนั้นจึงทำให้ได้ คำตอบที่ดีขึ้น จะเห็นได้ว่าวิธีการพื้นฐานของ เจเนติกอัลกอริทึมเป็นแบบการสุ่ม แต่มีหลักการและประสิทธิภาพจากการคาดเดาคำตอบใหม่จากสถิติคำตอบเดิมที่ดี

1.1.3 Language Model

จากการนำเอาค่าความน่าจะเป็นที่ได้จากผลลัพธ์ของ OCR มาใช้วิเคราะห์ความถูกต้องของประโยคเพียงอย่างเดียวไม่เพียงพอดังนั้นการเพิ่มประสิทธิภาพผลลัพธ์ของ OCR ให้ดียิ่งขึ้นจึงต้องมีการนำเอา Language Model มาใช้เพิ่มประสิทธิภาพในการวิเคราะห์ความถูกต้องของประโยค โดย Language Model คือแบบจำลองภาษา เป็นการพยามที่จะจับเอาลักษณะเฉพาะทางภาษาธรรมชาติ โดยทำให้อยู่ในรูปแบบซึ่งสามารถนำมาสร้างเป็นแบบจำลองได้ ภาษาธรรมชาติเป็นสิ่งที่มีความซับซ้อนโดยลักษณะของตัวมันอยู่ตามธรรมชาติ ซึ่งมันพัฒนาไปอย่างต่อเนื่องผ่านต่อไปจากรุ่นหนึ่งไปยังอีกรุ่นหนึ่งมาเป็นช่วงเวลานานโดย Language Model จะเป็นตัวบอกว่า ประโยคนั้นจัดเรียงเป็นอย่างไร เป็นไปได้ถูกต้องมากน้อยเพียงใด โดยกรรมวิธีหนึ่งซึ่งใช้ในการทำ Language Model ซึ่งมีความแม่นยำและมีความยืดหยุ่นสูง คือการนำ กรรมวิธีทางสถิติเข้ามาใช้โดยในการสร้างแบบจำลองจะใช้ค่าสถิติโดยใช้การฝึกฝน (training) corpora ของ text ขนาดใหญ่

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อพัฒนาความถูกต้องใน OCR ภาษาไทย ให้มีความถูกต้องเพิ่มขึ้น

2. เพื่อใช้ Genetic Algorithm มาช่วยเพิ่มความเร็วในการหาประโยคที่ถูกต้องจากกระบวนการ Token Passing

1.3 สมมติฐานของการศึกษา

สมมติฐานของงานวิจัยคือสามารถนำเอาทฤษฎีเจเนติกอัลกอริทึมมาช่วยในการหาประโยคที่ถูกต้องที่เป็นไปได้สูงสุด จากกระบวนการ Token Passing ได้อย่างมีประสิทธิภาพ

1.4 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้นำแนวความคิดในการหาประโยคที่ถูกต้องที่เป็นไปได้โดยเริ่มจากกระบวนการ Token Passing เพื่อสร้างกลุ่มคำที่เป็นไปได้ทั้งหมดจากนั้นนำเอาทฤษฎี Genetic Algorithm Holland(1973:88-105) มาช่วยในการหาประโยคที่เป็นไปได้สูงสุดโดยมีค่าความน่าจะเป็นของคำตอบจาก OCR และค่าความน่าจะเป็นที่คำนวณจาก Language Model เป็น Fitness Function Holland

1.5 ขอบเขตของการวิจัย

ในการวิจัยได้ทำการทดสอบประสิทธิภาพการหาประโยคที่ถูกต้องโดยเทียบกับ Full Search โดยมีรูปแบบจำลองภาษาและค่าความน่าจะเป็นของความถูกต้องที่ได้จาก OCR เป็นฟังก์ชั่นในการตัดสินใจคัดเลือกประโยคที่ถูกต้อง และลักษณะของคำที่ผิดที่เกิดในงานเอกสารที่เกิดจากความผิดพลาดจาก OCR จะเป็นลักษณะที่ผิดเฉพาะตัวอักษรที่ไม่ชัดเจน

1.6 ข้อตกลงเบื้องต้นของการวิจัย

1. การแก้ไขความผิดพลาดจาก OCR ได้แก้ไขเฉพาะความผิดพลาดที่เกิดในกรณีตัวอักษรที่ไม่ชัดเจนหรือตัวอักษรที่มีลักษณะใกล้เคียงกัน

2. คำศัพท์ที่ใช้ในงานวิจัยจะเป็นคำศัพท์ที่มีอยู่ในพจนานุกรมภาษาไทยเท่านั้นไม่สามารถใช้กับคำศัพท์ที่ไม่มีในพจนานุกรมภาษาไทยได้เช่นภาษาไทยที่เขียนทับศัพท์

1.7 โครงร่างของวิทยานิพนธ์

วิทยานิพนธ์นี้ประกอบด้วยเนื้อหาทั้งหมด 5 บท โดยแต่ละบทจะมีเนื้อหาดังนี้

บทที่ 1 เป็นการกล่าวถึงความเป็นมาของการทำวิจัยในการแก้คำผิดใน OCR ภาษาไทย ความมุ่งหมายและวัตถุประสงค์ของการศึกษา สมมติฐานของการศึกษา ทฤษฎีหรือแนวความคิดในการวิจัย และขอบเขตของการวิจัย

บทที่ 2 เป็นเอกสารและงานวิจัยที่เกี่ยวข้องทั้งหมด

บทที่ 3 เป็นวิธีการดำเนินงานวิจัย กล่าวถึงเครื่องมือที่ใช้ในงานวิจัย กระบวนการประยุกต์ทฤษฎีมาใช้ในงานวิจัยและอธิบายถึงโครงสร้างในการประยุกต์ใช้งาน

บทที่ 4 เป็นส่วนของผลการทดลองจากงานวิจัย

บทที่ 5 เป็นบทสรุปผลงานวิจัยของวิทยานิพนธ์นี้และข้อเสนอแนะแนวทางในการพัฒนาต่อไป

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ผู้วิจัยได้แบ่งเอกสารและงานวิจัยที่เกี่ยวข้องออกเป็นข้อๆ ดังนี้

2.1 Genetic Algorithm

2.2 Recognition

2.3 Thai OCR Error Correction

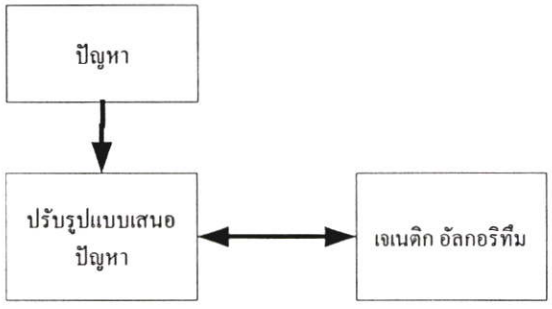
2.4 Token Passing Algorithm

2.5 Language Model

2.1 Genetic Algorithm

ในปัจจุบันนี้ปัญหาที่ต้องการคำตอบที่ดีที่สุด (Optimal Solution) ทางวิทยาศาสตร์ วิศวกรรมศาสตร์ คอมพิวเตอร์ หรือในการทำงานต่างๆ ที่เกิดขึ้นมามากมายนั้น สามารถหาคำตอบได้หลายๆ วิธี ซึ่งแตกต่างกันไปตามชนิดของปัญหาความคิด เทคนิค วิธีการวิเคราะห์ปัญหานั้นๆ และความแพร่หลายในการพัฒนาศักยภาพของคอมพิวเตอร์ให้รู้จักเรียนรู้เพื่อช่วยหาคำตอบหรือช่วยตัดสินคำตอบในขั้นต้นมีมากขึ้น โดยปัจจุบันนี้นักวิทยาศาสตร์ได้เริ่มนำความรู้เกี่ยวกับทฤษฎีหรือกฎเกณฑ์ทางธรรมชาติมาช่วยในการศึกษาวิจัย เช่นนิวรอลเน็ตเวิร์ก(Neural Network) ฟัซซี่ลอจิก(Fuzzy Logic) เป็นต้น เจเนติกอัลกอริทึมเป็นอีกวิธีหนึ่งที่จำลองรูปแบบวิธีการทางชีววิทยา ในการให้กำเนิดประชากรรุ่นใหม่ หรือขยายเผ่าพันธุ์ในรุ่นลูก รุ่นหลานต่อไป ซึ่งอาศัยพื้นฐานความคิดของวิวัฒนาการทางธรรมชาติถ่ายทอดลักษณะต่างๆ ทางพันธุกรรม โดยปฏิบัติตามกระบวนการทาง พันธุศาสตร์ เพื่อจะใช้ในการหาคำตอบที่ดีที่สุดหรือใกล้เคียงที่สุดของปัญหาโดยคอมพิวเตอร์

Holland (1973 : 88-105) เริ่มสนใจศึกษาในทฤษฎีวิวัฒนาการทางธรรมชาติ ในการกำเนิดประชากร สิ่งมีชีวิตในรุ่นๆ ต่อไป โดยกระบวนการธรรมชาติทางชีววิทยาประกอบด้วยการคัดเลือกทางธรรมชาติ คือสิ่งมีชีวิตใดแข็งแรงกว่าย่อมมีโอกาสอยู่รอดได้มากกว่านั่นหมายถึงการมีโครโมโซมซึ่งประกอบด้วยยีนส์ต่างๆ ที่มีลักษณะที่ดี นั้นจะมีโอกาสอยู่รอดได้มากกว่าโครโมโซมที่สามารถอยู่รอดได้ก็จะถูกถ่ายทอดยีนส์ที่มีลักษณะที่ดีเหล่านั้นไปยังลูกหลานได้มากกว่าเช่นกัน และกระบวนการทางพันธุศาสตร์ คือการกำเนิดโครโมโซมใหม่โดยการผสมพันธุ์เพื่อถ่ายทอดยีนส์จากการครอสโอเวอร์(Crossover) หรือจากการมิวเตชั่น(Mutation)



รูปที่ 2.1 หลักการเบื้องต้นของทฤษฎีเจเนติกอัลกอริทึม (Genetic Algorithm)

จากรูปที่ 2.1 แสดงหลักการเบื้องต้น โดยจะต้องมีการปรับปรุงรูปแบบของปัญหาในการนำเสนอเจเนติกอัลกอริทึมในลักษณะที่เหมาะสมเพราะเจเนติกอัลกอริทึมเป็นวิธีการค้นหาคำตอบโดยอาศัยวิธีการเลียนแบบการคัดเลือกทางธรรมชาติ และธรรมชาติทางพันธุกรรม โดยการรวมกันหรือสลับเปลี่ยนตัวแปรต่างๆ อันเป็นองค์ประกอบโครงสร้างของปัญหาที่ให้คำตอบที่ต้องการ ซึ่งอาศัยหลักการสุ่ม เพื่อปรับปรุงความสามารถในการค้นหาคำตอบที่ดียิ่งขึ้น การค้นหาคำตอบจากรุ่นหนึ่งไปรุ่นถัดไปตามวิวัฒนาการทางธรรมชาตินั้น คำตอบในรุ่นใหม่ที่เกิดขึ้นจากการสร้างความสัมพันธ์ของโครงสร้างต่างๆ ที่ประกอบด้วยค่าตัวแปรที่เหมาะสมดีในรุ่นก่อนๆ ดังนั้นจึงทำให้ได้คำตอบที่ดีขึ้น จะเห็นได้ว่าวิธีการพื้นฐานของ เจเนติกอัลกอริทึมเป็นแบบการสุ่ม แต่มีหลักการและประสิทธิภาพจากการคาดเดาคำตอบใหม่จากสถิติคำตอบเดิมที่ดีซึ่งแตกต่างจากวิธีการทั่วไปคือ

1. ค้นหาคำตอบภายใต้โครงสร้างของปัญหาอันเกิดจากการกำหนดรหัส (Coding) รูปแบบโครงสร้างจากกลุ่มตัวแปรต่างๆ ของปัญหานั้น ไม่ใช่ค้นหาคำตอบจากค่าของกลุ่มตัวแปรนั้น
2. ค้นหาคำตอบโดยพิจารณาจากประชากรคำตอบ หรือกลุ่มคำตอบ ไม่ใช่พิจารณาจากคำตอบใดคำตอบหนึ่ง
3. ค้นหาคำตอบจากผลลัพธ์ของกลุ่มค่าตัวแปรที่เป็นฟังชั่นเป้าหมายของปัญหา
4. ค้นหาคำตอบโดยอาศัยการถ่วงน้ำหนักและความเหมาะสมของแต่ละคำตอบจากกลุ่มคำตอบนั้นๆ

2.1.1 ฟังก์ชันเป้าหมายกับฟังก์ชันความเหมาะสม

การหาคำตอบที่ดีที่สุดของเจเนติกอัลกอริทึมเป็นการนำผลลัพธ์ที่ได้จากการหาคำตอบครั้งก่อนมาปรับปรุงให้ดีขึ้น วิธีการของเจเนติกอัลกอริทึมจะไม่พิจารณาจากขั้นตอนของการแก้ปัญหา แต่จะพิจารณาโดยตัดสินว่าคำตอบใหม่ที่ได้รับนั้นดีขึ้นหรือไม่ หรือเป็นคำตอบที่ใกล้เคียง

กับคำตอบที่ต้องการหรือไม่จากฟังก์ชันเป้าหมาย (Object Function : f) ในแต่ละปัญหาจะสามารถกำหนดฟังก์ชันเป้าหมายได้ตามรูปแบบของปัญหา โดยฟังก์ชันเป้าหมายเป็นฟังก์ชันที่แสดงความสัมพันธ์ของแต่ละตัวแปร พารามิเตอร์ เงื่อนไข หรือข้อกำหนดต่าง ๆ ของปัญหานั้นที่ระบุคำตอบใดคำตอบหนึ่งที่สามารถเป็นไปได้ ณ ค่าพารามิเตอร์ เงื่อนไข หรือข้อกำหนดชุดดังกล่าว สำหรับฟังก์ชันความเหมาะสม (Fitness Function : F) เป็นฟังก์ชันที่กำหนดค่าความเหมาะสม (fitness) ของแต่ละโครโมโซม เปรียบเสมือนค่าความสามารถในการอยู่รอดของแต่ละโครโมโซมและเป็นฟังก์ชันที่กำหนดโอกาส หรือสัดส่วนที่แต่ละโครโมโซมเหมาะสมจะถูกคัดเลือกไปใช้มากน้อยเพียงใด นั่นคือ ฟังก์ชันความเหมาะสมจะเป็นฟังก์ชันที่เสดงถึงค่าคำตอบที่เกิดขึ้นจากชุดตัวแปรของปัญหาในโครโมโซมนั้นว่าดีเพียงใด โดยทั่วไปแล้วเรามักใช้ฟังก์ชันเป้าหมายเป็นฟังก์ชันความเหมาะสม หรืออาจใช้ฟังก์ชันเป้าหมายที่ถูกปรับให้เหมาะสมในการใช้เจเนติกอัลกอริทึมเป็นฟังก์ชันความเหมาะสมก็ได้

2.1.2 รูปแบบโครโมโซม

จุดเริ่มต้นของการจำลองแบบทางธรรมชาติของเจเนติกอัลกอริทึมเพื่อใช้แก้ปัญหาเริ่มจากการมองปัญหาเทียบเท่ากับโครโมโซมชนิดหนึ่ง ประกอบด้วยยีนส์ลักษณะต่าง ๆ ซึ่งหมายถึงข้อมูลต่าง ๆ เมื่อแปลความหมายแล้วจะให้ค่าของคำตอบค่าหนึ่ง ในเจเนติกอัลกอริทึมยีนส์ที่อยู่ในโครโมโซมเป็นตัวแสดงค่าคำตอบคำตอบหนึ่งของปัญหา ที่แปรผันไปตามการประยุกต์ใช้งาน ซึ่งโดยทั่วไปยีนส์หมายถึงตัวแปร พารามิเตอร์ เงื่อนไข หรือ ข้อกำหนด ต่าง ๆ ที่เป็นองค์ประกอบของปัญหา การกำหนดรูปแบบของแต่ละโครโมโซมทำได้โดยการแปลงตัวแปร พารามิเตอร์ เงื่อนไข หรือข้อกำหนดต่าง ๆ ให้อยู่ในรูปลำดับของยีนส์บนโครโมโซมหรือเรียกว่าสตริง(String) อันประกอบด้วยบิท(Bit) หรือเรียกว่าอักขระ(Character) ซึ่งลักษณะต่าง ๆ ที่เป็นไปได้ของแต่ละยีนส์คือค่าของบิท(Bit Value) หรือค่าตัวแปร พารามิเตอร์ ต่างๆ ที่เป็นไปได้ การกำหนดรูปแบบของปัญหาให้เป็นไปตามธรรมชาติ โดยกำหนดรหัสในรูปแบบตัวเลขหรือตัวอักษรในช่วงที่จำกัดตามค่าตัวแปรหรือพารามิเตอร์ และประกอบรวมกันของยีนส์หรือโครโมโซมที่มีความยาวคงที่ เช่น หากต้องการหาค่าสูงสุดของฟังก์ชันทางคณิตศาสตร์ $y = x^2$ ที่ x เป็นจำนวนเต็มอยู่ในช่วง [0,31] แล้ววิธีการของเจเนติกอัลกอริทึมสามารถแก้ปัญหาได้โดยการกำหนดรูปแบบโครโมโซมจากการกำหนดรหัสตัวแปร x เป็นตัวไบนารี 0 หรือ 1 จำนวน 5 ตำแหน่ง ซึ่ง x จะมีค่าตั้งแต่ 00000 ถึง 11111 เป็นค่า 0 ถึง 31 ตามต้องการ เป็นต้น

2.1.3 วัฏจักรเจเนติกอัลกอริทึม

เมื่อกำหนดรูปแบบโครโมโซมและฟังก์ชันความเหมาะสมของปัญหาแล้ว เจเนติกอัลกอริทึมจะสามารถประมวลผลหาคำตอบของปัญหาได้ โดยสร้างวิวัฒนาการกลุ่มคำตอบในรุ่นต่อไปตามวัฏจักรการทำงานของเจเนติกอัลกอริทึม (Genetic Algorithm Cycle) ซึ่งมี 4 ขั้นตอน คือ

1. สร้างประชากรโครโมโซมรุ่นเก่า ตามรูปแบบโครโมโซมที่กำหนดไว้ โดยประชากรต้นกำเนิด (Initial Population) เกิดจากการสร้างชุดโครโมโซมต้นกำเนิดด้วยการสุ่มสร้างค่าแต่ละบิทของแต่ละโครโมโซม

2. วิเคราะห์ค่าความเหมาะสมแต่ละโครโมโซม โดยถอดรหัสค่าตัวแปร พารามิเตอร์ต่าง ๆ ของแต่ละบิทในโครโมโซม และคำนวณค่าความเหมาะสมจากฟังก์ชันความเหมาะสมที่กำหนดไว้

3. สร้าง Mating pool คือชุดโครโมโซมต้นแบบหรือชุดโครโมโซมพ่อแม่ ที่สามารถอยู่รอดเป็นต้นแบบ ซึ่งอาศัยการจำลองการคัดเลือกทางธรรมชาติ โดยพิจารณาถ่วงน้ำหนักจากค่าความเหมาะสมของแต่ละโครโมโซม หากโครโมโซมใดมีค่าความเหมาะสมมากก็จะมีโอกาสถูกคัดเลือกเป็นต้นแบบมาก

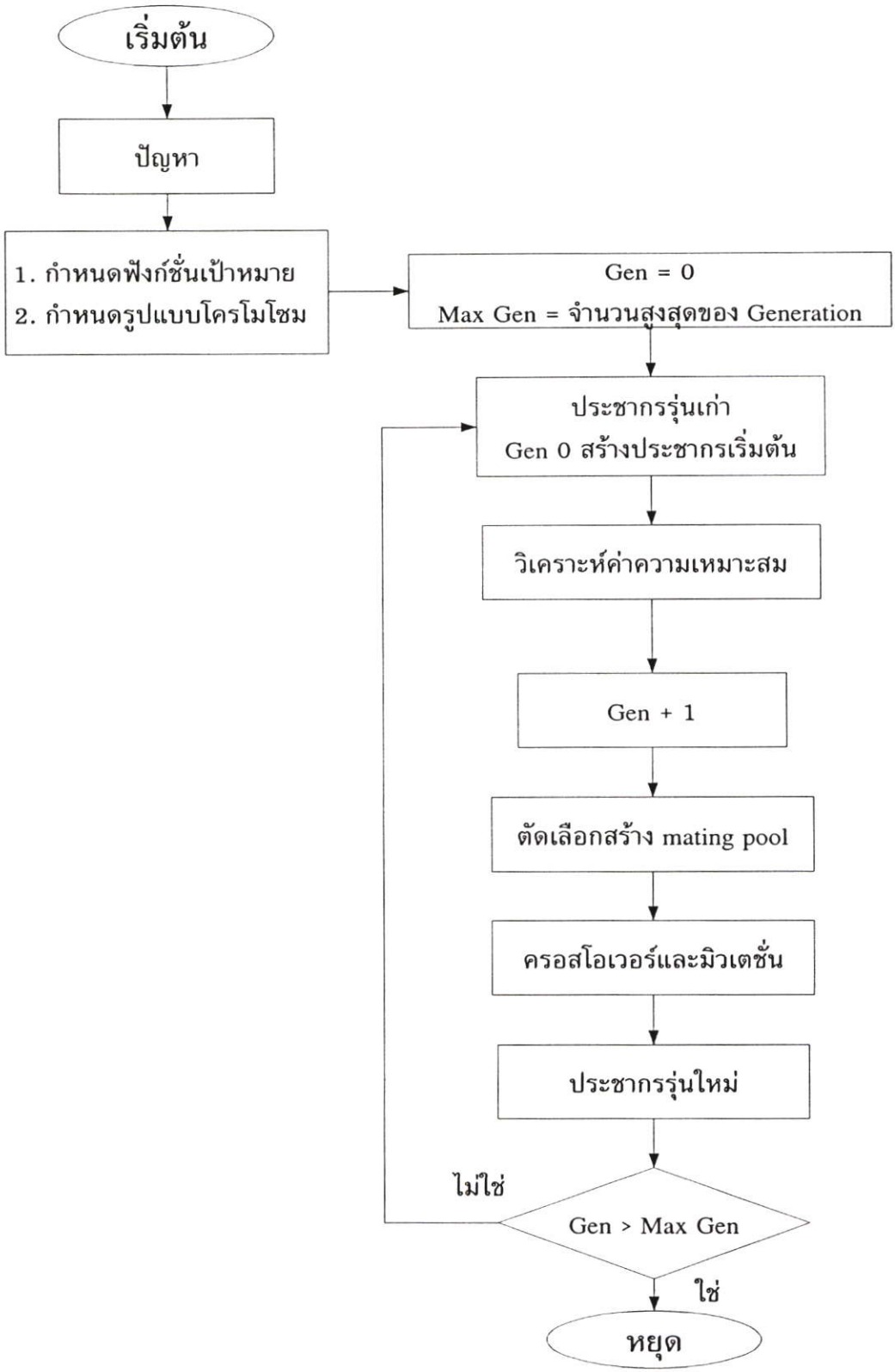
4. ดำเนินการทางพันธุศาสตร์โดยจับคู่โครโมโซมต้นแบบใน Mating Pool เพื่อสร้างประชากรในโครโมโซมรุ่นใหม่ ซึ่งตัวดำเนินการทางพันธุศาสตร์ประกอบด้วย ครอสโอเวอร์ โดยแลกเปลี่ยนค่ายีนส์บางส่วนของโครโมโซมซึ่งกันและกันหรือมิวเตชันโดยการสุ่มเปลี่ยนค่ายีนส์บางยีนส์ของแต่ละโครโมโซม เป็นต้น

การค้นหาคำตอบของเจเนติกอัลกอริทึม จะประมวลผลซ้ำตามวัฏจักรจนกว่าจะได้รับคำตอบที่พอใจตามเกณฑ์ที่ตั้งไว้ หรือในระยะเวลาตามจำนวนรุ่นที่ดำเนินการตามต้องการซึ่งแสดงอัลกอริทึมการทำงานของเจเนติกอัลกอริทึมดังนี้

อัลกอริทึมของเจเนติกอัลกอริทึม

```
{
        t := 0;
        Initpopulation P(t);
        // สร้างประชาการโครโมโซมต้นกำเนิดโดยการสุ่ม
        Evaluate P(t);
        // วิเคราะห์ค่าความเหมาะสมแต่ละโครโมโซมประชากรต้นกำเนิด
        While not terminate
        // ตรวจสอบเงื่อนไขความพอใจ
          Begin
                t := t + 1;
                P'(t) := Selectparents P(t-1);
                // คัดเลือกโครโมโซมต้นแบบจากประชาการรุ่นก่อน
                Recombine P'(t);
                // แลกเปลี่ยนส่วนยีนส์ภายในโครโมโซมต้นแบบ
                Mutate P'(t);
                // มิวเตชั่นโครโมโซมต้นแบบ
                Evaluate P'(t);
                // วิเคราะห์ค่าความเหมาะสมของประชากรรุ่นใหม่
                P(t) := P'(t);
                // ประชาการรุ่นใหม่กลายเป็นประชากรรุ่นเก่าต่อไป
          End;
}
```

ยุคแรก ๆ ของการเริ่มใช้งานเจเนติกอัลกอริทึมจะเป็นเจเนติกอัลกอริทึมแบบง่าย (Simple Genetic Algorithm : SGA) ซึ่งมีพื้นฐานและมีกระบวนการไม่มากนัก ง่ายในการศึกษาทำความเข้าใจในแต่ละขั้นตอน การทำงานของเจเนติกอัลกอริทึมแบ่งเป็นสองส่วน ดังไดอะแกรมในรูปที่ 2.2



รูปที่ 2.2 ไดอะแกรมการทำงานของทฤษฎีเจเนติกอัลกอริทึมแบบง่าย

ส่วนแรกคือขั้นตอนการเตรียมการ ส่วนที่สองคือขั้นตอนการทำงาน สำหรับส่วนของขั้นตอนการเตรียมการนั้นเป็นส่วนของการปรับรูปแบบของปัญหาให้เหมาะสมสำหรับการใช้เจเนติกอัลกอริทึม ประกอบด้วย 2 ส่วนดังนี้

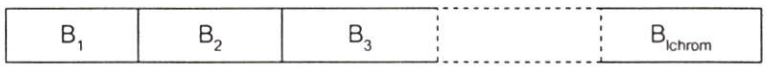
1. กำหนดฟังก์ชันความเหมาะสม เพื่อความสะดวกและง่ายต่อความเข้าใจขั้นตอนการทำงานต่าง ๆ จะกำหนดตัวอย่างปัญหาสำหรับอธิบายรายละเอียดการหาคำตอบของเจเนติกอัลกอริทึมแบบง่าย คือ ปัญหาการหาค่าสูงสุดของฟังก์ชัน $y = x^2$ ที่ x มีค่าระหว่างจำนวนเต็ม I[0,31] ดังนั้น

ตัวอย่าง : ฟังก์ชันเป้าหมายคือ $f = x^2$

กำหนดให้ฟังก์ชันความเหมาะสม คือ $F = x^2$

ซึ่งคำตอบที่ดีที่สุดคือค่า x ที่มีค่าความเหมาะสมสูงสุด (MAX (F))

2. กำหนดรูปแบบโครโมโซม รูปแบบของโครโมโซมที่จะใช้กับปัญหานี้ เป็นแบบไบนารี โดยค่าตัวแปรหรือพารามิเตอร์ของปัญหาจะถูกแปลงให้อยู่ในรูปของไบนารีโครโมโซม คือ ประกอบด้วย บิทที่มีค่าเป็น 0 หรือ 1 ซึ่งเป็นค่าในเลขฐานสอง และมีความยาว (Chromosome Length : lchrom) ตามแต่จะกำหนด ซึ่งแสดงด้วยสัญลักษณ์ได้ดังนี้



รูปที่ 2.3 ตัวอย่างรูปแบบของโครโมโซม

ซึ่ง $B_i \in I[0,1]$

ตัวอย่าง : วิธีการเข้ารหัสแบบไบนารีโดยแปลงค่าพารามิเตอร์ x ให้อยู่ในรูปไบนารี 5 บิท (lchrom = 5) ดังนั้นโครโมโซมของปัญหาจะมีค่าอยู่ในช่วง 00000 ถึง 11111 ซึ่งเมื่อถอดรหัส แล้วจะทำให้ x มีค่าอยู่ในช่วง 0 ถึง 31 ตามที่ต้องการ

ในส่วนของรายละเอียดขั้นตอนการทำงานของ เจเนติกอัลกอริทึมแบบง่ายจะเป็นขั้นตอนพื้นฐานเบื้องต้นแบบง่ายประกอบด้วย 4 ส่วนดังนี้

1. ประชากรรุ่นเก่า (Old Population) เป็นชุดโครโมโซมที่จะถูกคัดเลือกไปเป็นต้นแบบสำหรับสร้างประชากรรุ่นใหม่ (New Population) ในวิวัฒนาการ (Generation : gen) รุ่นต่อไป โดยประชากรเริ่มต้นที่ gen = 0 จะถูกสร้างขึ้นโดยการสุ่มตามจำนวนโครโมโซมในแต่ละรุ่น (Population Size : popsize) ที่กำหนด

ตัวอย่าง : ลำดับ	โครโมโซม
1	0 1 1 1 0
2	1 1 0 0 1
3	0 1 0 0 0
4	1 0 0 1 1

ชุดโครโมโซมรุ่นนี้เป็นชุดโครโมโซมรุ่นเริ่มต้นที่กำหนดให้ในแต่ละรุ่นประกอบด้วยจำนวนโครโมโซม จำนวน 4 โครโมโซม โดยแต่ละโครโมโซมประกอบด้วย ค่าไบนารี 0 หรือ 1 ที่เกิดจากการสุ่ม จำนวน 5 ครั้ง

2. วิเคราะห์ค่าความเหมาะสม เป็นขั้นตอนการถอดรหัสจากรูปแบบโครโมโซมที่กำหนดไว้ เพื่อคำนวณค่าความเหมาะสมตามฟังก์ชันความเหมาะสมของปัญหา ในที่นี้ฟังก์ชันเป้าหมายหรือฟังก์ชันความเหมาะสม คือ $F = x^2$ ดังนั้นการวิเคราะห์ค่าความเหมาะสมจึงเป็นการถอดรหัสเลขฐานสองของแต่ละโครโมโซมเป็นค่าตัวแปร x และคำนวณค่าความเหมาะสม คือ ค่า x^2 ซึ่งจะเห็นได้ว่าชุดโครโมโซมรุ่นเริ่มต้นมีค่าความเหมาะสมเป็น 196, 265, 64 และ 361 ตามลำดับ

ตัวอย่าง : ลำดับ	โครโมโซม	x	ค่าความเหมาะสม
1	0 1 1 1 0	14	196
2	1 1 0 0 1	25	625
3	0 1 0 0 0	8	64
4	1 0 0 1 1	19	361

3. การคัดเลือก เป็นขั้นตอนที่จำลองการคัดเลือกทางธรรมชาติเพื่อสร้าง Mating Pool โดยคัดเลือกชุดโครโมโซมรุ่นเก่าให้เป็นโครโมโซมต้นแบบ หรือ โครโมโซมพ่อ-แม่ เพื่อใช้สร้างโครโมโซมรุ่นลูกเป็นรุ่นต่อไป การคัดเลือกของเจเนติกอัลกอริทึมแบบง่าย เป็นแบบอ้างอิงค่าความเหมาะสม (Fitness-based Selection) โดยใช้ค่าความเหมาะสมเป็นตัวตัดสินว่า โครโมโซมใดในรุ่นเก่ามีโอกาสจะถูกเลือกเป็นโครโมโซมพ่อ-แม่มากน้อยเพียงใด โครโมโซมที่มีค่าความเหมาะสมที่ดีจะถูกกำหนดน้ำหนักค่าความน่าจะเป็นที่จะถูกเลือกแต่ละครั้งสูง การกำหนดค่าความน่าจะเป็นที่จะถูกเลือกต่อการสุ่มเลือกแต่ละครั้ง (Probability of Selected Value : pselect) ของแต่ละโครโมโซมกำหนดจากค่าความเหมาะสม เทียบกับผลรวมของค่าความเหมาะสมทั้งหมดดังสมการ

$$pselect_i = \frac{F_i}{\sum F}$$

ซึ่งสามารถคำนวณค่าที่คาดหวังว่าจะสุ่มได้ (Expected Value : E)ของแต่ละโครโมโซมในแต่ละรุ่น ดังสมการ

$$E_i = pselect_i * popsize = \frac{F_i}{F}$$

สำหรับวิธีการสุ่มโครโมโซมต้นแบบของเจเนติกอัลกอริทึมแบบง่ายนั้นเป็นแบบจำลองการหมุนวงล้อถ่วงน้ำหนัก (Roulette Wheel : RW) ซึ่งกำหนดขนาดแต่ละช่องของวงล้อนั้นตามความน่าจะเป็นที่จะสุ่มได้ในแต่ละครั้งของแต่ละโครโมโซมซึ่งมีวิธีการดังนี้

(1) หาค่าความเหมาะสมของแต่ละโครโมโซม

(2) หาค่าความน่าจะเป็นที่จะสุ่มได้ในแต่ละครั้งของแต่ละโครโมโซม

(3) หาค่าความถี่สะสม (q) ของค่าความน่าจะเป็นของแต่ละโครโมโซม ดังสมการ

$$q_i = \sum_{j=1}^{i} pselect_j$$

(4) สร้างเลขสุ่มจำนวนจริง (r) มีค่าอยู่ในช่วง [0.0, 1.0]

(5) เลือกโครโมโซมลำดับที่ r ซึ่ง r มีค่าอยู่ระหว่าง q_{i-1} และ q_i

ตารางที่ 2.1 ตัวอย่างของการหาค่าของสมการ

ลำดับ	โครโมโซม	X	ค่าความเหมาะสม (F)	ค่าความน่าจะเป็น ($pselect_i$)	จำนวนที่คาดหวัง (E_i)	จำนวนที่สุ่มได้จาก RW
1	0 1 1 1 0	14	196	0.157	0.628	1
2	1 1 0 0 1	25	625	0.502	2.008	2
3	0 1 0 0 0	8	64	0.051	0.204	0
4	1 0 0 1 1	19	361	0.290	1.160	1
	รวม		1246	1.000	4.000	
	ค่าเฉลี่ย		312	0.250	1.000	
	ค่าสูงสุด		625	0.502	2.008	

ตัวอย่างของการกำหนดค่าความน่าจะเป็นโดยกำหนดจากค่าความเหมาะสมเทียบกับผลรวมของค่าความเหมาะสมทั้งหมด จะเห็นได้ว่าในการคัดเลือกโครโมโซมต้นแบบจาก 4 โครโมโซมนี้ โอกาสที่จะสุ่มได้โครโมโซมลำดับที่ 1 ต่อการสุ่มแต่ละครั้งเท่ากับ 0.157 และโอกาสที่จะสุ่มได้

โครโมโซมลำดับที่ 2,3,4 ต่อการสุ่มแต่ละครั้งเท่ากับ 0.502, 0.051,และ 0.290 ตามลำดับ และจำนวนโครโมโซมต้นแบบที่สุ่มได้โดยจำลองการหมุนวงล้อดังนี้

ตารางที่ 2.2 ตัวอย่างของโครโมโซมต้นแบบ

ลำดับโครโมโซม	1	2	3	4
ค่าความเหมาะสม (F)	196	625	64	361
ค่าความน่าจะเป็นที่สุ่มได้แต่ละครั้ง ($pselect_i$)	0.157	0.502	0.051	0.290
ความถี่สะสมค่าความน่าจะเป็น (q_i)	0.157	0.659	0.710	1.000
สร้างเลขสุ่มในการหมุนวงล้อแต่ละครั้ง (r)	0.333	0.844	0.456	0.128
ลำดับโครโมโซมที่ถูกเลือก ($q_{i-1} \leq r \leq q_i$)	2	4	2	1

ซึ่งจำนวนที่สุ่มได้เป็นโครโมโซมต้นแบบใน mating pool ของแต่ละโครโมโซมเป็น 1, 2, 0 และ 1 ตามลำดับ จะเห็นได้ว่าโครโมโซมลำดับที่ 2 มีค่าความเหมาะสมสูงที่สุดจะมีโอกาสถูกคัดเลือกในจำนวนที่มากที่สุด ส่วนโครโมโซมลำดับที่ 3 มีค่าความเหมาะสมต่ำมากจึงมีโอกาสที่จะไม่ถูกเลือก

4. ดำเนินการทางพันธุศาสตร์ เป็นขั้นตอนที่จำลองแบบธรรมชาติทางพันธุกรรม ซึ่งตัวดำเนินการทางพันธุศาสตร์ของเจเนติกอัลกอริทึมแบบง่าย คือ ครอสโอเวอร์และมิวเตชันซึ่งมีรายละเอียดดังนี้

<u>ครอสโอเวอร์</u>

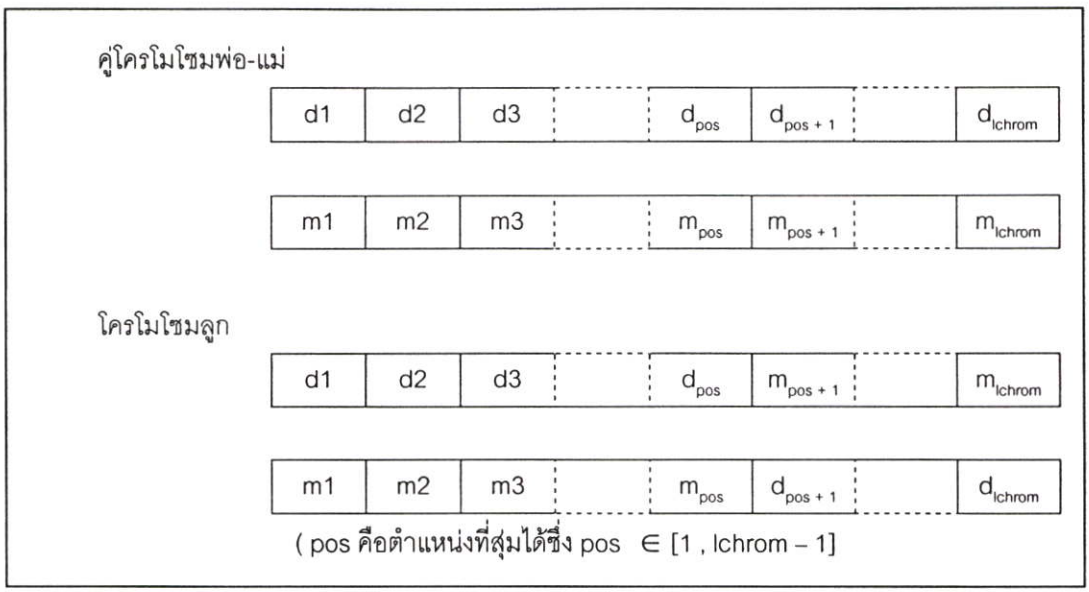
เป็นตัวดำเนินการในการแลกเปลี่ยนส่วนของโครโมโซมพ่อ-แม่ ตามการกำหนดอัตราความน่าจะเป็นของการครอสโอเวอร์ (Probability of Crossover : P_c) เพื่อสร้างชุดโครโมโซมรุ่นใหม่หรือโครโมโซมลูก มีขั้นตอนการทำงานคือ

ขั้นตอนแรก : สุ่มจับคู่โครโมโซมพ่อ-แม่ ใน mating pool ที่สร้างไว้จากการคัดเลือก

ขั้นตอนที่สอง: สร้างเลขสุ่มจำนวนจริง (r) มีค่าอยู่ในช่วง [0.0 , 1.0] โดยถ้า $r \leq P_c$ แล้วโครโมโซมพ่อ-แม่นั้นจึงจะมีการครอสโอเวอร์

ขั้นตอนที่สาม: ครอสโอเวอร์โดยการแลกเปลี่ยนส่วนของคู่โครโมโซมพ่อ-แม่นั้น ซึ่งการครอสโอเวอร์ของเจเนติกอัลกอริทึมแบบง่าย นั้นเป็นการครอสโอเวอร์แบบ 1 จุด (One-point Crossover)แสดงดังรูปที่ 2.4 ดังนี้

- สุ่มเลือกตำแหน่ง pos ซึ่งเป็นตำแหน่งที่จะครอสโอเวอร์ โดย pos มีค่าอยู่ในช่วง [1 ,lchrom-1]
- แลกเปลี่ยนค่าในแต่ละบิทของคู่โครโมโซมพ่อ-แม่ตั้งแต่ตำแหน่งที่ pos + 1 ถึง lchrom ซึ่งจะทำให้เกิดโครโมโซมลูกใหม่ 2 โครโมโซม

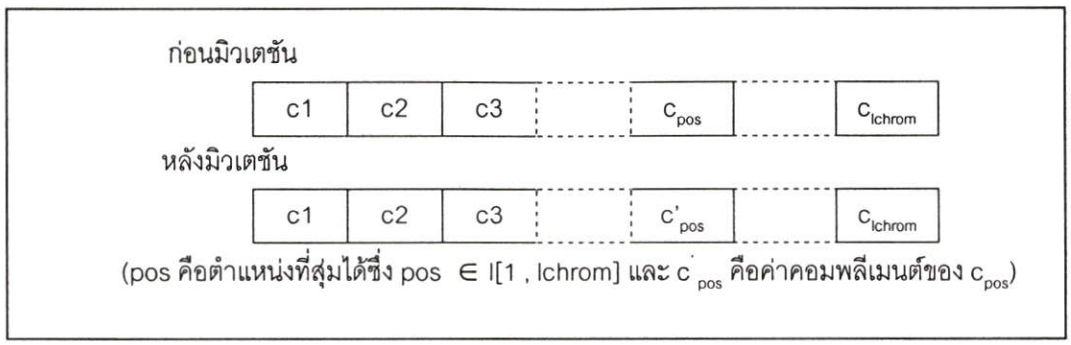


รูปที่ 2.4 ครอสโอเวอร์แบบ 1 จุด

จำนวนการครอสโอเวอร์ในแต่ละรุ่นดำเนินการขึ้นอยู่กับการกำหนดค่า P_c ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรแต่ละรุ่น popsize เท่ากับ 30 โครโมโซม และกำหนดให้ P_c = 0.6 แล้วจำนวนการครอสโอเวอร์ในแต่ละรุ่นเท่ากับ P_c*(popsize / 2) = 0.6 * (30 / 2) = 9 ครั้ง (การครอสโอเวอร์หนึ่งครั้งเกิดจากโครโมโซมสองโครโมโซม)

<u>มิวเตชัน</u>

เป็นตัวดำเนินการผ่าเหล่าตัวหนึ่งที่อาจช่วยให้โครโมโซม มีค่าความเหมาะสมดีขึ้นหลังจาก ครอสโอเวอร์ โดยกลับค่าบิทเป็นค่าใหม่ในตำแหน่งบิทที่สุ่มได้ ตามอัตราความน่าจะเป็นของการมิวเตชันในแต่ละบิท (Probability of Mutation : P_m) ที่กำหนด สำหรับการมิวเตชันของเจเนติกอัลกอริทึมแบบง่ายนั้นเป็นแบบไบนารีมิวเตชัน (Binary Mutation) โดยกลับค่าบิทเป็นค่าคอมพลีเมนต์คือจาก 0 เป็น 1 หรือ จาก 1 เป็น 0 ดังรูปที่ 2.5



รูปที่ 2.5 ไบนารีมิวเตชัน

จำนวนการมิวเตชันในแต่ละรุ่นขึ้นอยู่กับการกำหนดค่า P_m ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรแต่ละรุ่น popsize เท่ากับ 30 โครโมโซม ซึ่งแต่ละโครโมโซมประกอบด้วย 5 บิท และกำหนดให้ P_m = 0.02 แล้วจำนวนการมิวเตชันในแต่ละรุ่นเท่ากับ P_m * popsize * lchrom = 0.02 * 30 * 5 = 3 บิท

5. ประชากรรุ่นใหม่ เป็นชุดโครโมโซมลูกที่เกิดจากขั้นตอนของการวิวัฒนาการต่าง ๆ ทั้งหมด ซึ่งประชากรรุ่นใหม่ทั้งหมดที่เกิดขึ้น จะถูกถ่ายทอดกลายเป็นประชากรรุ่นเก่าสำหรับวิวัฒนาการในรุ่นถัดไป ซึ่งเรียกวิวัฒนาการแบบนี้ว่า การถ่ายทอดแบบทั่วไปหรือรีโพรดักชันแบบทั่วไป (General Reproduction) กระบวนการต่าง ๆ จะถูกปฏิบัติซ้ำ ๆ จนกระทั่งถึงรุ่นที่มากที่สุด(max generation)ที่ต้องการ

2.1.4 การประยุกต์ทฤษฎีเจเนติกอัลกอริทึม

เมื่อวิเคราะห์การทำงานของ SGA ซึ่งเป็น เจเนติกอัลกอริทึม(GA) ในยุคแรกๆ แล้วจะเห็นว่า วิธีการของ GA เป็นการหาคำตอบแบบสุ่ม ซึ่งเป็นวิธีการที่ไม่มีการบันทึกหรือจดจำคำตอบที่ดีที่สุดของรุ่นก่อน จึงทำให้การหาคำตอบของ SGA ได้คำตอบที่ดีมากขึ้นหรือน้อยลง ดังนั้นหากสามารถพัฒนาวิธีการค้นหาคำตอบของ GA ให้ดีขึ้นแล้วจะเป็นการปรับปรุงสมรรถนะของ GA ยิ่งขึ้น สำหรับหาค่าสูงสุดของฟังชั่น $y = x^2$ เราปรับปรุงการค้นหาคำตอบของ GA โดย

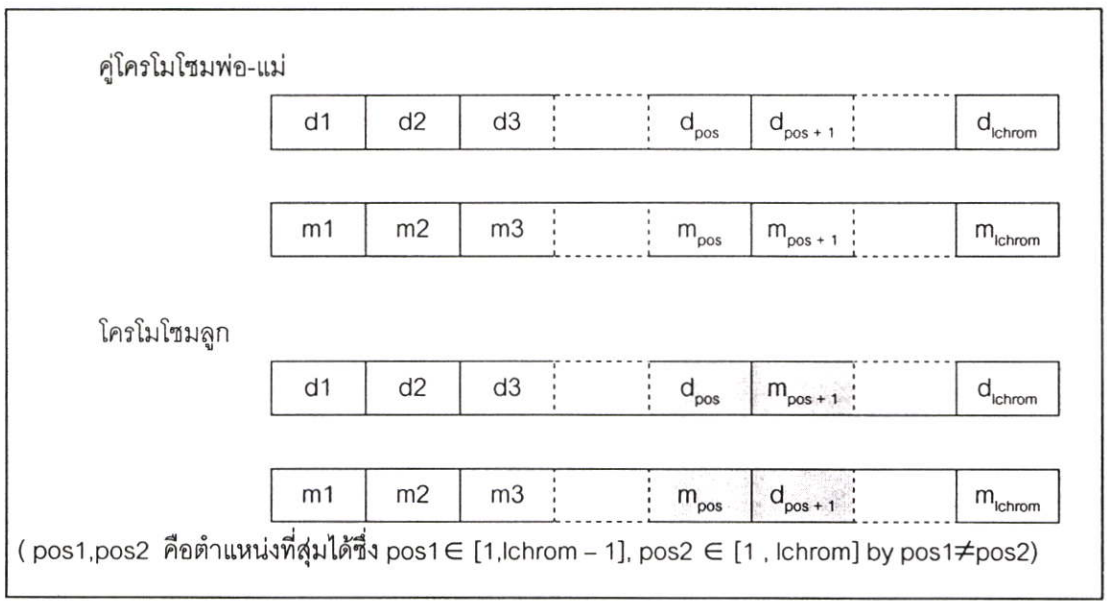
1. รีโพรดักชั่นแบบรักษาค่าความเหมาะสมที่ดี

เนื่องจากในการค้นหาคำตอบของ SGA นั้น มีโอกาสที่จะสูญเสียโครโมโซมรุ่นเก่า ที่มีค่าความเหมาะสมที่ดีไปได้ ซึ่งจะทำให้คำตอบในรุ่นถัดไปนั้นดีมากหรือน้อยลง ดังนั้นหากปรับปรุง SGA ให้ควบคุมการค้นหาคำตอบ โดยรักษาโครโมโซมที่ดีไว้แล้ว จะช่วยวิวัฒนาการหาคำตอบในรุ่นถัดไปดีขึ้นเรื่อยๆ โดยมีวิธีดังนี้

- กำหนดจำนวนโครโมโซมที่ดีที่สุด ของรุ่นเก่าที่ต้องการรักษาเป็น 1,2,3,4.....
- ถ้าจำนวนโครโมโซมที่กำหนดเป็น 1 ให้สร้างชุดโครโมโซมรุ่นใหม่ทั้งหมดที่มีค่าความเหมาะสมน้อยที่สุด
- ถ้าจำนวนโครโมโซมที่กำหนดเป็น 2,3,..... ให้คัดลอกโครโมโซมที่ดีที่สุดในรุ่นเก่า ตามจำนวนที่กำหนดมาเป็นโครโมโซมรุ่นใหม่ แล้วจึงสร้างโครโมโซมรุ่นใหม่ ส่วนที่เหลือต่อไป

2. ครอสโอเวอร์แบบ 2 จุด

การแลกเปลี่ยนส่วนของโครโมโซม พ่อ แม่ บางครั้งหากแลกเปลี่ยนค่าบิทเพียงบางช่วงของโครโมโซมแล้วจะสร้างโครโมโซมที่ดีกว่า เช่น การหาค่าสูงสุดของฟังชั่น $y = x^2$ ของคู่โครโมโซมพ่อแม่ 01110 และ 11001 ซึ่งมีค่าความเหมาะสมเป็น 196 และ 625 หากเปลี่ยนค่าบิทตำแหน่งที่ 3 และ ตำแหน่งที่ 4 เท่านั้นจะทำให้เกิดโครโมโซมลูกคือ 01*00* และ 11*11* มีค่าความเหมาะสมเป็น 64 และ 961 ซึ่งโครโมโซม 11111 เป็นคำตอบที่ดีที่สุดที่ต้องการ ดังนั้นการประยุกต์ SGA โดยพัฒนาตัวดำเนินการครอสโอเวอร์ให้เป็นแบบ 2 จุด จะทำให้ SGA ค้นหาคำตอบได้ดียิ่งขึ้น ดังรูปที่ 2.6



รูปที่ 2.6 ครอสโอเวอร์แบบ 2 จุด

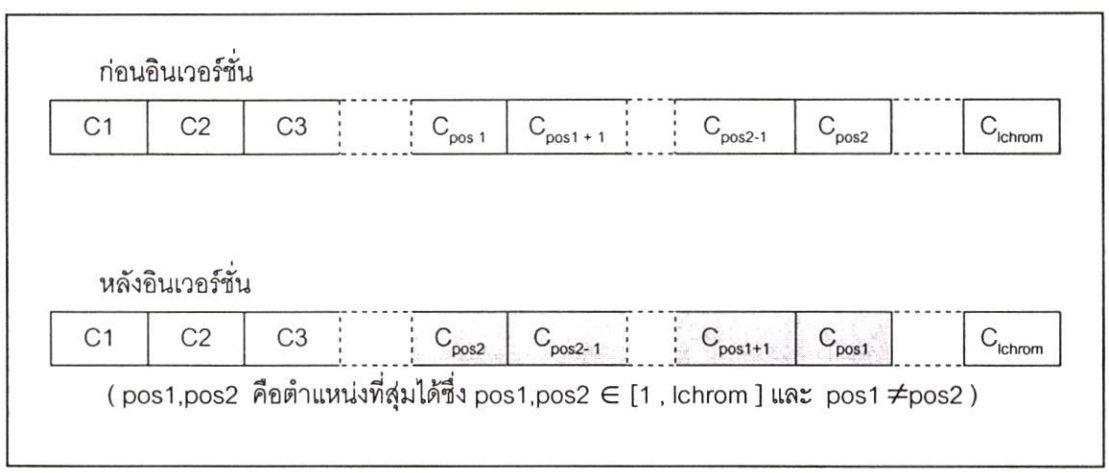
- สุ่มเลือกตำแหน่ง pos1 , pos2 คือตำแหน่งเริ่มต้น และตำแหน่งสุดท้ายที่จะครอสโอเวอร์ตามลำดับซึ่ง pos1 มีค่าอยู่ในช่วง [1,lchrom-1] และ pos2 มีค่าอยู่ในช่วง [1,lchrom] โดยที่ pos1 มีค่าน้อยกว่า pos2
- แลกเปลี่ยนค่าในแต่ละบิทของคู่โครโมโซมพ่อแม่ ตั้งแต่ตำแหน่งที่กำหนด pos1+1 ถึง pos2

3. ไบนารีมิวเตชั่นแบบกำหนดบิท

เนื่องจากการหาคำตอบของ SGA นั้น กระบวนการไบนารีมิวเตชั่นอาจทำให้โครโมโซมที่เปลี่ยนแปลงไปหาคำตอบที่ดีลดลงและทำให้โครโมโซมสูญเสียโครโมโซมที่ดีไป เช่นโครโมโซม 11110 มีค่าความเหมาะสมเป็น 900 หากสุ่มได้บิทตำแหน่งที่ 1 เกิดมิวเตชั่นแล้ว โครโมโซมที่เกิดขึ้นจากการมิวเตชั่นคือ 01110 ทำให้มีค่าความเหมาะสมลดลงเป็น 196 แต่ในบางครั้งข้อดีหรือจุดเด่นของปัญหาจะสามารถนำมาปรับให้เข้ากับการค้นหาคำตอบที่ดีขึ้นได้ สำหรับในการหาค่าสูงสุดของฟังชั่น $y = x^2$ นี้ค่าบิทของโครโมโซมที่เป็น 1 จะทำให้ค่าความเหมาะสมสูงขึ้นเสมอ ดังนั้นหากปรับปรุงไบนารีมิวเตชั่น ให้เป็นแบบกำหนดค่าแน่นอนให้กับบิทเกิดมิวเตชั่น โดยกำหนดให้บิทที่เกิดมิวเตชั่นมีค่าบิทเป็น 1 เสมอ จะช่วยปรับแนวทางการค้นหาคำตอบของ SGA ดีขึ้น

4. อินเวอร์ชั่น (Inversion)

เป็นตัวดำเนินการที่ประยุกต์เพิ่มเติมใน SGA (Simple Genetic Algorithm)โดยจำลองลักษณะของการอินเวอร์ชั่นในทางพันธุศาสตร์ที่เป็นลักษณะของการกลับหัวกลับหางส่วนของยีนส์ภายในโครโมโซม ที่อาจช่วยให้เกิดโครโมโซมที่ดีขึ้นได้ โดยการกลับส่วนค่าบิทภายในช่วงตำแหน่งของโครโมโซมที่สุ่มได้ตามอัตราค่าความน่าจะเป็นของการอินเวอร์ชั่นแต่ละโครโมโซม (Probability of Inversion: Pi) ที่กำหนดดังรูปที่ 2.7



รูปที่ 2.7 อินเวอร์ชั่น

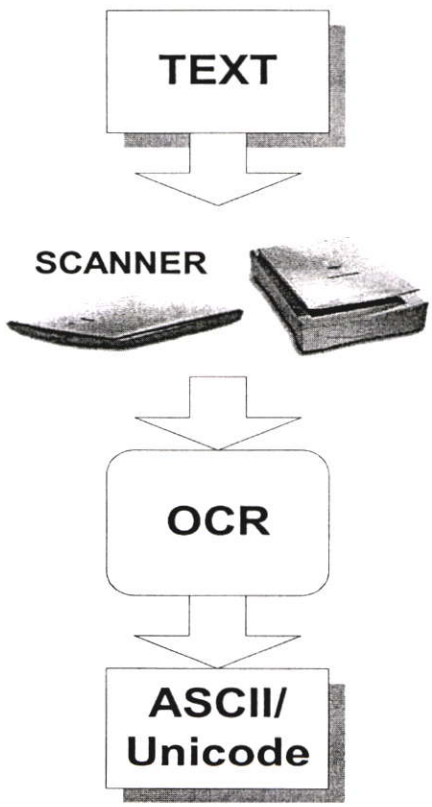
- สุ่มเลือกตำแหน่ง pos1 และ pos2 คือตำแหน่งเริ่มต้นและตำแหน่งสุดท้ายที่จะอินเวอร์ชั่นตามลำดับซึ่ง pos1,pos2 มีค่าอยู่ในช่วง [1,lchrom] โดยที่ pos1 < pos2
- กลับค่าบิทในช่วงของตำแหน่ง pos1 ถึง pos2 ของโครโมโซมโดยสลับค่าบิท pos1 กับ pos2, pos1 +1 กับ pos2-1,pos1+2 กับ pos2-2,.....

2.2 Recognition

ได้มีการค้นคว้างานวิจัยเรื่อง Pattern Recognition กันอย่างกว้างขวางในปัจจุบัน และได้มีการนำมาใช้งานกันอย่างกว้างขวางทั้งทางด้านธุรกิจ(การจดจำตัวอักษร) ทางด้านการแพทย์ (การค้นหาสิ่งผิดปกติ) ทางด้านระบบอัตโนมัติ(หุ่นยนต์) ทางด้านการทหาร ทางด้านการสื่อสาร และอื่นๆ อีกมากมาย

สาขาต่างๆ ที่นำไปใช้งานเช่น การจดจำตัวอักษร การค้นหาเป้า การวิเคราะห์สัญญาณและภาพทางด้านชีวการแพทย์ การส่งระยะไกล การค้นหาภาพหน้าเหมือน และลายพิมพ์นิ้วมือ เศรษฐกิจสังคม การจดจำคำพูดและความเข้าใจ การจดจำชิ้นส่วนเครื่องยนต์และทำการตรวจสอบโดยอัตโนมัติ

การรู้จำตัวอักษรเป็นการแปลงรูปแบบของภาษาจากลักษณะรูปภาพ/เครื่องหมายที่อยู่บนแผ่นกระดาษหรือพื้นผิวใด ๆ เป็น สัญลักษณ์ที่ใช้แสดงถึงตัวอักษรนั้น ๆ เช่นอยู่ในรูปแบบของ Unicode หรือ ASCII ดังแสดงในรูปที่ 2.8



รูปที่ 2.8 การแปลงข้อความบนกระดาษเป็น ASCII

ในการรู้จำภาษาไทยค่อนข้างยุ่งยากกว่าภาษาอื่นเนื่องจากว่าภาษาไทยนั้นไม่มีตัวแบ่งระหว่างคำเหมือนภาษาอังกฤษหรือเครื่องหมายเว้นระหว่างคำ ดังนั้นในภาษาไทยอาจจะจำเป็นต้องมีการตัดแบ่งคำก่อนทำการวิเคราะห์ความถูกต้องของคำ สำหรับการแก้ไขคำผิด มีกรรมวิธีในการแก้ไขอยู่หลายวิธี ขึ้นอยู่กับชนิดของความผิดพลาดและลักษณะการเขียนโดยเฉพาะประโยคในภาษาไทย นั้นจะไม่มีการแบ่งแยกระหว่างคำให้เห็นอย่างชัดเจน จนกว่าจะจบประโยค หรือมีการเว้นวรรค หรือ ขึ้นย่อหน้าใหม่ การตรวจสอบคำสะกดภาษาไทยนั้นจึงจะต้องทำกับตัวอักขระทุกตัวที่มีความคลุมเครือและกับทุกคำที่มีขอบเขตคลุมเครือเท่าที่เป็นไปได้ ดังนั้นการทราบขอบเขตคำในประโยคจึงมีความจำเป็นอย่างยิ่งในการตัดแบ่งคำด้วยลักษณะการเขียนในภาษาไทย เป็นการยากที่จะทำให้ทราบว่าเราควรจะตัดแบ่งคำอย่างไร หรือ มีคำใดที่มีคำผิดอยู่ในประโยค

จากผลงานวิจัยที่ผ่านมาการตรวจสอบตัวสะกดในภาษาไทยจะใช้หลักการทางอักขระวิธีและวิธีการตรวจสอบตัวสะกดโดยใช้หลักการเปรียบเทียบกับพจนานุกรม ซึ่งการตรวจสอบทางอักขระวิธีนี้ จะสามารถตรวจสอบได้รวดเร็ว แต่จะมีความถูกต้องน้อย เพราะอาจจะพบคำที่ผิดแต่เป็นไปตามกฎ ส่วนการตรวจสอบกับพจนานุกรมจะเป็นการนำคำในประโยคมาเปรียบเทียบกับคำในพจนานุกรม ซึ่งมีข้อดีคือ มีความแม่นยำ ถูกต้องสูง แต่ความเร็วในการประมวลผลจะช้ากว่าแบบแรก

จากการจดจำตัวอักษรถ้าหากเอกสารต้นฉบับมีคุณภาพไม่ค่อยดีเท่าทีควร หรือคุณภาพที่ได้จากการสแกนต่ำ อาจจะทำให้การรู้จำของตัวอักษรเกิดความผิดพลาดสูงขึ้น ดังนั้นในงานวิจัยนี้จึงเสนอแนวความคิดในการแก้ไขคำผิดที่ได้จาก OCR เพื่อให้มีประสิทธิภาพที่ดียิ่งขึ้น

2.3 Thai OCR Error Correction

การแก้คำผิดในเอกสารที่เกิดจากความผิดพลาดจาก OCR นั้น ในงานวิจัยนี้ได้ทำการแก้ไขความผิดพลาดในกรณีตัวอักษรผิด เช่น ถารรณรงค์ ซึ่งประโยคที่ถูกคือ การรณรงค์ ในส่วนของงานวิจัยได้แบ่งหัวข้อ ที่เกี่ยวข้องกับงานวิจัยเป็นหัวข้อย่อยดังต่อไปนี้

2.3.1 การอ่านตัวอักขระที่ไม่ชัดเจน

ในการอ่านตัวอักขระที่ไม่ชัดเจนเราสามารถแยกแยะความแตกต่างระหว่างตัวอักษรได้อย่างง่ายดาย ถึงแม้ เอกสารนั้นจะมีคุณภาพค่อนข้างต่ำหรือผิดพลาดบางส่วน เนื่องจากเมื่อเราเจอคำที่อ่านแล้วเราไม่สามารถอ่านได้หรือไม่ชัดเจนเราก็จะนึกถึงตัวอักษรที่ใกล้เคียงกับตัวนั้นๆ และลองสับเปลี่ยนในใจ ถ้าแก้ไขแล้วสามารถอ่านได้ใจความก็น่าจะเป็นคำที่เราคิดไว้ หรือถ้าเปลี่ยนไปแล้วไม่เข้ากับใจความรอบข้างก็จะลองเทียบเคียงกับคำอื่นที่ลักษณะใกล้เคียงกัน ซึ่งจะ

ต้องใช้ฐานข้อมูลเดิมเป็นตัวตัดสินใจในการอ่านอักขระที่ไม่ชัดเจนด้วย โดยการอ่านตัวอักขระที่ไม่ชัดเจนนั้นวิธีในการหาตัวอักขระที่ถูกต้องนั้นมีหลายวิธีซึ่งประกอบไปด้วยวิธีต่างๆ ดังนี้

- ใช้ความรู้เกี่ยวกับการสร้างคำ
- ใช้ตัวอักษรที่ลักษณะใกล้เคียงกัน
- ใช้ความรู้เกี่ยวกับความหมายของคำ
- ใช้ความคุ้นเคยกับคำและประโยค

สำหรับลักษณะของตัวอักษรที่คล้ายกัน เช่น

- ก ภ ถ
- น บ ม ฆ
- ด ต ค ศ
- ท ฑ ห
- ผ พ บ
- ช ซ

จากตัวอย่างข้างบนนี้จะพบว่ามีแบบตัวอักษรหลายแบบซึ่งคล้ายกันมากโดยเฉพาะบางคู่ สำหรับบางแบบของตัวอักษร มีความแตกต่างกันน้อยมาก ดังตัวอย่างดังนี้

- ภาพ --> กาพ
- บรรทัด --> บรรทัค
- การละครั้ง --> ภารละครั้ง
- นอนหลับ --> นอมหลัน
- ทดลอง --> ทคลอง
- นโยบาย --> มโยผาย

ซึ่งตัวอักษรที่ถูกต้องน่าจะอยู่ในผลลัพธ์ของ OCR ซึ่งมีค่าคะแนนรองลงมา เช่น

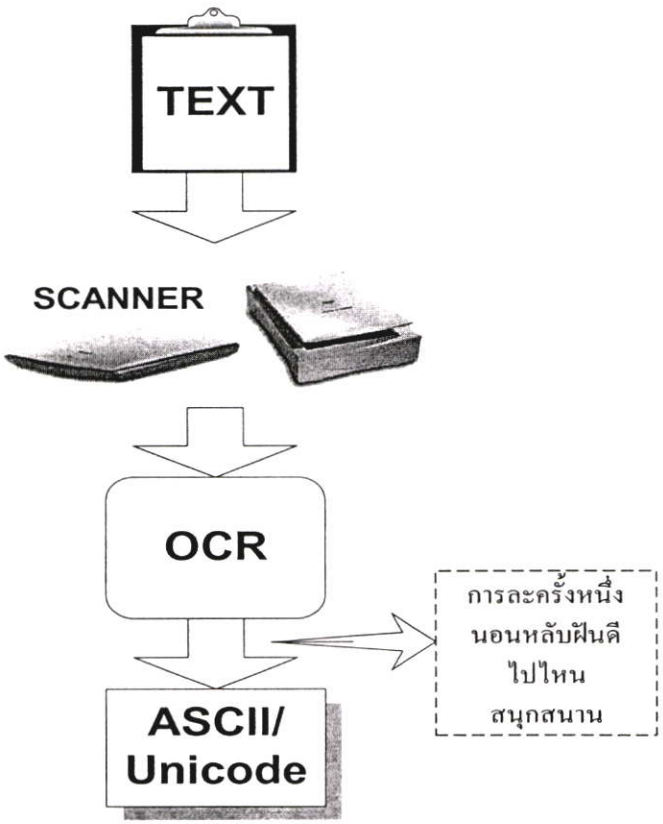
ก 0.70 ภ 0.68 ถ 0.50 ฤ 0.42 ฏ 0.32 ฏ 0.33 .....

า 0 .80 ว 0.75 จ 0.40 ๅ 0.30 ๆ 0.30

พ 0 .76 ผ 0.74 ฟ 0.71 ฝ 0.64 ฬ 0.50

ตามปกติ OCR ทั่วไปจะเลือกเอาตัวที่มีคะแนนสูงที่สุดในแต่ละตัวออกมา เช่นจากตัวอย่างข้างบนจะได้คำว่า กาพ ออกมา ซึ่งจะเป็นคำที่ผิด

ดังนั้นจึงเกิดแนวคิดในการแก้ไขคำที่ผิดพลาด โดยแทนที่จะใช้ตัวอักษรที่วิเคราะห์ได้จาก OCR เพียงตัวเดียวต่อหนึ่งตำแหน่งก็จะรับมาทั้งชุดในรูปแบบของคำหรือประโยคเพื่อใช้แก้ไขความผิดพลาดในขั้นตอนต่อไป



รูปที่ 2.9 การรับตัวอักษรเป็นชุดจาก OCR

การคำนวณเพื่อจัดการกับตัวอักษรที่ทำให้เกิดความกำกวม อยู่บนพื้นฐานของสมมุติฐานของ เนื้อความของสารสนเทศนั้น ว่ามีเพียงใด ถ้าตัวอักษรนั้นมีความเป็นอันหนึ่งอันเดียวกันและทำให้เกิดเป็นคำที่มีความหมายและสอดคล้องกัน สารสนเทศนั้นจะสูงในทางตรงกันข้าม ถ้าหากว่าตัวอักษรนั้นไม่สามารถอ่านได้หรือ อ่านออกมาแล้วไม่สามารถตีความได้หรือไม่มีความสอดคล้องกับคำอื่นใกล้เคียงกัน สารสนเทศนั้นจะ ต่ำ

ยกตัวอย่างเช่นหากมีข้อความ "กต" จะถือว่าสารสนเทศหรือข้อความนี้มีค่า ต่ำ เนื่องจาก กต ไม่มีความหมายหรือสารสนเทศที่สื่อความหมายใดๆ จากตัวอย่างที่ยกมานี้ง่ายในการตรวจสอบเนื่องจากเป็นคำ แต่ถ้าเป็นประโยคยาวกว่าหนึ่งคำปัญหาหนึ่งคือการไม่ทราบขอบเขตที่แน่นอนของคำ เนื่องจากภาษาไทยนั้นเขียนคำติดต่อกันไปโดยไม่มีการเว้นวรรคหรือใช้สัญลักษณ์

พิเศษเพื่อแยกระหว่างคำ ดังนั้นในการตรวจสอบคำสะกดในภาษาไทยจึงจำเป็นต้องผ่านการตัดแบ่งคำ

2.3.2 การตัดแบ่งคำ

สำหรับในภาษาไทยได้มีงานวิจัยในการตัดแบ่งคำโดยสองวิธีหลักคือ

1. ใช้อักษรวิธี Theeramunkong and Usanavasin (2001)
2. ใช้วิธีตรวจสอบคำจากพจนานุกรม Promchan (1998)

การใช้อักษรวิธี เป็นการตรวจสอบการสะกดคำโดยใช้ไวยากรณ์ในการเขียนภาษาไทย ข้อดีคือ ไม่จำเป็นต้องมีการฝึกฝนก่อน สามารถใช้งานได้ทันทีและสามารถทำงานได้รวดเร็ว ส่วนข้อเสียคือ คำที่ตัดออกมาอาจเป็นคำที่สะกดผิด หรือตัดคำออกมาโดยไม่มีความหมาย ใด เช่น คำว่า เผียง และ และ แผร เป็นคำที่เมื่อตรวจสอบด้วยการเขียนนั้นถูกต้องแต่ ไม่เป็นคำที่มีความหมายในภาษาไทย

ส่วนกรรมวิธีการตรวจสอบขอบเขตคำภาษาไทยด้วยการเทียบคำกับพจนานุกรมและประสิทธิภาพของวิธีต่าง ๆ ในงานวิจัย Performance Comparison of Thai Word Separation Algorithm Promchan(1998) คือ

- ตัดคำที่สั้นที่สุด
- ตัดคำที่ยาวที่สุด
- ใช้คำที่ถูกใช้บ่อยที่สุด
- ใช้ back-tracking
- Maximal-matching เลือกใช้คำที่ทำให้เกิดจำนวนคำน้อยที่สุดในประโยค (วรรคภายในบรรทัด)
- ใช้พจนานุกรมพิเศษช่วย สำหรับคำที่ทำให้เกิดความกำกวมซึ่งอาจพิจารณาตัดคำโดยใช้มนุษย์เป็นผู้เลือกการตัดคำ โดยพจนานุกรมสองชุดโดยชุดหนึ่งจะมีคำที่ทำให้เกิดความกำกวมในการตัดประโยค เช่น

วัฒนธรรมไทย = วัฒน ธรรม ไทย = วัฒนธรรม ไทย

รอยกร่าง = รอ ยก ร่าง

มีดอก = มี ดอก

โคลงเรือ = โค ลง เรือ = โคลง เรือ

มาสอบ = มา สอบ

สำหรับงานวิจัยนี้การตัดคำจะใช้ Token passing จะเป็นการเลือกคำที่เป็นไปได้ทั้งหมดซึ่งสามารถทำให้เกิดประโยคที่ต่อเนื่องไปจนจบประโยค โดยคำที่เป็นไปได้เปรียบเทียบกับพจนานุกรมซึ่งจะได้อธิบายในหัวข้อต่อไป

2.3.3 ลักษณะคำที่ผิดสำหรับ OCR

- คำผิดที่ไม่ปรากฏในพจนานุกรม สามารถตรวจสอบได้

เช่น "ถาร เพาะ ปลูก" ซึ่งจะเห็นว่าคำว่า "ถาร" จะไม่มีในพจนานุกรม ซึ่งจะทำให้เราสามารถตรวจสอบได้

เช่น "นฆ นาน" ซึ่งจะเห็นว่าคำว่า "นฆ" จะไม่มีในพจนานุกรม ซึ่งจะทำให้เราสามารถตรวจสอบได้

- คำผิดที่ไม่สามารถตรวจสอบได้ด้วยพจนานุกรม

เช่น "ภาร เพาะ ปลูก" ซึ่งจะเห็นคำว่า "ภาร" เป็นคำที่มีในพจนานุกรม ซึ่งเป็นคำที่ถูกแต่เมื่อมามองในรูปประโยคจะเป็นประโยคที่ผิด

เช่น "นน นาน" ซึ่งจะเห็นว่าคำว่า "นน" จะเป็นคำที่มีในพจนานุกรม ซึ่งเป็นคำที่ถูกแต่เมื่อมามองในรูปประโยคจะเป็นประโยคที่ผิด

- คำที่ถูกต้องแต่ตรวจสอบไม่พบในพจนานุกรมเนื่องจากเป็นการลดรูปของคำ

เช่น "ชีววัสดุ" ซึ่งจะเป็นคำที่ไม่เจอในพจนานุกรม แต่เป็นคำที่ถูกต้องเนื่องจากอาจจะเป็นคำที่ใหม่ที่ยังไม่มีปรากฏในพจนานุกรมหรือเป็นการลดรูปของคำ

- คำที่ไม่ผิดซึ่งเป็นคำที่ยืมมาจากภาษาต่างประเทศ หรือเป็นชื่อเฉพาะ

เช่น "กรดอะมิโนล" ซึ่งเป็นคำที่ไม่เจอในพจนานุกรมเนื่องจากเป็นคำที่ยืมมาจากภาษาต่างประเทศ

2.3.4 การแก้ไขคำผิด

ได้มีงานวิจัยอยู่หลายแนวทางด้วยกันในการแก้ไขคำผิดได้แก่

- การใช้ Token Passing Algorithm ในการแก้คำผิดใน OCR ภาษาไทย Kruatrachue (2001 : 599-602) เป็นการนำเอาตัวอักษรที่ใกล้เคียงกันที่มีค่าความนะเป็นที่ถูกต้องใกล้เคียงกันที่ได้จาก OCR มาเข้าสู่กระบวนการหาคำที่เป็นไปได้โดยเทียบกับ Dictionary เพื่อหาคำที่ถูกต้องจากนั้นจะได้คำที่เป็นไปได้จำนวนหลายคำเมื่อนำมาสร้างอยู่ในรูปของประโยคจะทำให้เกิดประโยคที่เป็นไปได้จำนวนมาก ดังนั้นขั้นตอนสุดท้ายจึงมีการใช้วิธี Prunning เพื่อตัดบางคำออกไปเพื่อลดจำนวนประโยคที่จะทำการค้นหา จากนั้นใช้ Language Model เป็นตัววิเคราะห์ความถูกต้องของประโยค

- การใช้ไวยากรณ์ในการตรวจสอบความถูกต้องและแก้ไขคำผิด Uwe (1998) เป็นการนำเอาไวยากรณ์มาช่วยวิเคราะห์ในการตรวจสอบความถูกต้องและแก้คำผิดโดยรูปแบบประโยคที่ถูกต้องจะต้องถูกต้องตามหลักไวยากรณ์ คือมีประธาน กริยา และกรรมของประธาน หรือคำขยายต่างๆ เป็นตัวช่วยในการตรวจสอบความถูกต้องและแก้คำผิด

- การใช้ระบบผู้เชี่ยวชาญ Kazem (1994 : 270-278) ระบบผู้เชี่ยวชาญในการแก้คำผิดจะเกี่ยวข้องกับการจัดการความรู้ของภาษา และถูกออกแบบให้ช่วยในการตัดสินใจโดยทำการเลือกค่าคะแนนน้ำหนักต่างๆ จากระบบผู้เชี่ยวชาญ โดยใช้หลักการทำงานด้วยระบบปัญญาประดิษฐ์

- การใช้ winnow Golding (1999 : 107-130) การแก้คำผิดโดยวิธีนี้จะทำการรวบรวมคำที่มีรูปร่างคล้ายกัน หรือคำที่มักจะผิดหรือเพี้ยน โดยรวบรวมไว้เป็นกลุ่มเข้าด้วยกันเพื่อทำการแยกแยะรายละเอียดในขั้นตอนต่อไปว่าคำเหล่านี้มีความหมายสัมพันธ์กับคำรอบข้างอย่างไรบ้าง ซึ่งคำที่อยู่รอบๆ เหล่านี้อาจจะรวมถึงชนิดของคำด้วยเช่น คำนาม คำกริยา คำวิเศษ และอาจรวมถึงหมวดหมู่ของคำด้วย

- การใช้ tri-gram และ winnow Mekavin (1998) ในส่วนของการแก้คำผิดโดยวิธีนี้ปัญหามาจากจำนวนคำภายในกลุ่มคำที่เกิดขึ้นนั้นจะมีคำที่เป็นไปได้นั้นจำนวนมาก ซึ่งในการทำงานจริงหากใช้ทุกคำที่เป็นไปได้มาตรวจสอบด้วย Winnow แล้ว ทุกคำที่เป็นไปได้จะทำให้เกิดการทำงานที่ไม่จำเป็น ดังนั้นจึงเอา tri-gram เข้ามาตรวจสอบแล้วเลือกเฉพาะคำ ที่น่าจะเป็นไปได้แล้วทำการตรวจสอบด้วย Winnow ต่อไป

- การใช้ โครงข่ายปราสาทเทียม Neural network John(1994 : 322-333) ในส่วนของการแก้คำผิดโดยวิธีนี้เป็นการนำเอาทฤษฎีโครงข่ายปราสาทเทียมมาช่วยในการแก้ไขคำผิด โดยได้ใช้อินพุตสำหรับโครงข่ายปราสาทเทียมประเภทเพอเซปตรอนหลายชั้น(Multilayer Perceptron) และใช้กระบวนการเรียนรู้แบบแพร่กระจายกลับหลัง(Backpropagetion) เพื่อให้ได้ตัวอักษรที่ถูกต้องที่น่าจะเป็นไปได้สูงสุดโดยจำลองรูปแบบลักษณะของการวิเคราะห์เป็นโครงข่ายใยประสาทที่เชื่อมต่อกันระหว่างเซลล์ประสาทจำนวนมากมาย ทำหน้าที่เป็นจุดศูนย์กลางการควบคุมกิจกรรมต่างๆ โดยเรียนรู้และศึกษาการทำงานของสมองชีวภาพเพื่อกำหนดแนวทางสำหรับการสร้างแบบจำลองขึ้นมา แล้วพยายามสมมติฐานลักษณะการทำงานโดยจำลองเป็นโมเดลคณิตศาสตร์ที่มีลักษณะเดียวกันแล้วดำเนินการคำนวณ

2.4 Token Passing Algorithm

กรรมวิธีการตัดแบ่งคำด้วยโทเกนอัลกอริทึม จะกระทำการตรวจสอบคำสะกดโดยจะนำตัวอักษรของคำเพิ่มเข้ามาในโทเกนทีละตัวแล้วนำไปเปรียบเทียบกับคำในพจนานุกรม เมื่อพบว่าเป็นคำหรือมีโอกาสที่จะเกิดเป็นคำ โทเกนนั้นก็จะยังเก็บเอาไว้ และถ้าโทเกนใดตรวจไม่พบคำหรือไม่มีโอกาสที่จะเกิดเป็นคำแล้วโทเกนนั้นก็จะถูกทิ้งไป ดังตัวอย่างในตารางที่ 2.3

ตารางที่ 2.3 ตัวอย่างเปรียบเทียบจำนวนคำที่ทำการตรวจสอบระหว่างการมีขอบเขตของคำและไม่มีขอบเขตของคำ (ภาษาไทยและภาษาอังกฤษ)

มีเว้นช่องว่างระหว่างคำ (Space between words)	ไม่มีเว้นช่องว่างระหว่างคำ (No word boundaries)
จะ ต้อง Has to	จะต้อง Hasto
2 มีสองคำที่ต้องทำการตรวจสอบนั่นคือ Has , to	คำแรกที่ตรวจสอบ: Ha*, Has, Hast, Hasto คำที่สองที่เป็นไปได้: (เริ่มจาก Ha) st, sto (start from Has) to
	ประโยคที่เป็นไปได้ Has to

มีเว้นช่องว่างระหว่างคำ	ไม่มีเว้นช่องว่างระหว่างคำ
เขียน ติด กัน	เขียนติดกัน
มีสามคำที่ต้องทำการตรวจสอบนั่นคือ เขียน, ติด, กัน	คำแรกที่จะตรวจสอบ เข, เขี, เขีย, เขียน, เขียนต, เขียนติ, เขียนติด,เขียนติดก , เขียนติดกั, เขียนติดกัน
	คำที่สองที่เป็นไปได้ (เริ่มจาก เขียน) ติ, ติด
	คำที่สามที่เป็นไปได้ (เริ่มจาก ติ) ดก (เริ่มจาก ติด) กัน
	ประโยคที่เป็นไปได้ เขียน ติด กัน

* Shows a word or beginning of a word in a dictionary

สำหรับในงานวิจัยนี้, ผลลัพธ์จาก OCR จะเป็นกลุ่มของตัวอักษรออกมา 5 ลำดับตัวอักษรซึ่งมีค่าความน่าจะเป็นของแต่ละตัวสูงกว่า 70% ขึ้นไป ด้วยเหตุนี้กรรมวิธีในการตรวจสอบคำสะกดจึงใช้เวลามากขึ้นเนื่องจากแต่ละตัวอักษร ถูกแทนที่ด้วยชุดของตัวอักษร ที่มีลักษณะใกล้เคียงกัน ดังแสดงให้เห็นในตารางที่ 2.4

ตารางที่ 2.4 อักขระ 5 ตัวที่เป็นไปได้และค่าความน่าจะเป็นจากการรู้จำของแต่ละตัวอักษร (ภาษาไทยและภาษาอังกฤษ)

H (0.87)	a (0.88)	s (0.86)	t (0.85)	o (0.89)
	O (0.77)	5 (0.79)	f (0.79)	a (0.78)
	e (0.75)		l (0.79)	e (0.76)
	d (0.74)		1 (0.74)	d (0.73)
	0 (0.72)			0 (0.73)

เ	ข	อิ	ย	น	ต	อิ	ด
0 .87	0.88	0.86	0.88	0.85	0.89	0.86	0.87
โ	ซ	อื	บ	บ	ด	อี	ต
0.78	0.77	0.79	0.78	0.79	0.78	0.78	0.79
ไ	ซ	อื	น	ม	ค	อื	ค
0.77	0.75	0.77	0.76	0.79	0.76	0.75	0.78
ใ	บ	อี	ม	ข	ศ	อื	ศ
0.75	0.74	0.75	0.73	0.74	0.73	0.73	0.75
ร	ป	อ์	ษ	ษ	ศ	อ์	ศ
0.74	0.72	0.72	0.70	0.70	0.73	0.72	0.73

จากตารางที่ 2.4 แสดงผลลัพธ์ที่ได้จาก OCR อ่านประโยค "เขียนติด" ในที่นี้สระ อิ ในตำแหน่งที่ 2 นั้นมีผลลัพธ์จากการรู้จำได้คะแนนสูงเป็นอันดับแรก และถ้าคิดเฉพาะอักษรซึ่งมีคะแนนสูงสุดแล้วจะได้เป็นประโยคว่า "เขยนติด" ซึ่งเป็นกรณีของอักษรผิดไม่ถูกต้อง จากประโยคที่ถูกต้องแล้วน่าจะเป็นสระ อี ดั้งนั้นจึงได้มีการนำเอากระบวนการ โทเกนพาสซิ่ง อัลกอริทึมมาใช้ในการหาคำที่น่าจะเป็นไปได้ที่ถูกต้องเพื่อที่จะได้ผลลัพธ์ที่ถูกต้องออกมา

อัลกอริทึมของโทเกนพาสซิ่ง จะเป็นการเอาตัวอักขระที่เป็นไปได้ที่มีค่าความถูกต้องที่มากกว่า 0.70 ขึ้นไป หรือ สูงสุด 5 ตัวอักษร (เนื่องจากว่าถ้าเอามากเกินไปจะทำให้เกิดคำที่เป็นไปได้จำนวนมากทำให้ยากในการหาประโยคที่ถูกต้อง) จากนั้นโทคเกนจะถูกสร้างและส่งผ่านไปยังอักขระแต่ละตัวดังแสดงในตารางที่ 2.5 และตารางที่ 2.6 โทเกนที่เก็บไว้นั้นจะประกอบด้วยคำ (หรือส่วนเริ่มต้นของคำ)ที่พบในพจนานุกรม สำหรับโทเกนที่ไม่มีคำปรากฏในพจนานุกรมจะถูกทิ้งไป และเมื่อโทเกนนั้นได้ทำการตรวจสอบคำสะกดและค้นหาคำในพจนานุกรมแล้วอาจสร้างโทเกนขึ้นใหม่ได้ โดยอันหนึ่งเป็นของโทเกนซึ่งสิ้นสุดลงด้วยขอบเขตของคำ และอีกอันหนึ่งเป็นโทเกนที่คำยังไม่สิ้นสุดซึ่งอาจจะยังสามารถต่อไปเป็นคำอื่นซึ่งมีอยู่ในพจนานุกรมได้อีก

ตารางที่ 2.5 ตัวอย่างของโทคเกนพาสซิ่ง (ภาษาอังกฤษ)

1	2	3	4	5 (char Position)
H H	a Ha*	s s, Has, Hos,Hes	t t, st, Hast, Host, Hest	o to, lo
	o Ho	5	f ff,sf	a a, la, ta
	e He		1 ll, Hasl	e Haste, fe
	d		1	d
	0			0
	Ha,ho,he	Has	Hast, Host,st	to,lo,a,la,ta,fe ,Haste
1	3	4	9	7(num. of token)

ตารางที่ 2.6 ตัวอย่างของโทคเกนพาสซิ่ง (ภาษาไทย)

0	1	2	3	4	5	6	7
เ เ	ข เข, เข_, โข_, ไข_,	อิ เขิ, เปิ	ย เขีย, เชีย, เซีย_,เซีย, เปีย_, เบีย	น เขียน_, เซียน _, เบียน_	ด เขียน_ด, เขียน_ต, เบียน_ต, เขียบ_ต ,เขียบ_ต ,เขียม_ต,	อิ เขียน_ติ,เขียน_ติ, เบียน_ติ,เขียบ_ติ, เขียบ_ติ,เขียม_ติ, เขียน_ติ_,เขียน_ติ _,เบียน_ติ_,เขียบ_ ติ_,เขียบ_ติ_,เขียม _ติ_,เขียน_คิ,เขียน _คิ,เบียน_คิ,เขียบ_ คิ,เขียบ_คิ,เขียม_คิ	ด เขียน_ติด_,เขียน_ ติด_,เบียน_ติด_, เขียบ_ติด_,เขียบ_ ติด_,เขียม_ติด_, เขียน_ตีด_,เขียน_ ตีด_,เบียน_ตีด_, เขียบ_ตีด_,เขียบ_ ตีด_,เขียม_ตีด_, เขียน_คิด_,เขียน_ คิด_,เบียน_คิด_, เขียบ_คิด_,เขียบ_ คิด_,เขียม_คิด_
โ โ	ซ เซ, ไซ_	อึ 	บ เขิบ_, เปิบ_	บ เขียบ_, เซียบ_	ด เขียน_ด, เขียน_ต, เบียน_ต, เขียบ_ต, ,เขียบ_ต, ,เขียม_ต	อี เขียน_ตี,เขียน_ตี, เบียน_ตี,เขียบ_ตี, เขียบ_ตี,เขียม_ตี, เขียน_ตี_,เขียน_ตี _,เบียน_ตี_,เขียบ_ ตี_,เขียบ_ตี_,เขียม _ตี_,เขียน_ตี,เขียน _ตี,เบียน_ตี,เขียบ_ ตี,เขียบ_ตี,เขียม_ตี	ต เขียน_ตีด_,เขียน_ ตีด_,เบียน_ตีด_, เขียบ_ตีด_,เขียบ_ ตีด_,เขียม_ตีด_
ไ ไ	ซ เซ, เซ_, โซ, ไซ	อึ 	น เขิน_	ม เขียม_	ค นค_, เขียน_ค, เขียน_ค, เบียน_ค, เขียบ_ค, ,เขียบ_ค, ,เขียม_ค	อึ เขียน_ตี_,เขียน_ตี _,เบียน_ตี_,เขียบ_ ตี_,เขียบ_ตี_,เขียม _ตี_,เขียน_ตี,เขียน _ตี,เบียน_ตี,เขียบ_ ตี,เขียบ_ตี,เขียม_ตี	ค
ใ ใ	บ โบ_, ใบ_, รบ_ ,เบ	อี เขี, เซี, เปี , เซี, เบี	ม 	ข 	ศ 	อี เขียน_ตี, เขียน_ตี, เบียน_ตี, เขียบ_ตี, ,เขียบ_ตี, ,เขียม_ตี	ค
ร ร	ป โป_, ไป_, เป	อ์ 	ษ 	ษ 	ศ 	อ์ 	ศ

เมื่อเข้าสู่กระบวนการโทเกนพาสซิ่งอัลกอริทึม แล้วจะได้คำที่เป็นไปได้จำนวนหลายคำประกอบกันเป็น Word Graph ดังรูป 2.10 ซึ่งจะไปกอบไปด้วยตำแหน่งของตัวอักษรและค่าคะแนนของแต่ละคำนั้นซึ่งได้มาจากการคำนวณดังสมการ โดยค่าคะแนนของแต่ละคำนี้จะเป็นค่าหนึ่งของ Fitness Function ของกระบวนการเจเนติกอัลกอริทึม

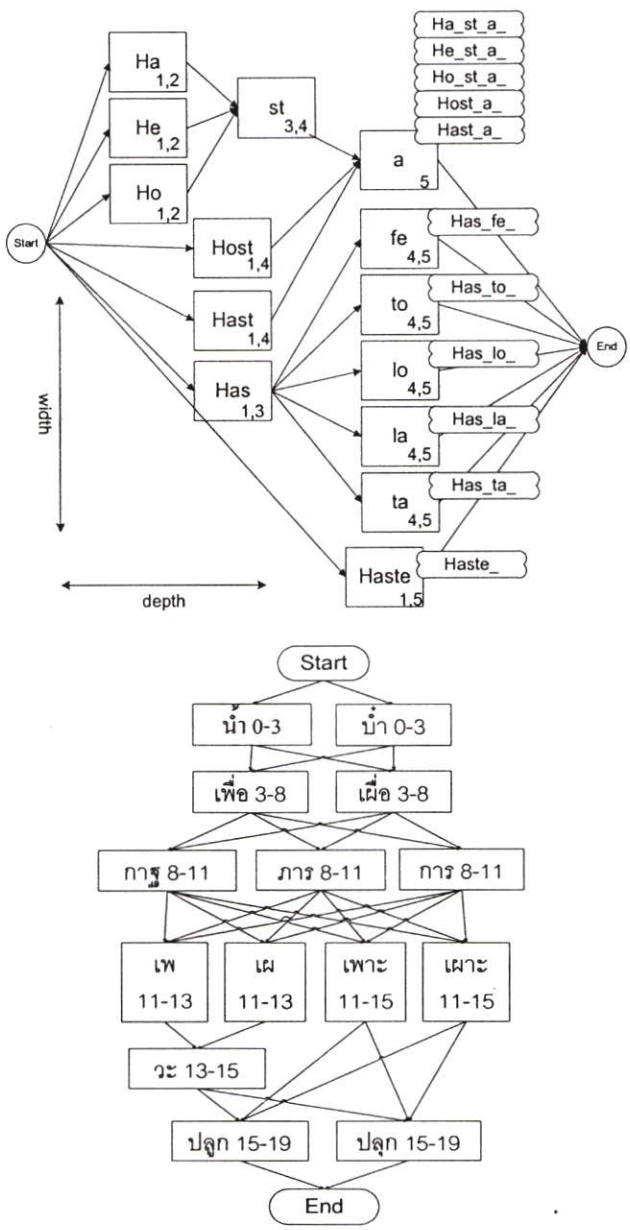
$$P(W \mid S) = n\prod_{i}^{n} P(c_i)$$

โดยที่

c = อักขระ

n = จำนวนตัวอักษรของคำ

S = สายอักขระ $c_1c_2...c_n$ โดยที่ c_n คืออักขระตัวสุดท้ายของคำ



รูปที่ 2.10 Word Graph ที่ได้จากโทเกนพาสซิ่งอัลกอริทึม

2.5 Language Model

จากผลในการสร้าง word graph จะพบว่ามีความเป็นไปได้หลายแบบในการสร้างประโยคด้วยกันและจากการนำค่าความน่าจะเป็นที่ได้จากผลลัพธ์ของ OCR มาใช้วิเคราะห์ความถูกต้องเพียงอย่างเดียวยังไม่เพียงพอดังนั้นการเพิ่มประสิทธิภาพผลลัพธ์ของ OCR ให้ดียิ่งขึ้นจึงต้องมีการนำเอา Language Model มาใช้เพิ่มประสิทธิภาพ เช่น จากผลลัพธ์ที่ได้

ตารางที่ 2.7 แสดงประโยคที่ได้จาก words graph และคะแนนรวมของแต่ละประโยค (คะแนนรวมที่คิดจากค่าคะแนนของแต่ละตัวอักษรที่ได้จาก OCR)

No.	ประโยค	คะแนนรวม
1	น้ำ เพื่อ ภาร เพาะ ปลูก	0.88
2	น้ำ เพื่อ การ เพาะ ปลูก	0.84
3	น้ำ เผื่อ ภาร เพาะ ปลูก	0.86
4	น้ำ เพื่อ ภาร เผาะ ปลูก	0.81
5	น้ำ เผื่อ การ เพาะ ปลูก	0.80
6	น้ำ เพื่อ ภาร เพ วะ ปลูก	0.75
7	น้ำ เผื่อ ภาร เผาะ ปลูก	0.74
8	น้ำ เพื่อ การ เผาะ ปลูก	0.70
9	น้ำ เพื่อ การ เพ วะ ปลูก	0.70
10	น้ำ เผื่อ ภาร เพ วะ ปลูก	0.66
11	น้ำ เพื่อ ภาร เผ วะ ปลูก	0.63
12	น้ำ เผื่อ การ เผาะ ปลูก	0.63
13	น้ำ เพื่อ การ เผ วะ ปลูก	0.62
14	น้ำ เผื่อ ภาร เผ วะ ปลูก	0.62
15	น้ำ เผื่อ การ เพ วะ ปลูก	0.60

จากตารางที่ 2.7 จะเห็นว่าประโยคที่ถูกต้องคือประโยคที่สองแต่คะแนนรวมที่คิดจากค่าคะแนนของแต่ละตัวอักษรที่ได้จาก OCR นั้นยังต่ำกว่าประโยคที่หนึ่ง ซึ่งเราสามารถเพิ่มประสิทธิภาพของความถูกต้องได้โดยใช้แบบจำลองทางภาษาเข้ามาช่วย ซึ่งจะทำให้ได้ค่าความถูกต้องของประโยคสูงมากยิ่งขึ้นดังผลลัพธ์ตารางที่ 2.8

ตารางที่ 2.8 แสดงประโยคที่ได้จาก words graph และคะแนนจาก Language Model

No.	ประโยค	Perplexity
1	น้ำ เพื่อ ภาร เพาะ ปลูก	79.3
2	น้ำ เพื่อ การ เพาะ ปลูก	11.04
3	น้ำ เผื่อ ภาร เพาะ ปลูก	79.3
4	น้ำ เพื่อ ภาร เผาะ ปลูก	299.26
5	น้ำ เผื่อ การ เพาะ ปลูก	46.71
6	น้ำ เพื่อ ภาร เพ วะ ปลูก	432.37
7	น้ำ เผื่อ ภาร เผาะ ปลูก	761.3
8	น้ำ เพื่อ การ เผาะ ปลูก	88.67
9	น้ำ เพื่อ การ เพ วะ ปลูก	156.9
10	น้ำ เผื่อ ภาร เพ วะ ปลูก	941.41
11	น้ำ เพื่อ ภาร เผ วะ ปลูก	432.37
12	น้ำ เผื่อ การ เผาะ ปลูก	374.96
13	น้ำ เพื่อ การ เผ วะ ปลูก	156.9
14	น้ำ เผื่อ ภาร เผ วะ ปลูก	941.41
15	น้ำ เผื่อ การ เพ วะ ปลูก	521.75

จากตารางที่ 2.8 จะเห็นว่าค่า Perplexity ที่ได้จาก Language Model ยิ่งมีค่าน้อยเท่าไหร่แสดงว่าประโยคนั้นมีโอกาสถูกต้องมากที่สุด ซึ่งจากตารางจะเห็นว่าประโยคที่สองมีค่า Perplexity ต่ำที่สุด ซึ่งก็จะเห็นว่าเป็นประโยคที่ถูกต้อง ดังนั้นเพื่อประสิทธิภาพของการตรวจแก้คำผิดใน OCR ภาษาไทย ในงานวิจัยจึงเป็นการเอาค่าที่ได้จาก Language Model มาเป็นส่วนหนึ่งของ Fitness Function ในเจเนติกอัลกอริทึม ซึ่งจะกล่าวในส่วนต่อไป

การสร้างแบบจำลองภาษา เป็นการพยามที่จะจับเอาลักษณะเฉพาะทางภาษาธรรมชาติ โดยทำให้อยู่ในรูปแบบซึ่งสามารถนำ มาสร้างเป็นแบบจำลองได้ ภาษาธรรมชาติเป็นสิ่งที่มีความซับซ้อนโดยลักษณะของตัวมันอยู่ตามธรรมชาติ ซึ่งมันพัฒนาไปอย่างต่อเนื่องผ่านต่อไปจากรุ่นหนึ่งไปยังอีกรุ่นหนึ่งมาเป็นช่วงเวลานาน โดย Language Model จะเป็นตัวบอกว่า ประโยคนั้นจัดเรียงเป็นอย่างไร เป็นไปได้ถูกต้องมากน้อยเพียงใด

2.5.1 สาเหตุในการใช้กรรมวิธีทางสถิติ

แม้ว่าในส่วนของงานด้าน Natural Language Processing จะเป็นการประยุกต์ใช้ rule-based เป็นหลัก แต่ในด้าน Language Model rule-based นั้น เป็นการยึดติดอยู่กับลักษณะเฉพาะทางกฎเกณฑ์มากเกินไปจนขาดความยืดหยุ่น อาจจะต้องใช้กฎ, ความรู้ (knowledge) เป็นจำนวนมากในการเพิ่มความถูกต้อง

กรรมวิธีหนึ่งซึ่งใช้ในการทำ Language Model ซึ่งมีความแม่นยำและมีความยืดหยุ่นสูง คือการนำ กรรมวิธีทางสถิติเข้ามาใช้โดยในการสร้างแบบจำลองจะใช้ค่าสถิติโดยใช้การฝึกฝน (training) corpora ของ text ขนาดใหญ่

2.5.2 แบบจำลองทางสถิติและแบบจำลอง Knowledge-based

แบบจำลองทางสถิติมีข้อดีกว่าแบบจำลอง Knowledge-based คือค่าความน่าจะเป็นที่ได้จากแบบจำลอง มีประโยชน์มากกว่าคำตอบเพียง "ใช่"/"ไม่ใช่" ซึ่งค่าความน่าจะเป็นที่ได้นำไปใช้ประกอบกับเทคนิคอื่น ๆ ต่อไปได้มีประสิทธิภาพกว่า การสร้างแบบจำลองทางสถิติสามารถพัฒนาและทำการฝึกฝนได้ด้วยโปรแกรมทางคอมพิวเตอร์ซึ่งสามารถทำได้รวดเร็วกว่าแบบจำลองทาง knowledge โดยเฉพาะเมื่อต้องการสร้างแบบจำลองสำหรับเรื่องในหัวข้อใหม่ ในทางปฏิบัติ, ฐานความรู้ส่วนมาก จะใช้เวลาในการคำนวณในขณะทำงานมากกว่า แบบจำลองทางสถิติ

ส่วนข้อเสียของแบบจำลองทางสถิติ เนื่องจากแบบจำลองทางสถิติไม่มีการตรวจสอบความหมายของข้อความ ทำให้บางประโยคซึ่งอ่านดูไม่ถูกต้องตามสามัญสำนึก แต่อาจจะถูกลงความเห็นว่ายอมรับได้ จากโปรแกรม เนื่องจากความสัมพันธ์ของคำกลุ่มนั้นมีความความน่าจะเป็นสูง แบบจำลองทางสถิติต้องการจำนวนของข้อมูลมาทำการฝึกขนาดใหญ่ และการส่งผ่านแบบจำลองระหว่างเรื่องต่างหัวข้อหรือต่างภาษากันไม่สามารถทำได้ เสมอไป

2.5.3 เอนโทรปี (Entropy)

ทฤษฎีสารสนเทศข้อมูลเป็นการใช้คณิตศาสตร์อย่างเป็นระบบและมีแบบแผน และในส่วนของข้อความบนหนังสือก็เป็นรูปแบบหนึ่งของการสื่อสาร ดังนั้นทฤษฎีสารสนเทศจึงมีส่วนเกี่ยวข้องกับงานวิจัย

จากทฤษฎี สารสนเทศที่ออกมา X_i ขึ้นอยู่กับความน่าจะเป็น ถ้าความน่าจะเป็น $P(x_i)$ มีค่าน้อย, เราจะได้ระดับดีกรีของสารสนเทศที่สูง, เนื่องจากผลลัพธ์ที่ได้ออกมามีโอกาส เกิดขึ้นน้อย และถ้าความน่าจะเป็น $P(x_i)$ มีค่าสูง ข้อมูลสารสนเทศที่รับมานั้นจะมีขนาดเล็ก, เนื่องจากผลที่ได้ออกมานั้นสามารถประเมินได้ว่าดีมาก ดังนั้น เราสามารถกำหนดค่าสารสนเทศได้ดังนี้

$$I(x_i) = \log \frac{1}{P(x_i)}$$

ใช้ค่า logarithm เนื่องมาจาก มีข้อมูลของสองเหตุการณ์ที่เป็นอิสระต่อกัน (เป็นการ joint probability ซึ่งในการคำนวณจะเป็นผลคูณ) สามารถนำมาบวกกันได้ และเมื่อใช้ ค่าลอการิทึมฐานสอง หน่วยของ information ก็จะเรียกว่าบิท (bit)

$$H(X) = E[I(X)] = \sum_s P(x_i) \log_2 \frac{1}{P(x_i)} = E[-\log_2 P(x)]$$

entropy H(X) คือจำนวนเฉลี่ยของ information ซึ่งต้องการในการกำหนด what kind of symbol ที่เกิดขึ้น แบบจำลองที่ดีจะให้ค่า entropy ที่ต่ำเมื่อนำไปทดสอบ, หรือกล่าวได้อีกอย่างหนึ่งว่าข้อความที่นำมาทดสอบนั้นมีความเป็นไปได้สูง เมื่อมีค่า entropy ต่ำ และค่า entropy นี้ยังเป็นค่าเฉลี่ยของความไม่แน่นอนของ symbol สมมุติว่า แซมเปิลสเปส S คือจำนวนตัวอักษรขนาด $\|S\| = N$ ค่า entropy H(X) จะเข้าสู่ค่าสูงสุดเมื่อ p.f. has a uniform distributaion เช่น

$$P(x_i) = P(x_j) = \frac{1}{N} \qquad \text{สำหรับ ทุก i และ j}$$

จากสมการนี้สามารถแปลความหมายของ uncertainty ซึ่งมีค่าสูงที่สุดเมื่อค่าความน่าจะเป็นของเหตุการณ์ทั้งหมดซึ่งสามารถเกิดขึ้นได้เท่า ๆ กัน ไม่มีเหตุการณ์ใดมีค่าความน่าจะเป็นสูงกว่าเหตุการณ์อื่น ๆ และมันยังสามารถพิสูจน์ได้ว่าค่า entropy H(X) ไม่เป็นลบและจะมีค่าเป็นศูนย์ ถ้าเพียงค่าความน่าจะเป็น มีค่าเป็น หนึ่ง

$$H(X) \geq 0$$

branching factor ของภาษาเป็นการวัดระดับความยากลำบากของระบบภาษาที่ โดยมีความสัมพันธ์กับขนาดของจำนวนคำที่ใช้เป็นคำศัพท์ที่จำเป็นต้องใช้ในการแยกแยะประโยคที่ได้มา

จากคำนิยามของ entropy ที่กล่าวมาเราสามารถกำหนดค่า branching factor ได้โดยคำนวณจาก

$$PP(X) = 2^{H(X)}$$

เราเรียก PP(X) ว่า perplexity ของ X , และค่า perplexity นี้ยังเป็นตัวที่เรามักใช้วัดคุณภาพของ Language Model โดยทั่วไปอีกด้วย (Bahl et al.,1983) ซึ่ง model ที่ให้ค่า perplexity ออกมาต่ำ model that is in some sense "good" และค่า perplexity เองก็ยังเป็นส่วนกลับค่าเฉลี่ยทางเลขคณิตของความน่าจะเป็นของแต่ละคำในประโยคอีกด้วย

2.5.4 แบบจำลองภาษาทางสถิติ N-gram

ในแบบจำลองภาษาทางสถิติ N-gram เราสามารถคำนวณค่าความน่าจะเป็นของ P(W) ของคำ W โดยแสดงให้เห็นว่า W มีความถี่เท่าไรในการเกิดคำ W ในฐานข้อมูลนั้น เนื่องจากเรามีประโยคและคำเป็นจำนวนมากการคำนวณค่าความน่าจะเป็นจากคำ เพียงคำเดียว ดังเช่น P(นอน) จึงไม่เหมาะสมเท่าไหร่ ดังนั้นเราจึงคำนวณจาก P (มอน หลับ ฝัน), ซึ่งไม่น่าจะมีประโยคที่เขียนเช่นนี้ แต่ถ้าเป็นคำว่า P(นอน หลับ ฝัน) ก็จะมีค่าความน่าจะเป็นของคำที่มากขึ้น เราสามารถพูดได้ง่ายๆ คือแบบจำลองภาษาทางสถิติ N-gram นั้นจะเป็นการคำนวณค่าความน่าจะเป็นของคำโดยดูจากความถี่ที่เกิดคำนั้นจากฐานข้อมูล
P(W) สามารถเขียนใหม่ได้เป็น

$$\begin{aligned} P(W) &= P(w_1, w_2, ..., w_n) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_n \mid w_1, w_2, ..., w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i \mid w_1, w_2, ..., w_{i-1}) \end{aligned}$$

ซึ่ง $P(w_i \mid w_1, w_2, ..., w_{i-1})$ คือความน่าจะเป็นของ w_i ที่สืบเนื่องมาจากลำดับของคำ $w_1, w_2, ..., w_{i-1}$ ซึ่งเกิดขึ้นก่อนหน้า w_i ซึ่งจากสมการข้างบนนี้ w_i จะขึ้นอยู่กับคำทั้งหมดที่อยู่ก่อนหน้า w_i สำหรับคำศัพท์ขนาด v จะมี v^{i-1} history , ดังนั้น $P(w_i \mid w_1, w_2, ..., w_{i-1})$ จึงมีขนาด v^i ค่าที่นำมาใช้ในการคำนวณ ซึ่งในความเป็นจริงแล้วความน่าจะเป็นของ $P(w_i \mid w_1, w_2, ..., w_{i-1})$ ไม่สามารถคำนวณได้จากทุกคำก่อนหน้า w_i , และ $w_1, w_2, ..., w_{i-1}$ ที่เป็นประโยคต่อเนื่องยาว ๆ นั้นมีโอกาสเกิดขึ้นน้อยมาก หรือแทบจะไม่พบเลย ในทางปฏิบัติเราจึงแบ่งแยกกลุ่มคำออกเป็นช่วงละเท่า ๆ กันโดยกลุ่มคำสามารถแบ่งได้โดยใช้คำก่อนหน้า

$w_1, w_2, ..., w_{i-1}$ เป็นจำนวนช่วงเท่า ๆ กัน ซึ่งจะได้ว่า $P(w_i \mid w_1, w_2, ..., w_{i-1})$ หาจากกลุ่มคำโดยแบ่งเป็นส่วนย่อย ๆ จำนวน i คำ เช่นถ้าเราใช้การแบ่งเป็นกลุ่มละสามคำ (tri - gram) เราจะหาความน่าจะเป็นของคำจาก $P(w_i \mid w_{i-2}, w_{i-1})$

ซึ่งในทางปฏิบัติเราสามารถหาค่าความน่าจะเป็นได้โดยการนับจำนวนเหตุการณ์ ที่เราสนใจใน corpus, ให้ $C(w_{i-2}w_{i-1}w_i)$ เป็นการนับจำนวน $w_{i-2}w_{i-1}w_i$ ที่เกิดขึ้นใน corpus ที่นำมาฝึก เช่นเดียวกันกับ $C(w_{i-2}w_{i-1})$ ดังนั้นจึงหาได้ว่า

$$P(w_i \mid w_{i-2}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}$$

แต่การหาค่าความน่าจะเป็นด้วยวิธีการนี้ จะพบว่าจะเกิดปัญหาถ้าเกิด ไม่พบลำดับของคำบางคำในการฝึกจะทำให้หาค่าความน่าจะเป็นออกมาเท่ากับ 0

2.5.5 Smoothing

ในการสร้าง N-gram Language Model เงื่อนไขที่สำคัญอย่างหนึ่งก็คือความจำกัดของข้อมูลที่นำมาใช้ ฝึกฝน นั่นคือปัญหาในกรณีที่เราไม่มีข้อมูลของคำบางคำที่ไม่ค่อยถูกใช้ เราสามารถประมาณค่า maximum likelihood สำหรับค่าความน่าจะเป็นของเหตุการณ์ ε ซึ่งเกิดขึ้นเป็นจำนวน $C(\varepsilon)$ จากทั้งหมด R เป็น

$$P(\varepsilon) = \frac{C(\varepsilon)}{R}$$

และด้วยข้อมูลที่นำมาฝึกซึ่งมีอยู่จำกัด, ทำให้ค่าความน่าจะเป็นที่ได้จากสมการข้างบนนี้ สูงเกินความเป็นจริงสำหรับ events ที่ observed และ ต่ำเกินไปสำหรับเหตุการณ์ที่ unobserved สำหรับ N-gram Language Model เราไม่สามารถหลีกเลี่ยงปัญหาที่จะพบคำศัพท์ที่ไม่พบในการฝึกได้ ยกตัวอย่างเช่นถ้ามีคำศัพท์ 32000 คำ สำหรับ tri-gram จะมี tri-gram ที่เป็นไปได้ทั้งหมดเท่ากับ 3.2768e+13 tri-gram ถ้า corpus ที่นำมาฝึก มีขนาด 100 ล้านคำ ซึ่งเป็นเพียง ขนาด 0.00030517578125% ของ tri-gram ที่เกิดขึ้นได้เท่านั้นเอง และสำหรับกรณีที่ไม่พบคำศัพท์ที่เกิดขึ้น จากสมการข้างบน จะได้ความน่าจะเป็นเท่ากับศูนย์ ซึ่งไม่น่าจะเกิดขึ้น

ดังนั้นจึงมีการนำเอาเทคนิคที่เรียกว่าการทำ smoothing เพื่อไม่ให้เกิดการ bias ค่า maximum likelihood และเพื่อให้แน่ใจว่าไม่มีค่าความน่าจะเป็นเท่ากับศูนย์ เทคนิค smoothing ได้ถูกนำมาใช้ในการแก้ไข กรณีซึ่งเกิดค่าความน่าจะเป็นเท่ากับศูนย์ สำหรับในกรณีซึ่ง พบคำ ซึ่งไม่ปรากฏใน language model คำว่า smoothing มาจากกระบวนการของเทคนิค smoothing ซึ่ง

ทำให้เกิดการกระจายค่าของความน่าจะเป็นมีความราบเรียบมากขึ้น โดยปรับค่าความน่าจะเป็นที่ต่ำเช่น ศูนย์ให้มีค่าสูงขึ้น ปรับค่าความน่าจะเป็นที่สูงให้ต่ำลงเป็นต้น นอกจากการนำเทคนิค smoothing มาใช้จะช่วยป้องกันในการเกิดค่าความน่าจะเป็น เท่ากับศูนย์แล้ว เทคนิคนี้ยังช่วยให้โมเดลมีความแม่นยำเพิ่มขึ้น

จากตัวอย่าง เพื่อง่ายในการทำความเข้าใจ จะยกสมการ smoothing อย่างง่ายสำหรับกรณี bi-gram Lidstone(1920), Johnson(1932),และ Jeffreys(1948)

$$p(w_i \mid w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}\left[1 + c(w_{i-1}w_i)\right]} = \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

V = จำนวนคำศัพท์ทั้งหมด

โดยที่คำใด ๆ ที่ไม่ปรากฏในพจนานุกรมที่ใช้สำหรับ แบบจำลองจะถูกมองเป็น unknown word แต่ค่าที่ได้จากสมการข้างบนนี้ก็ยังให้ความแม่นยำของแบบจำลองได้ไม่ดีขึ้นมาก นักจึงได้มีการพัฒนาต่อไปเป็น

$$P_{smooth}(w_i \mid w_{i-n+1}...w_{i-1}) = \begin{cases} \alpha(w_i \mid w_{i-n+1}...w_{i-1}) & if\ C(w_{i-n+1}...w_i) > 0 \\ \gamma(w_{i-n+1}...w_{i-1}) P_{smooth}(w_i \mid w_{i-n+2}...w_{i-1}) & if\ C(w_{i-n+1}...w_i) = 0 \end{cases}$$

จากสมการ ถ้า n-gram สามารถนับจาก corpus ที่ฝึกได้โดยไม่เป็นศูนย์ เราจะใช้สมการ $\alpha(w_i \mid w_{i-n+2}...w_{i-1})$ นอกนั้นเราจะทำการ backoff ไปยังลำดับของ n-gram ที่ต่ำกว่า โดยใช้สมการ $P_{smooth}(w_i \mid w_{i-n+2}...w_{i-1})$ โดยที่เงื่อนไขของค่า scaling factor $\gamma(w_{i-n+1}...w_{i-1})$ จะเป็นตัวที่ทำให้ผลรวมของการกระจายทั้งหมดมีค่าเป็นหนึ่ง เราเรียกอัลกอริทึมในลักษณะ แบบนี้ว่าแบบจำลอง backoff

อัลกอริทึมอื่นสำหรับการทำ smoothing ได้แก่ linear interpolation สำหรับอันดับ n-gram ที่อันดับสูงกว่าหรือต่ำกว่า

$$P_{smooth}(w_i \mid w_{i-n+1}...w_{i-1}) = \lambda P_{ML}(w_i \mid w_{i-n+1}...w_{i-1}) + (1-\lambda) P_{smooth}(w_i \mid w_{i-1+2}...w_{i-1})$$

โดยที่ λ คือค่า weight ของการ interpolation ซึ่งขึ้นอยู่กับ $w_{i-n+1}...w_{i-1}$ ดังนั้นเราจึงเรียกแบบจำลองแบบนี้ว่าแบบจำลอง interpolated

สิ่งที่แตกต่างกันระหว่างแบบจำลอง backoff และ interpolated คือสำหรับค่าความน่าจะเป็นของ n-grams ซึ่งสามารถพบได้ใน corpus (nonzero counts) แบบจำลอง interpolated จะใช้ข้อมูลจากการกระจายของ ลำดับ n-grams ที่ต่ำกว่ามาร่วมในการคำนวณด้วยส่วน backoff นั้นจะไม่ใช้

2.5.6 **สัมประสิทธิ์การลดทอน** (discounting)

การลดทอนค่า เป็นส่วนที่เพิ่มเติมในหลักการของการแก้ไข การ bias ของเหตุการณ์ที่สังเกต ของการประมาณค่าโดยใช้ maximum likelihood probability estimates เหตุการณ์ที่นับได้จะถูกปรับลดค่าโดยคูณด้วยค่าสัมประสิทธิ์การลดทอน $d_{C(\varepsilon)}$ ซึ่ง 0 <= $d_{C(\varepsilon)}$ <=1

สำหรับทุก $C(\varepsilon) \geq 1$ ดังนั้นค่าลดทอนแบบง่ายในการนับคือ

$$C^*(\varepsilon) = d_{C(\varepsilon)} C(\varepsilon)$$

ให้ r คือจำนวนครั้งที่นับได้ และ r* คือจำนวนที่ปรับแก้แล้วเขียนสมการใหม่ได้เป็น

$$r^* = rd_r$$

สำหรับการกำหนดค่าสัมประสิทธิ์การลดทอน หรือค่า d_r มีอยู่หลายวิธีด้วยกันได้แก่

Good-Turing discounting

ถ้าเรากำหนดค่า n_r เป็นจำนวนหรือเหตุการณ์ ซึ่งเกิดขึ้นเป็นจำนวน r ครั้ง, รูปแบบสำหรับการลดทอนค่าแบบ Good-turing สามารถเขียนได้ดังนี้

$$d_r = \frac{\frac{(r+1)n_{r+1}}{rn_r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}}$$

สำหรับ r < k (ซึ่งโดยทั่วไปค่า $k \approx 7$) และ d_r = 1 สำหรับ r ที่สูงกว่า k

Linear discounting

ในการหาค่าสัมประสิทธิ์การลดทอนแบบ linear ดังสมการ

$$d_r = 1 - \frac{n_1}{R}$$

โดยที่ R คือจำนวนคำทั้งหมดของข้อมูลที่นำมาฝึก

Absolute discounting

สัมประสิทธิ์การลดทอน หาได้จากการลบจำนวนที่นับได้ออกด้วย b แล้วหารด้วยจำนวนที่นับได้นั้นอีกครั้งหนึ่ง

$$d_r = \frac{r - b}{r}$$

ซึ่งการสมมุติค่า $b = \frac{n_1}{n_1 + 2n_2}$ ให้ค่าออกมาดีที่สุด

Witten-Bell discounting

การหาค่าสัมประสิทธิ์การลดทอนของ Witten-Bell โดยสัดส่วนในการลดทอนค่าไม่ขึ้นอยู่กับจำนวนที่นับได้ แต่จะขึ้นอยู่กับ t โดย t เป็นจำนวนเหตุการณ์ที่แตกต่างกันเช่นสำหรับ bi-gram "A B", t เป็นจำนวน ของ bi-gram . "A *" ทั้งหมดในแบบจำลอง

$$d_r(t) = \frac{R}{R + t}$$

เมื่อนำเอา เทคนิคการลดทอนค่ารวมเข้ากับเทคนิคการทำ smoothing backoff จะได้สมการเป็น

$$P_{kate}(w_i \mid w_{i-1}) = \begin{cases} C(w_{i-1}w_i)/C(w_{i-1}) & if\ r > k \\ d_r C(w_{i-1}w_i)/C(w_{i-1}) & if\ k \geq r > 0 \\ \alpha(w_{i-1})P(w_i) & if\ r = 0 \end{cases}$$

โดยที่ $d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}}$ และ $\alpha(w_{i-1}) = \frac{1 - \sum_{w_i:r>0} P_{Katz}(w_i \mid w_{i-1})}{1 - \sum_{w_i:r>0} P(w_i)}$

2.5.7 เทคนิคการลดขนาดของแบบจำลองภาษา

Count-Cutoffs

เป็นเทคนิคที่ง่ายและธรรมดาที่สุด โดยสมมุติว่าค่าความน่าจะเป็นของคำ z ที่ตามด้วยคำ x และ y หาได้จาก

$$P(z \mid xy) = \begin{cases} \frac{C(xyz) - D(C(xyz))}{C(xy)} & if\ C(xyz) > 0 \\ \alpha(xy)P(z \mid y) & otherwise \end{cases}$$

โดยที่ C(xyz) หมายถึงการนับจำนวน xyz ที่เกิดขึ้นทั้งหมดในข้อมูลที่นำมาฝึก ฟังชันก์ α เป็นค่าคงที่สำหรับ normalization ส่วนฟังก์ชัน D(C(xyz)) คือ discount ฟังชันก์

เทคนิคการ cutoff, สมมุติว่าถ้า cutoff ที่ 3 , ถ้า C(xyz) <= 3 จะไม่นำมาใส่ใน แบบจำลองทางภาษา ถึงแม้ C(z) จะมากกว่า 3 ก็ตาม สำหรับเทคนิคนี้จะมีผลมากในแบบจำลองที่มีขนาดเล็ก, จะพบว่าค่า perplexity เพิ่มขึ้นอย่างเห็นได้ชัดเมื่อเพิ่มค่า cutoff

N-gram Pruning

เมื่อลำดับแบบจำลอง n-gram สูงขึ้น, ขนาดของแบบจำลองก็จะจะมีขนาดใหญ่โตขึ้นด้วย การ pruning เพื่อลดขนาดของ n-gram จึงมีความจำเป็นยิ่ง โดยแบบจำลองที่ prune แล้วจะต้องให้ความแตกต่างของผลลัพธ์ระหว่าง แบบจำลองที่ prune แล้วและของเดิม แตกต่างกันน้อยที่สุด

Class n-gram

นอกจากนี้ยังได้มีการใช้เทคนิคการแบ่งแยกประเภทของคำเพื่อเพิ่มเติมความถูกต้องของแบบจำลอง โดยจัดคำแยกเป็นกลุ่มเช่น

วัน เดือน ปี, ...

ส้ม แตง ฝรั่ง ขนุน ทุเรียน, ...

น้ำเงิน ฟ้า ส้ม แดง เขียว เหลือง, ...

บทที่ 3
วิธีดำเนินการวิจัย

ในการประยุกต์ทฤษฎีเจเนติกอัลกอริทึมช่วยในเพิ่มศักยภาพของ OCR ผู้วิจัยได้ดำเนินการวิจัยตามหัวข้อต่อไปนี้

3.1 เครื่องมือที่ใช้ในการวิจัย

3.2 กระบวนการประยุกต์ทฤษฎีเจเนติกอัลกอริทึมมาใช้ในงานวิจัย

3.3 รวบรวมข้อมูลและวิเคราะห์ข้อมูลเพื่อหาประสิทธิภาพ

3.1 เครื่องมือที่ใช้ในการวิจัย

3.1.1 โปรแกรมทดสอบในการคัดเลือกประโยคที่ถูกต้องโดยทฤษฎีเจเนติกอัลกอริทึม

3.1.2 อุปกรณ์ที่ใช้สร้างเครื่องมือในการวิจัยประกอบด้วย

- เครื่องคอมพิวเตอร์ประเภท PC
- ระบบปฏิบัติการ Windows 2000 Professional
- โปรแกรมภาษา Visual C++ 6.0 Enterprise Edition
- แผ่น CD-ROM MSDN Library
- โปรแกรม Matlab Version
- แผ่น Diskette และ CD-ROM

3.2 กระบวนการประยุกต์ทฤษฎีเจเนติกอัลกอริทึมมาใช้ในงานวิจัย

3.2.1 ศึกษาทฤษฎีเจเนติกอัลกอริทึม โทเกนพาสซิ่งอัลกอริทึม(Token Passing Algorithm) และรูปแบบจำลองภาษา(Language Model)

3.2.1.1 ศึกษาทฤษฎีเจเนติอัลกอริทึมที่จะนำมาใช้ในการสร้างเครื่องมือ

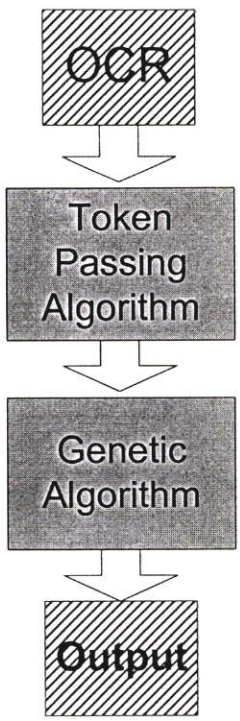
3.2.1.2 ศึกษางานวิจัยที่เกี่ยวกับเจเนติกอัลกอริทึม จากวารสารหรืองานประชุมวิชาการระดับนานาชาติ เพื่อที่จะนำไปประยุกต์ใช้กับงานวิจัย

3.2.1.3 ศึกษาทฤษฎีโทเกนพาสซิ่งอัลกอริทึม(Token Passing Algorithm)

3.2.1.4 ศึกษารูปแบบจำลองภาษา(Language Model)

3.2.2 ออกแบบโครงสร้างของงานวิจัยในการตรวจแก้คำผิดใน OCR ภาษาไทย

เป็นการออกแบบระบบโครงสร้างของงานวิจัยทั้งหมด ประกอบไปด้วย ส่วนของ OCR, โทเกนพาสซิ่งอัลกอริทึม, เจเนติกอัลกอริทึม ดังรูปที่ 3.1



รูปที่ 3.1 ขบวนการตรวจแก้คำผิดใน OCR ภาษาไทย

- OCR

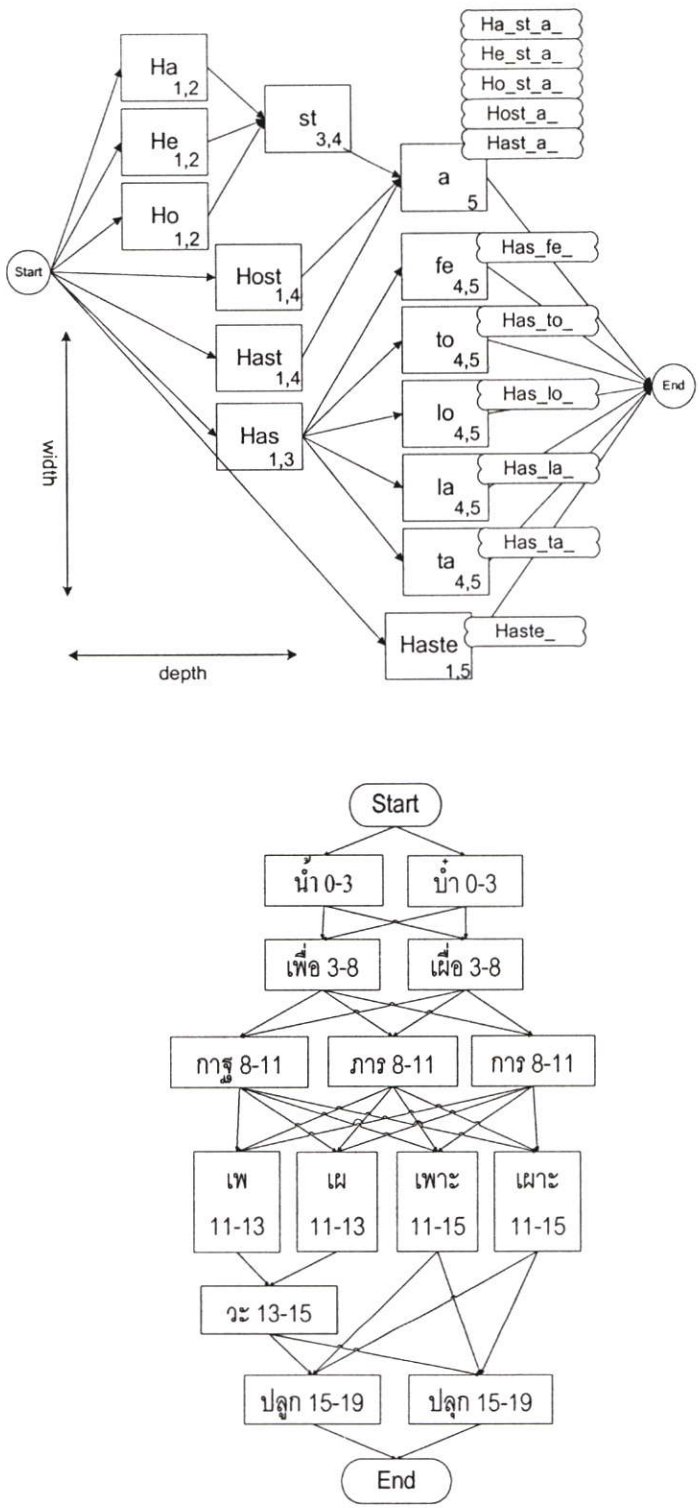
ในส่วนนี้จะเป็นส่วน Input ของงานวิจัยซึ่งจะประกอบไปด้วยตัวอักษรที่มีอัตรา Recognition Probabilities ที่มีค่าสูงกว่า 0.70 ดังตารางที่ 3.1

ตารางที่ 3.1 ค่า Recognition Probabilities ที่สามารถ Recognize ได้

H (0.87)	a (0.88)	s (0.86)	t (0.85)	o (0.89)
	O (0.77)	5 (0.79)	f (0.79)	a (0.78)
	e (0.75)		l (0.79)	e (0.76)
	d (0.74)		1 (0.74)	d (0.73)
	0 (0.72)			0 (0.73)

- Token Passing Algorithm

ในส่วนนี้จะเป็นส่วนหาคำที่เป็นไปได้จากตัวอักขระที่ได้รับมา ซึ่งลักษณะการหาคำได้อธิบายไว้ในบทที่ 2 โดย Output ที่ได้จะเป็น Word Graph ของคำที่สามารถเป็นไปได้ดังรูปที่ 3.2



รูปที่ 3.2 Word Graph ที่ได้จากขบวนการโทเกนพาสซิ่งอัลกอริทึม

- Genetic Algorithm

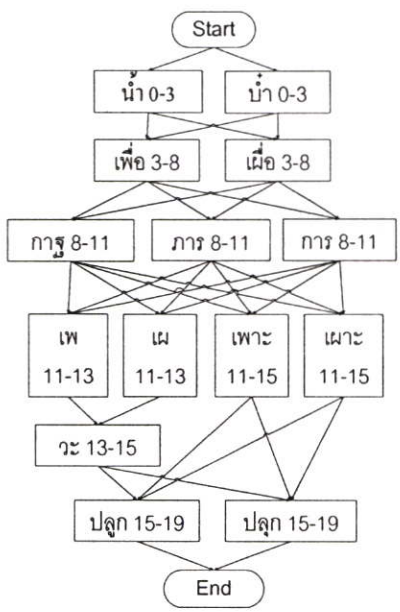
ในส่วนนี้จะเป็นส่วนหาประโยคที่ถูกต้องจาก Word Graph ซึ่งจะได้อธิบายในส่วนหัวข้อถัดไป

- Output

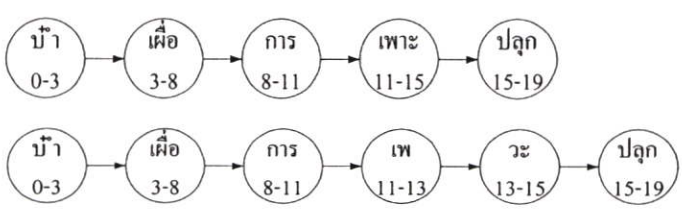
ในส่วน Output จะได้คำตอบที่เป็นประโยคที่ถูกต้องที่ทำการแก้ไขคำผิด

3.2.3 **นำทฤษฎีเจเนติกอัลกอริทึมมาใช้ในการหาประโยคที่ถูกต้องโดยปรับปรุงรูปแบบของปัญหาให้อยู่ในรูปแบบโครโมโซมตามแบบแผนทฤษฎีเจเนติก อัลกอริทึม และกำหนดฟังก์ชั่นเป้าหมายหรือฟังก์ชั่นความเหมาะสมของปัญหา**

3.2.3.1 ทำการปรับรูปแบบของปัญหาในการหาประโยคที่ถูกต้องให้เข้ากับทฤษฎีเจเนติกอัลกอริทึม จัดให้อยู่รูปแบบของโครโมโซม โดยความยาวโครโมโซมได้จากจากเส้นทางของคำที่เป็นไปได้ จากโหนดเริ่มต้นจนถึงเส้นทางโหนดปลายทางดังรูป ซึ่งความยาวโครโมโซมบางทีอาจจะไม่เท่ากันซึ่งแล้วแต่ค่าที่ได้มาจากวิธีโทเกนพาสซิ่งอัลกอริทึม



length chromosome = Start node to end node



รูปที่ 3.3 Word Graph ที่ได้จากกระบวนการ Token Passing และตัวอย่างโครโมโซมที่ถูกเลือก

3.2.3.2 กำหนดฟังก์ชั่นความเหมาะสม (Fitness Function) ที่จะถูกเลือกของโครโมโซมนั้นหรืออาจพูดได้ว่าเป็นฟังก์ชั่นที่กำหนดค่าความเหมาะสมของแต่ละโครโมโซมเปรียบเสมือนค่าความสามารถในการอยู่รอดของแต่ละโครโมโซมและเป็นฟังก์ชั่นที่กำหนดโอกาสหรือสัดส่วนที่แต่ละโครโมโซมเหมาะสมจะถูกคัดเลือกไปใช้มากน้อยเพียงใด ดังสมการ

$$Fitness_1 = \frac{\sum_{i=1}^{n} \Pr obability(node\ i)}{n}$$

$$Fitness_2 = Perplexity(sentence)$$

$$Fitness_T = \left(\frac{Fitness_1 * 30}{100}\right) + \left(\frac{11 - \log_{10}(Fitness_2)}{10} * \frac{70}{100}\right)$$

โดย n คือ จำนวน โหนดที่ผ่าน (จำนวนความยาวโครโมโซม)

$Fitness_1$ คือ ค่าเฉลี่ยค่าความน่าจะเป็นของคำที่ได้จาก OCR

$Fitness_2$ คือ ค่าความคลุมเครือของประโยคที่ได้จากรูปแบบจำลองภาษา(Language Model)

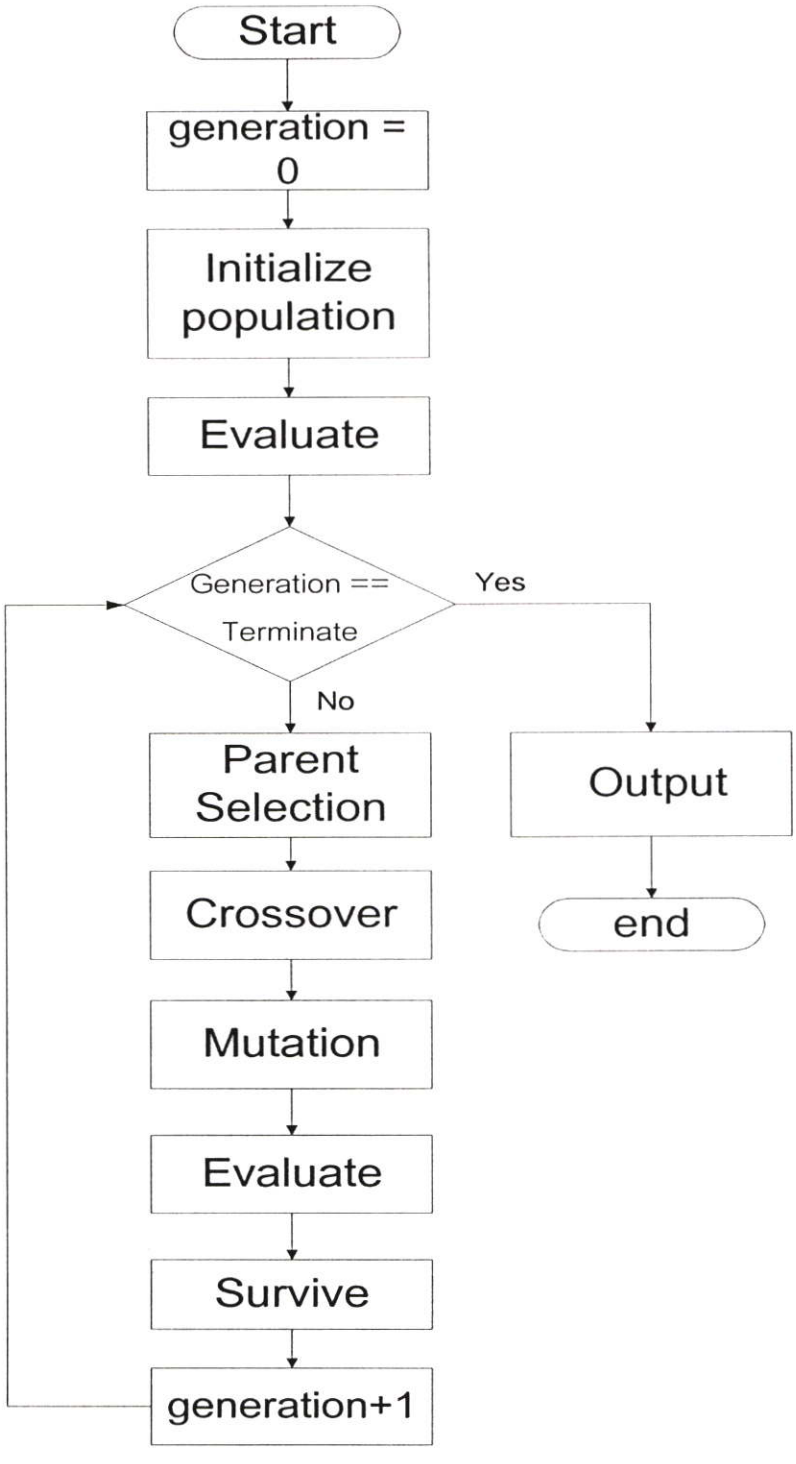
$Fitness_T$ คือ ค่าเฉลี่ยของความถูกต้องของประโยค

3.2.4 สร้างโปรแกรมคอมพิวเตอร์ตามทฤษฎีเจเนติกอัลกอริทึมเพื่อหาประสิทธิภาพในการหาประโยคที่ถูกต้อง

ดำเนินการเขียนโปรแกรมตามโครงสร้างของทฤษฎีของเจเนติกอัลกอริทึม ดังโครงสร้างของทฤษฎีดังรูปที่ 3.4 และ 3.5

```
Procedure GA
        {
         t := 0;
         Initialize population P(t);
         // สร้างประชาการโครโมโซมต้นกำเนิดโดยการสุ่ม
         Evaluate P(t);
          // วิเคราะห์ค่าความเหมาะสมแต่ละโครโมโซมประชากรต้นกำเนิด
         While not terminated (do)
          // ตรวจสอบเงื่อนไขความพอใจ
         {
            Pp(t) := Parents_selection  P(t);
            // คัดเลือกโครโมโซมต้นแบบจากประชาการรุ่นก่อน
            P'(t) := Crossover  Pp(t);
            // แลกเปลี่ยนส่วนยีนส์ภายในโครโมโซมต้นแบบ
            P"(t) := Mutation  P'(t);
            // มิวเตชั่นโครโมโซมต้นแบบ
            Evaluate P"(t);
            // วิเคราะห์ค่าความเหมาะสมของประชากรรุ่นใหม่
            P(t+1) := Select(P"(t) U Q);
            // เลือกประชากร
            t := t+1;
          }
         }
```

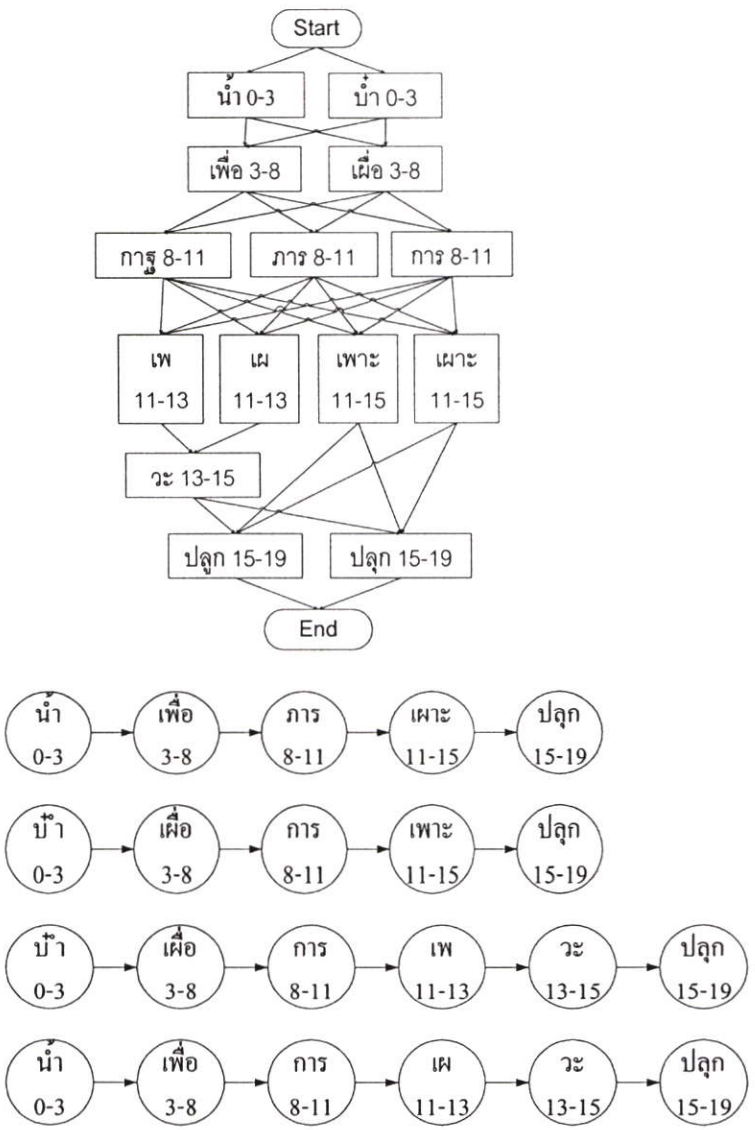
รูปที่ 3.4 โครงสร้างของเจเนติกอัลกอริทึม



รูปที่ 3.5 ไดอะแกรมโครงสร้างเจเนติกอัลกอริทึม

ส่วนต่างๆ ของไดอะแกรมรูปที่ 3.5 ซึ่งเป็นกระบวนการของทฤษฎีเจเนติกอัลกอริทึมที่ใช้ในงานวิจัยในการแก้ข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทย โดยอธิบายเป็น 7 หัวข้อย่อยดังนี้

1. Initialize Population เป็นการสร้างประชากรต้นแบบโดยการสุ่มจาก Graph ตั้งแต่ Start Node ถึง End Node ดังรูปตัวอย่างที่ 3.6

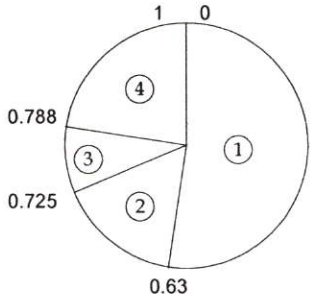


รูปที่ 3.6 รูปตัวอย่างของการสร้างประชากรต้นแบบโดยการสุ่มจาก Word Graph

2. Evaluate วิเคราะห์ค่าความเหมาะสมแต่ละโครโมโซม จากสมการของ Fitness Function

3. Parent Selection เป็นการเลือกโครโมโซมพ่อ-แม่เพื่อที่จะกระทำการดำเนินการทางพันธุศาสตร์ โดยวิธีการสุ่มโครโมโซมต้นแบบของเจเนติกอัลกอริทึมแบบง่ายนั้นเป็นแบบจำลอง

การหมุนวงล้อถ่วงน้ำหนัก (Roullet Wheel : RW) ซึ่งกำหนดขนาดแต่ละช่องของวงล้อนั้นตามความน่าจะเป็นที่จะสุ่มได้ในแต่ละครั้งของแต่ละโครโมโซม

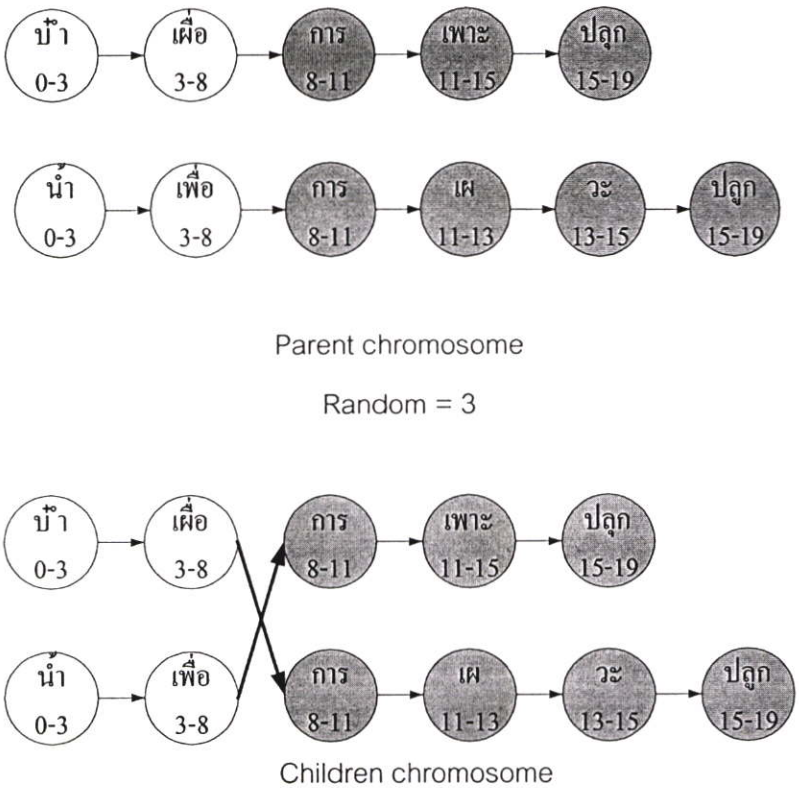


รูปที่ 3.7 รูปตัวอย่างของ Roulette Wheel

4. Crossover เป็นการดำเนินในการแลกเปลี่ยนส่วนของโครโมโซมพ่อ-แม่ ตามการกำหนดอัตราความน่าจะเป็นของการครอสโอเวอร์ (Probability of Crossover : Pc) เพื่อสร้างชุดโครโมโซมรุ่นใหม่หรือโครโมโซมลูก โดยมีโครงสร้างดังรูปที่ 3.8 และตัวอย่างของการครอสโอเวอร์ดังรูปที่ 3.9

```
Procedure crossover;
{
   Crossover_count := population = 0;
   total _ crossover = Pc * (Pop_size/ 2)
         while Crossover_count  < total _ crossover
         {
         Parent1 := Random using Roulette wheel P(t);
         Parent2 := Random using Roulette wheel P(t);
         If(parent1 and parent2 have crossing point)
         {   Position  :=  Random from all crossing point;
            Children:=Crossover(Parent1,Parent2,Position);
            Crossover_count++; population +=2;        }
         }
         while (population < Pop_size)
         {   Children := Random using Roulette wheel P(t);
            population ++;                              }
}
```

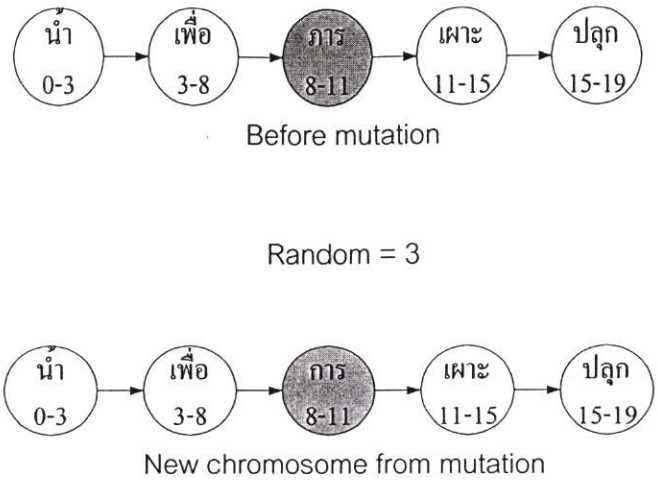
รูปที่ 3.8 รูปโครงสร้างของการ Crossover



รูปที่ 3.9 รูปตัวอย่างของการ Crossover

จำนวนการครอสโอเวอร์ในแต่ละรุ่นดำเนินการขึ้นอยู่กับการกำหนดค่า P_c ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรแต่ละรุ่น popsize เท่ากับ 6 โครโมโซม และกำหนดให้ P_c = 0.666 แล้วจำนวนการครอสโอเวอร์ในแต่ละรุ่นเท่ากับ P_c*(popsize / 2) = 0.666 * (6 / 2) = 2 ครั้ง (การครอสโอเวอร์หนึ่งครั้งเกิดจากโครโมโซมสองโครโมโซม)

5. Mutation เป็นตัวดำเนินการผ่าเหล่าตัวหนึ่งที่อาจช่วยให้โครโมโซม มีค่าความเหมาะสมดีขึ้นหลังจากการครอสโอเวอร์ โดยกลับค่าบาง Node ของโครโมโซมให้เป็นค่าใหม่ ตามอัตราความน่าจะเป็นของการมิวเตชั่นในแต่ละบิท (Probability of Mutation: Pm) โดยในการมิวเตชั่นนั้นจะตัวที่มาแทนจะต้องอยู่ในตำแหน่งของลำดับตัวอักษรเดียวกันจึงจะสามารถทำการมิวเตชั่นได้ดังตัวอย่างรูปที่ 3.10



รูปที่ 3.10 รูปตัวอย่างของการมิวเตชั่น

จำนวนการมิวเตชันในแต่ละรุ่นขึ้นอยู่กับการกำหนดค่า P_m ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรแต่ละรุ่น popsize เท่ากับ 6 โครโมโซม ซึ่งความยาวโครโมโซม 5 และกำหนดให้ $P_m = 0.1$ แล้วจำนวนการมิวเตชันในแต่ละรุ่นเท่ากับ $P_m * popsize * lchrom = 0.1 * 6 * 5 = 3$

6. Evaluate วิเคราะห์ค่าความเหมาะสมแต่ละโครโมโซมลูกที่ได้จากการการดำเนินการทางพันธุศาสตร์ จากสมการของ Fitness Function

7. Survive เป็นกระบวนการในการตัดสินใจคัดเลือกโครโมโซมที่ดีไว้ เนื่องจากในบางครั้งการค้นหาคำตอบของเจเนติกอัลกอริทึมนั้นมีโอกาสที่จะสูญเสียโครโมโซมในรุ่นเก่า ที่มีค่าความ เหมาะสมที่ดีไปได้ ซึ่งจะทำให้ได้คำตอบในรุ่นถัดไปนั้นดีมากหรือน้อยลง ดังนั้น ควรมีการปรับปรุงให้ควบคุมการค้นหาคำตอบ โดยรักษาโครโมโซมที่ดีไว้แล้วจะช่วยให้วิวัฒนาการคำตอบในรุ่นถัดไปดีขึ้นเรื่อยๆ

3.3 รวบรวมข้อมูลและวิเคราะห์ข้อมูลเพื่อหาประสิทธิภาพ

ฐานข้อมูลที่ใช้ในงานวิจัยได้มากจากหนังสือพิมพ์โดยจัดเก็บในลักษณะ Text File เพื่อนำมาใช้ในรูปแบบของภาษา(Language Model) และการหาประสิทธิภาพของทฤษฎีเจเนติกอัลกอริทึมในการหาประโยคที่ถูกต้องจากได้เปรียบเทียบกับทฤษฎีการหาประโยคที่ถูกต้องแบบ Full Search

บทที่ 4
ผลการทดลอง

การวิจัยครั้งนี้เป็นการวิจัย เพื่อแก้ไขข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทย ให้มีความถูกต้องมากขึ้นโดยทฤษฎีเจเนติกอัลกอริทึม ผลการทดสอบเริ่มจากการนำอินพุทสตริงจาก OCR ที่มีค่าความน่าจะเป็นสูงสุดของตัวอักษรแต่ละตัวมาสร้างเป็นคำที่เป็นไปได้โดยให้อยู่ในรูปของ Word Graph โดยวิธีการโทเกนพาสซิ่งอัลกอริทึม จากนั้นก็จะหาประโยคที่ถูกต้องโดยการประยุกต์ใช้ทฤษฎีเจเนติกอัลกอริทึมมาช่วยในการค้นหาประโยคที่ถูกต้องจาก Word Graph โดยค่า Fitness Function จะเป็นการนำเอาค่าความน่าจะเป็นของคำที่ได้จาก โทเกนพาสซิ่งอัลกอริทึม และ ค่าความคลุมเคลือ(Perplexity) จาก Language Model มาช่วยในหาประโยคที่ถูกต้อง

จากตารางที่ 4.1 ได้สรุปถึงผลการทดสอบสมการ Fitness function ของ ทฤษฎีเจเนติกอัลกอริทึมโดยให้มีการจำลองตัวอักษรที่ผิดในส่วนของอินพุทที่ได้จาก OCR เป็นจำนวน 0, 10, 20, 30 ตัวอักษร จากจำนวนตัวอักษรทั้งหมด 282 ตัวอักษร

ดังตัวอย่างที่ผิด 30 ตัวอักษร

"*เริ่นต้นมะ แม้ว่าจะยังไม่นีข้อสรุปจาภรัจบาลว่าจะสร้างโรงไฟฟ้าหินภรูดหรือไม่ แต่ล่าสุดภลุ่มชาวบ้ามผู้ดัดค้านการก่อสร้างโดรงการโรงไฝฟ้าดังกล่าว ได้รวมตัวกัน เข้ายึดพื้มที่สาธารณะชายหะเล บริเวญโดธงภารก่อสร้างโรงไฟฝ้าหินกรูด เพื่อลร้างบ้ามพักรับรฮง สถานหี่ประชุม แสะเวทีสาธารณะ จขการทดสฮบ*"

โดยประโยคที่ถูกต้องคือ

"*เริ่มต้นนะ แม้ว่าจะยังไม่มีข้อสรุปจากรัฐบาลว่าจะสร้างโรงไฟฟ้าหินกรูดหรือไม่ แต่ล่าสุด กลุ่มชาวบ้านผู้คัดค้านการก่อสร้างโครงการโรงไฟฟ้าดังกล่าว ได้รวมตัวกัน เข้ายึดพื้นที่สาธารณะชายทะเล บริเวณโครงการก่อสร้างโรงไฟฟ้าหินกรูด เพื่อสร้างบ้านพักรับรอง สถานที่ประชุม และเวทีสาธารณะ จบการทดสอบ*"

จากนั้นการทดลองให้มีการกำหนดค่า Generation = 30, Crossover ratio = 0.6, Mutation ratio = 0.05

ตารางที่ 4.1 ผลการทดสอบสมการ Fitness Function

Wrong Char / Fitness		0	10	20	30
$Fitness_1$	จำนวนผิด	0	10	20	30
	%ความถูกต้อง	100%	96.45%	92.90%	89.36%
$Fitness_2$	จำนวนผิด	11	9	15	15
	%ความถูกต้อง	96.09%	96.80%	94.68%	94.68%
$Fitness_T$	จำนวนผิด	2	4	7	6
	%ความถูกต้อง	99.29%	98.58%	97.57%	97.81%

$$Fitness_1 = \frac{\sum_{i=1}^{n} \Pr obability(node\ i)}{n}$$

$$Fitness_2 = Perplexity(sentence)$$

$$Fitness_T = \left(\frac{Fitness_1 * 30}{100}\right) + \left(\frac{11 - \log_{10}(Fitness_2)}{10} * \frac{70}{100}\right)$$

โดย n คือ จำนวน โหนดที่ผ่าน (จำนวนความยาวโครโมโซม)

$Fitness_1$ คือ ค่าเฉลี่ยค่าความน่าจะเป็นของคำที่ได้จาก OCR

$Fitness_2$ คือ ค่าความคลุมเครือของประโยคที่ได้จากรูปแบบจำลองภาษา(Language Model)

$Fitness_T$ คือ ค่าเฉลี่ยของความถูกต้องของประโยค

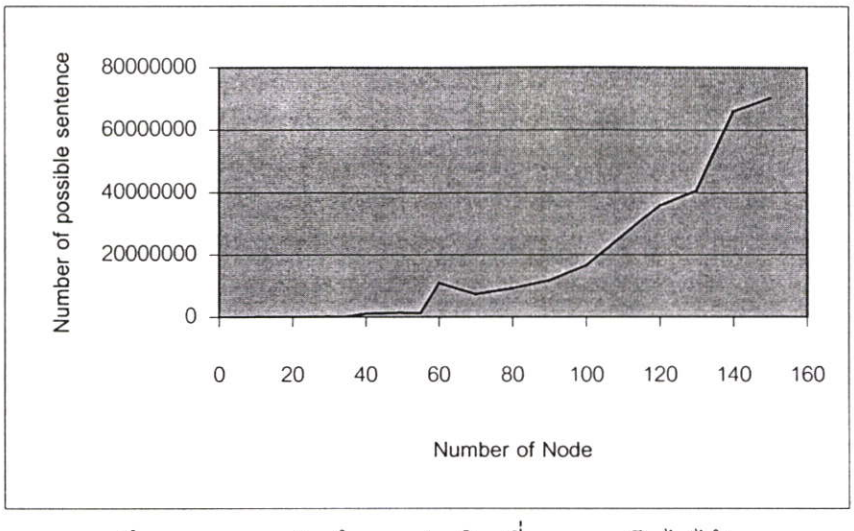
จากตารางที่ 4.2 จะประกอบไปด้วยประโยคที่ทำการทดสอบที่ได้จาก OCR โดยเริ่มจากตัวอักษรที่มีค่าความน่าจะเป็นไปได้สูงสุด 5 ตัวอักษรมาเข้าสู่ขบวนการโทเกนพาสซิ่งอัลกอริทึม จากนั้นได้คำที่เป็นไปได้ทั้งหมดที่อยู่ในรูปแบบของ Word Graph ออกมา ดังเช่นประโยค "หากปีไหน" ซึ่งจะเกิดคำที่เป็นไปได้อยู่ 14 คำเรียงกันเป็น Word Graph และเมื่อทำการหาประโยคที่ถูกต้องจาก Word Graph นั้นโดยการไปทุกเส้นทางของ Word Graph จะได้ประโยคจำนวน 48 ประโยคที่เป็นไปได้ จากนั้นก็นำไปตรวจสอบหาความถูกต้องของประโยคจาก ค่าความน่าจะเป็นของคำที่ได้จากขบวนการโทเกนพาสซิ่งอัลกอริทึม และ ค่า Perplexity จาก Language Model ซึ่งจะรวมกัน โดยในตารางคือค่า Optima (ค่าของประโยคที่ตรงกับ Input หรือค่าของประโยคที่ถูกต้อง) จากนั้นทำการทดสอบเป็นจำนวน 10 ครั้งในแต่ละประโยคโดยใช้ทฤษฎีเจเนติกอัลกอริทึม เพื่อหาค่าความน่าจะเป็นของคำที่ได้จากขบวนการโทเกนพาสซิ่งอัลกอริทึมและค่า Perplexity จาก Language Model โดยจะเก็บค่าที่ดีที่สุดและค่าไม่ดีที่สุดไว้ ดังแสดงในตาราง สุดท้ายจากตารางจะมีค่า Frequency ซึ่งจะเป็นค่าโอกาสที่จะได้ค่าที่ถูกต้องที่ตรงกับอินพุทของตัวอักษรนั้น

จากผลการทดลองเรากำหนดค่า Popuation size = 30 ,Generation = 20, Crossover ratio = 0.6, Mutation ratio = 0.05 โดยการทดสอบจะทำการ Run Program 10 ครั้ง ต่อ 1 Sentence โดยผลการทดสอบเราสรุปดังตารางที่ 4.2

ตารางที่ 4.2 ผลการทดลองในการ Run Program 10 ครั้ง ต่อ 1 Sentence

Input Text and Number of node	Posible sentence	Optima	GA		frequency	O/P
			Best	wrost		
หากปีไหน (14)	48	0.817	0.817	0.817	100%	หากปีไหน
ห้ามหญิงเที่ยวผับ (25)	234	0.724	0.724	0.724	100%	ห้ามหญิงเที่ยวผับ
วัฒนธรรมการกินอยู่ (52)	2,037	0.8789	0.8789	0.8068	90%	วัฒนธรรม การ กิน อยู่
ในการประชุมระหว่างประเทศ (58)	2,298	0.9176	0.9176	0.8452	90%	ใน การ ประชุม ระหว่าง ประเทศ
แนะหนังสือดีขนมลูก (55)	1,560	0.7807	0.7807	0.733	90%	แนะ หนังสือ ดี ขนม ลูก
ผลการทดลองที่ได้จากการคำนวณ (74)	323,680	0.94708	0.94708	0.82	60%	ผล การ ทดลอง ที่ ได้ จาก การ คำนวณ
ทุกคนสบายดีหรือเปล่า (80)	27,272	0.85244	0.85244	0.8003	70%	ทุก คน สบาย ดี หรือ เปล่า
งานเทศกาลเฉลิมฉลองวันปีใหม่ (86)	370,320	0.77539	0.77539	0.705	40%	งาน เทศกาล เฉลิม ฉลอง วัน ปี ใหม่
ก่อนกินทำให้สุกถั่วงอกดิบอันตราย (104)	1,741,824	0.8327	0.8327	0.77	20%	ก่อน กิน ทำ ให้ สุก ถั่ว งอก ดิบ อันตราย
นโยบายกวาดล้างยาเสพติดโลก (117)	1,330,428	0.8472	0.8472	0.80	20%	นโยบาย กวาดล้าง ยา เสพ ติด โลก
หลังจากที่ตำรวจเข้าตรวจค้นและได้หลักฐานสำคัญหลายอย่างที่จะสาวไปถึงตัว ผู้บงการใหญ่ (217)	>10,000,000	0.8467	0.8467	0.75	10%	หลังจาก ที่ ตำรวจ เข้า ตรวจ ค้น และ ได้ หลัก ฐาน สำคัญ หลาย อย่าง ที่ จะ สาว ไป ถึง ตัว ผู้ บง การ ใหญ่

จำนวน Node ที่ได้จากขบวนการ Token Passing Algorithm ในตารางที่ 4.2 ยิ่งมีจำนวนมากเท่าไหร่จะทำให้สามารถเกิดประโยคที่เป็นไปได้จำนวนมากขึ้น ดังรูปที่ 4.1



รูปที่ 4.1 แสดงกราฟจำนวน Node กับ จำนวนประโยคที่สามารถเป็นไปได้

ในการทดลองขั้นต่อไปพวกเราได้ทดสอบโดยการเปลี่ยนค่า Population Size(Popsize) เพื่อดูว่าการเปลี่ยนแปลงค่า Population Size มีผลต่อการเปลี่ยนแปลงในการหาประโยคที่ถูกต้องหรือไม่ โดยในผลการทดลองได้ทดสอบเอาประโยคที่มีจำนวน Node จำนวน 74 node ที่สามารถเกิดประโยคทั้งหมดที่เป็นไปได้จำนวน 323,680 คำมาทำการทดสอบ โดยกำหนดค่าอัตราต่างๆ คงเดิมเหมือนในการทดลองในตารางที่ 4.1 เปลี่ยนแปลงเพียงแต่ค่า Popsize โดยสรุปได้ดังตารางที่ 4.3

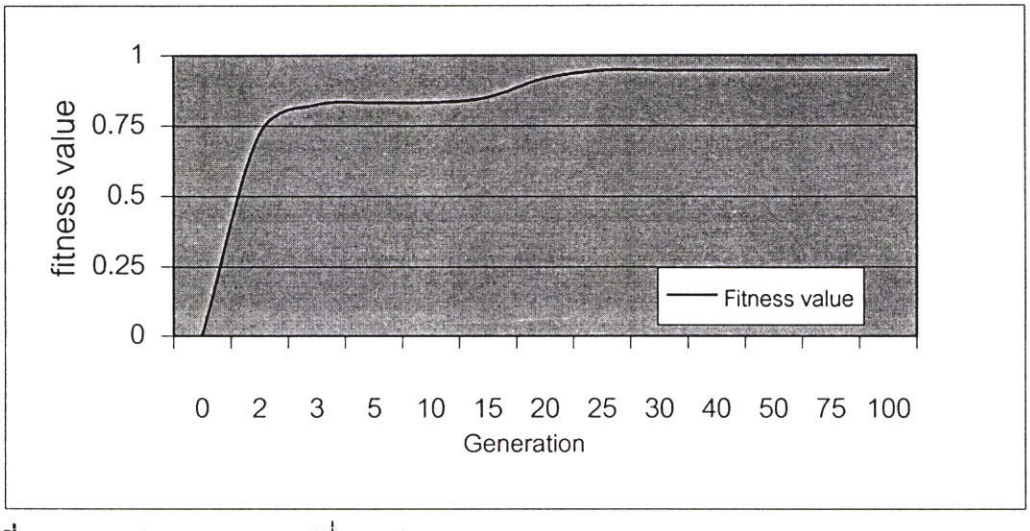
ตารางที่ 4.3 ผลการทดลองในการเปลี่ยนแปลงค่า Population Size

Popsize	Frequency
6	10%
10	20%
16	40%
26	50%
30	60%
50	80%

จากนั้นได้ทดสอบโดยการเปลี่ยนค่า Generation เพื่อดูว่าการเปลี่ยนแปลงค่า Generation มีผลกระทบต่อการเปลี่ยนแปลงในการหาประโยคที่ถูกต้อง โดยกำหนดค่าอัตราต่างๆ คงเดิมเหมือนในการทดลองในตารางที่ 4.2 เปลี่ยนแปลงเพียงแต่ค่า Generation โดยสรุปผลการทดลองดังตารางที่ 4.4 และดังรูปกราฟที่ 4.2

ตารางที่ 4.4 ผลการทดลองในการเปลี่ยนแปลงค่า Generation

Generation	Frequency
3	10%
5	30%
10	40%
15	50%
20	60%
25	60%
30	60%
40	70%
50	80%



รูปที่ 4.2 กราฟแสดงผลการเปลี่ยนแปลง Generation

ในการทดสอบขั้นต่อไปเราได้ทำการทดสอบในส่วนของ Genetic Algorithm ในการค้นหาประโยคที่ถูกต้องโดยเปรียบเทียบกับ Full Search (Possible sentences) เพื่อหาประสิทธิภาพของการนำเอาทฤษฎี Genetic Algorithm มาช่วยในการหาประโยคที่ถูกต้อง โดยสรุปได้ดังตารางที่ 4.5

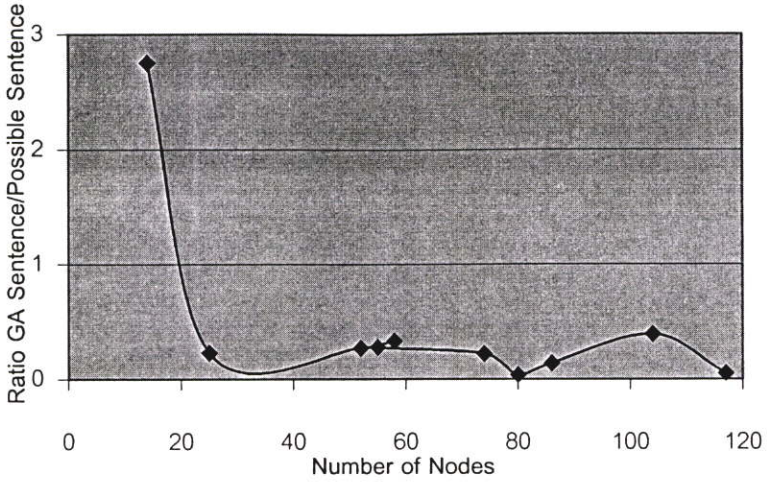
ตารางที่ 4.5 ผลการทดลองในการเปรียบเทียบเพื่อหาประสิทธิภาพของทฤษฎี

Number of Node	Possible sentence	Generation of optimal solution		Mean Generation	Y
		Min	Max		
14	48	1	13	4.4	2.75
25	234	1	3	1.8	0.230
52	2,037	4	39	18.2	0.268
58	2,298	9	254	254	0.331
55	1,560	1	43	14.4	0.276
74	323,680	72	5912	2385.4	0.221
80	27,272	6	358	358	0.039
86	370,320	19	7253	1719	0.1392
104	1,741,824	13	8210	2268.4	0.390
117	1,330,428	25	7304	2325	0.0524

จากตารางที่ 4.5 ค่า Y เป็นค่าวัดประสิทธิภาพที่ได้จากการนำเอาทฤษฎี Genetic Algorithm มาช่วยในการหาประโยคที่ถูกต้อง โดยค่า Y หาจาก

$$Y = \frac{(Popsize \times Mean\ Generation)}{Possible\ Sentences}$$

จากนั้นเรานำผลการทดลองที่ได้จากตารางที่ 4.5 มาพล็อตกราฟได้ดังรูปที่ 4.3 เพื่อแสดงให้เห็นถึงประสิทธิภาพในการเปรียบเทียบระหว่าง Genetic Algorithm และ Full Search



รูปที่ 4.3 กราฟแสดงประสิทธิภาพของทฤษฎี Genetic Algorithm โดยเทียบกับ Full Search

ในลำดับถัดไปได้ทดสอบประสิทธิภาพในการหาตัวอักษรที่ถูกต้องโดยให้อินพุทเป็นตัวอักษรที่ผิดจำนวน 0, 20, 40, 60 ตัวอักษรของประโยคทั้งหมดจำนวน 282 ตัวอักษร จากนั้นทดสอบดูผลความถูกต้องของตัวอักษรที่ทำการแก้ไขว่ามีความถูกต้องมากน้อยเพียงใด

อินพุทของตัวอักษรที่ทดสอบดังเช่น "จบการทดสอบ" ซึ่งจะเอาคำผิดใส่เข้าไปในประโยค จากนั้นนำตัวอักษรที่ใกล้เคียงของตัวอักษรมาทำการหาคำที่เป็นไปได้ดังเช่น

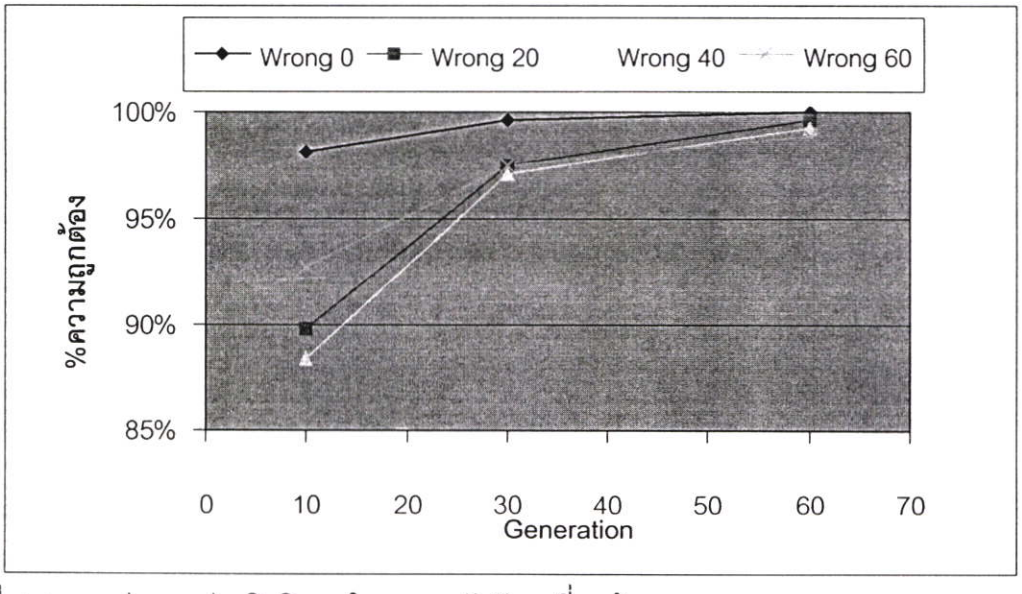
ฐ → จ
บ → ข → ษ → ม → น
ภ → ถ → ก
ว → า
ร → โ → เ
ท → ห
ค → ด → ต → ศ
ส → ล
อ → ฉ → ฮ
น → บ → ม → ข → ษ

จากคำใกล้เคียงจากนั้นก็จะสร้างคำขึ้นมาโดยกระบวนการโทเกนพาสซิ่งอัลกอริทึม และเข้าสู่กระบวนการหาประโยคที่ถูกต้องโดยใช้เจเนติกอัลกอริทึม ซึ่งจากผลการทดลองโดยการเปลี่ยนค่า อินพุทโดยเพิ่มตัวอักษรที่ผิดเข้าไป ซึ่งสรุปผลได้ดังตารางที่ 4.6

ตารางที่ 4.6 ผลการทดลองในการหาประสิทธิภาพในการหาตัวอักษรที่ถูกต้อง

Wrong Char \ Generation		10	30	60
0	%ความถูกต้องทั้งหมด	98.17%	99.63%	100%
	จำนวนผิดจากทั้งหมด	14	1	0
20	%ความถูกต้องทั้งหมด	89.81%	97.45%	99.63%
	จำนวนผิดจากทั้งหมด	28	7	1
40	%ความถูกต้องทั้งหมด	83.36%	97.09%	99.27%
	จำนวนผิดจากทั้งหมด	32	8	2
60	%ความถูกต้องทั้งหมด	92.72%	97.45%	98.90%
	จำนวนผิดจากทั้งหมด	20	7	3

จากนั้นเรานำผลการทดลองที่ได้จากตารางที่ 4.6 มาพล็อตกราฟได้ดังรูปที่ 4.4 เพื่อแสดงให้เห็นถึงประสิทธิภาพในการหาตัวอักษรที่ถูกต้อง



รูปที่ 4.4 กราฟแสดงประสิทธิภาพในการหาตัวอักษรที่ถูกต้อง

ในส่วนผลการทดลองลำดับต่อไปเป็นการทดลองเปรียบเทียบผลการทดลองระหว่างการแก้ไขข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทยด้วยวิธีโทเกนพาสซิ่งอัลกอริทึม Kruatrachue(2001 : 599 – 602) โดยการ Prunning กับวิธีแก้ไขข้อผิดพลาดของตัวอักษรที่ได้จาก OCR ภาษาไทยด้วยโดยเจเนติกอัลกอริทึม

ตารางที่ 4.7 ผลการทดลองในการเปรียบเทียบความถูกต้องในการแก้ไขข้อผิดพลาด

Input Character	Wrong Char	Number of Sentence (Prunning)	% Correct Sentence		Number of Sentence (Genetic)	% Correct Sentence	
			Wrong	%		Wrong	%
25 Char	0	32	0	100%	150	0	100%
	3	32	0	100%	150	0	100%
	6	32	0	100%	150	0	100%
50 Char	0	1044	0	100%	900	0	100%
	5	1044	0	100%	900	0	100%
	10	1044	0	100%	900	0	100%
75 Char	0	3983	4	94.6%	1500	1	100%
	8	3983	4	94.6%	1500	2	98%
	15	3983	4	94.6%	1500	5	93.3%
90 Char	0	2068	5	94.4%	1500	0	100%
	10	2068	5	94.4%	1500	0	100%
	20	2068	5	94.4%	1500	5	94.4%

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการดำเนินการวิจัยตามขั้นตอน เราได้นำเอาทฤษฎี เจเนติกอัลกอริทึมมาช่วยในหาประโยคที่ถูกต้องโดยเริ่มจากการเอาตัวอักขระที่เป็นไปได้สูงสุด 5 ตัวอักษร มาใช้ในการหาคำที่เป็นไปได้โดยขบวนการโทเกนพาสซิ่งอัลกอริทึม โดยคำที่ได้มานั้นจะเป็นคำที่มีอยู่ใน Dictionary จากนั้นก็นำไปเข้าสู่กระบวนการ เจเนติกอัลกอริทึม เพื่อหาประโยคที่ถูกต้อง จากนั้นทำการตรวจเช็คในรูปคำข้างเคียงโดยใช้รูปแบบจำลองทางภาษา (Language Model) เพื่อเพิ่มประสิทธิภาพให้ดียิ่งขึ้น ในงานวิจัยได้สรุปผลการวิจัยตามแนวทางวัตถุประสงค์ คือเพื่อพัฒนาความถูกต้องใน OCR ภาษาไทยให้มีความถูกต้องมากขึ้นและ เพื่อใช้ทฤษฎีเจเนติกอัลกอริทึม มาช่วยเพิ่มความเร็วในการหาประโยคที่ถูกต้อง โดยสรุปได้ดังนี้

สรุปในส่วนการตรวจแก้คำผิดใน OCR ภาษาไทย ซึ่งจะทำให้ได้ผลการทดลองมีประสิทธิภาพมากขึ้นดังผลการทดสอบที่ทดลองใส่ตัวอักษรที่ผิดจำนวน 0 – 60 ตัวอักษร ในส่วน Input จากผลการทดลองแสดงให้เห็นว่าได้พัฒนาความถูกต้องใน OCR ภาษาไทย เนื่องจากสามารถแก้คำผิดจากตัวอักษรที่ไม่ถูกต้องจำนวนมากให้ถูกต้องมากขึ้น ได้เป็นอย่างดี เนื่องจากมีการตรวจสอบทั้งในส่วนระดับของคำและในส่วนของคำข้างเคียง (Language Model) ทำให้เพิ่มประสิทธิภาพให้ที่ได้สามารถแก้คำผิดใน OCR ภาษาไทย ได้อย่างมีประสิทธิภาพ

สรุปในส่วนการนำเอาทฤษฎีเจเนติกอัลกอริทึมมาช่วยเพิ่มความเร็วในการหาประโยคที่ถูกต้องนั้น เนื่องจากว่าหลังจากเอาตัวอักขระที่เป็นไปได้สูงสุด 5 ตัวอักษร มาใช้ในการหาคำที่เป็นไปได้โดยขบวนการโทเกนพาสซิ่งอัลกอริทึมแล้ว ถ้าในกรณีที่เป็นประโยคยาวๆ จะทำให้เกิดประโยคที่เป็นไปได้จำนวนมาก จึงมีการนำเอาทฤษฎีเจเนติกอัลกอริทึม มาช่วยค้นหาประโยคที่ถูกต้องที่เป็นไปได้ที่สุด โดยผลการทดลองเราได้เปรียบเทียบกับ Full Search จะแสดงให้เห็นว่าในกรณีที่อักขระที่รับเข้ามานั้นเป็นประโยคที่ยาวเมื่อเอาทฤษฎีเจเนติกอัลกอริทึม มาใช้จะช่วยให้มีความรวดเร็วในการหาคำตอบที่ถูกต้องยิ่งขึ้นแทนที่จะหาในทุกเส้นทางโดย Full Search แต่ในกรณีที่ประโยคสั้นๆ แบบ Full Search จะมีประสิทธิภาพที่ดีมากกว่า ซึ่งสรุปได้ว่าการนำเอาทฤษฎีเจเนติกอัลกอริทึม มาช่วยนั้นจะทำให้เราสามารถหาประโยคที่ถูกต้องได้อย่างรวดเร็วมากขึ้นในกรณีที่ประโยคนั้นมีความยาวของประโยคตั้งแต่ปานกลางถึงยาวมาก

5.2 ข้อเสนอแนะ

ในงานวิจัยในการใช้ Genetic Algorithm ในการแก้คำผิดใน OCR ภาษาไทยนั้น ผู้วิจัยขอเสนอแนะดังนี้

- ในงานวิจัยนี้เป็นแก้คำผิดจาก OCR ในกรณีของตัวอักษรผิดหรือไม่ชัดเจนที่เกิดจาก OCR เท่านั้น ส่วนในกรณีที่มีตัวอักษรขาดหรือเกินนั้นยังไม่สามารถทำได้ จึงต้องทำการพัฒนาต่อไปในอนาคต

- ในการเพิ่มประสิทธิภาพความถูกต้องให้มากยิ่งขึ้นเราสามารถทำได้โดยสร้าง Language Model ที่ดี และกำหนดค่า Fitness Function ให้เหมาะสม ซึ่งจะทำให้ความถูกต้องเพิ่มมากขึ้น

- ศึกษาทฤษฎี Evolutionary Computation เพิ่มเติม ซึ่งอาจจะช่วยเพิ่มประสิทธิภาพในการค้นหาคำตอบที่ถูกต้องและรวดเร็วมากยิ่งขึ้น

- ในส่วนของการหาคำตอบที่ถูกต้องจาก Genetic Algorithm นั้น ยังมีความคลุมเครือในส่วนของการกำหนด Generation ที่เหมาะสมกับประโยคเนื่องจากในบางประโยคถ้ากำหนดจำนวน Generation มากเกินไปก็จะทำให้สูญเสียเวลาในการค้นหามากไป แต่ถ้ากำหนดน้อยไปก็อาจจะไม่เจอประโยคที่ถูกต้อง ดังนั้นจึงต้องมีการพัฒนาต่อไป

บรรณานุกรม

กาญจนี วงศ์วิภาพร. 2541. "การจัดตารางสอนของโรงเรียนแบบอัตโนมัติโดยจีเนติกอัลกอริทึม." , วิทยานิพนธ์วิศวกรรมศาสตร์มหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง..

ยุทธนา ลีลาศวัฒนกุล. 2544. **คู่มือการเขียนโปรแกรมและใช้งาน Visual C++ 6.0 ฉบับโปรแกรมเมอร์**. กรุงเทพฯ : อินโฟเพรส.

วีระศักดิ์ วัฒนายากร และเอื้อน ปิ่นเงิน. 2542. "ระบบตรวจสอบลายเซ็นแบบออนไลน์โดยใช้จีเนติกอัลกอริทึม." วิทยานิพนธ์เทคโนโลยีสารสนเทศมหาบัณฑิต, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

รวีวรรณ ชินะตระกูล. 2542. **การทำวิจัยทางการศึกษา**. กรุงเทพฯ : ที.พี.พริ้นท์.

Ackley, D. H. 1985. "A Connectionist Algorithm for Genetic Search." **Proceedings of an International Conference on Genetic Algorithm and Their Application.** 7 : 121-135.

Berry, R. J. 1965. **Genetics.** London : English University Press.

Clarson, P. and Rosenfeld, R. **Statistical Language Modeling Using the CMU-CAMBRIDGE Toolkit.** [Online]. Available : http://www.svr-www.eng.cam.ac.uk/~prc14/toolkit.html

Cooley, J. W. and Tukey, J. W. 1965. "An Algorithm for the Machine Computation of Complex Fourier Series." **Math Computation.** 19 : 297-381

Feller, W. 1968. **An Introduction to Probability Theory and its Application.** New York : Wiley.

Fogel, D. B. 1992. "An Analysis of Evolutionary Programming." 43-51. **Proceedings of the First Annual Conference on Evolutionary Programming.**

Goldberg, D. E. 1989. **Genetic Algorithms in Search, Optimization, and Machine Learning.** Massachusetts : Addison-Wesley Publishing Company, Inc.

Golding, A. R. et. al. 1999. "A Winnow base Approach to Context Sensitive Spelling Correction." 107-130. **ICML96.** Boston.

Holland, J.H. 1973. "Genetic Algorithm and the Optimal Allocation of Trials." **SIAM Journal of Computing.** 2(2) : 88-105.

John, M. et. al. 1994. "Disambiguation and Spelling Correction for a Neural Network based Character Recognition System." 322-333. **Proceedings Document Recognition.**

Juan, C. et. al. 2000. "Stochastic Error Correcting Parsing for OCR Post Processing." **International Conference on Pattern Recognition.**

Kazem, T. et. al. 1994. "An Expert System for Automatically Correction OCR output." 270-278. **International Symposium on Electronic Imaging Science and Technology.**

Kruatrachue, B. et. al. 2001. "Thai OCR Error Correction Using Token Passing Algorithm." 599-602. **IEEE Pacific Rim Conference: Communications, Computer and Signal Processing.**

Lawrence, D. 1989. "Adapting Operator Probabilities in Genetic Algorithm." 60-69. **Proceedings of the Third International Conference on Genetic Algorithm.**

Man, K. F. and others. 1997. **Genetic Algorithms for Control and Signal Processing.** London : Springer.

Man, K. F. and others. 1999. **Genetic Algorithms Concepts and Designs.** London : Springer.

Mekavin, S. et. al. 1998. "Progress of Combining Trigram and Winnow in Thai OCR Error Correction." **IEEE Conference.**

Michalewicz, Z. 1996. **Genetic Algorithms + Data Structures = Evolution Program.** 3rd. ed. London : Springer.

Promchan, P. 1998. "Performance Comparison of Thai Word Separation Algorithm." **NCSEC98 Conference.**

Radcliffe, A. 1981. "A Problem Solving Technique Based on Genetics." **Creative Computing.** 3(2) : 78-81.

Shaefer, C. G. 1987. "The ARGOT Strategy: Adaptive Representation Genetic Optimizer Technique." 50-58. **Proceedings of the Second International Conference on Genetic Algorithm.**

Theeramunkong, T. and Usanavisin, S. 2001. "Non Dictionary Based Thai Word Segmentation Using Decision Trees." **First International Conference on Human Language Technology Research.**

Uwe, Q. 1998. "Tool for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Value."

Vasconcelos, J. A. et. al. 2001. "Improvements in Genetic Algorithms." 3414-3417. **IEEE Transactions on Magnetics.**

Young, S.J. et. al. 1989. "Token Passing: a Simple Conceptual Model for Connected Speech Recognition System." Cambridge University Engineering Department.

ภาคผนวก

ภาคผนวก ก

ลักษณะตัวอักษรที่ใกล้เคียงกัน

ลักษณะตัวอักษรที่ใกล้เคียงกัน

ก ภ ถ	ท ห	อ ฉ ฮ
ข บ	ธ ร	ฮ อ ฉ
ฃ ข บ	น บ ม ข ษ	ๆ ๅ
ค ด ต ศ	บ ข ษ ม น	ะ
ฅ ด ค ศ	ป ฝ ฟ พ	◌ั ◌็ ◌่
ฆ ม	ผ บ พ	า ว
ง	ฝ ฟ พ	ำ า ว
จ ฐ	พ ผ	◌ิ ◌ี ◌ึ ◌ื
ฉ อ ฮ	ฟ ฝ พ ผ	◌ิ ◌ี ◌ึ ◌ื
ช ซ	ภ ถ ก	◌ิ ◌ี ◌ึ ◌ื
ฌ ช ข	ม น บ ย ษ	◌ิ ◌ี ◌ึ ◌ื
ญ ณ ญ	ย บ น ม ษ	◌ุ
ญ ญ ณ	ร โ เ	◌ู
ฎ ฏ	ฤ ๅ	เ โ ใ ไ ร
ฏ ฎ	ล ส	โ ใ ไ ร
ฐ จ	ๅ ฤ	ใ ไ เ
ฑ ท	ว า	ไ ใ เ
ฒ ต	ศ ค ด ต	า า ว ๆ
ณ ญ ฌ	ษ บ ข	
ด ต ค ศ	ส ล	◌ั ◌็ ◌่
ต ด ค ฒ ศ	ห ท ฑ	◌ั ◌็ ◌่
ถ ภ ก	ฬ ฟ ฝ	

ภาคผนวก ข

ผลการทดสอบความถูกต้องในการหาประโยคที่ถูกต้อง

ประโยคที่ทำการทดสอบ

Wrong = 0(**จำนวนตัวอักษรที่ผิด 0 ตัว**)

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค
อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย
ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงดูดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย
ให้มากยิ่งขึ้น งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า แก่ประชา
ชนทั่วไปสินค้าจะมีหลากหลายราคา
แข่งขันได้คุณภาพ โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่
สินค้านำมาแสดงจะเป็นสินค้า ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

Wrong = 20(**จำนวนตัวอักษรที่ผิด 20 ตัว**)

ถรมส่งเสริม**ภ**ารส่งออกกระทรวงพาณิชย์**จ**ัด**ง**านแสดงสิ**ม**ค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาส**ัน**พันธ์สัญลักษณ์ตราสิ**ม**ค้าไทย
ให้เป็นที่รู้จักและยอ**นร**ับแก่ผู้บริโภค
อีกทั้งเป็นการกระตุ้**มใ**ห้ผู้ส่งออก ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสิ**ม**ค้าไทย
ได้เห็นถึงความสำคั**ณ**
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงดูดผู้เข้าร่**วน**งานชาวต่างประเ**หศ**
เพื่อสร้างค่านิย**น**ในการบริโภคสินค้าไทยของชาวไทย
ให้มากยิ่งขึ้น งานแส**ค**งต**ธ**าสินค้า**ไหย**นี้จะจั**ค**ขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนั**ภ**ธุรกิจชาวต่างประเทศจา**ภ**ทั่วโลก
ได้มีโอกาลเจร**ฐ**าการค้า**กั**ษผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า แก่ประชา
ชนทั่วไปสินค้าจะมีหลากหลายราคา
แข่งขันได้คุณภาพ โดยจะมีบริษัทผู้ส่งอ**ฮ**กเข้าร่วมงานในพื้นที่
สินค้านำมาแสดง**ฐ**ะเป็นสินค้า ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

Wrong = 40(**จำนวนตัวอักษรที่ผิด** 40 **ตัว**)

ถรมส่งเสรินภารส่งออก**ถระทรวง**พาณิชย์**จัค**งานแสดง**สิม**ค้าสัญลักษณ์ต**ธา**สินค้า**ไหย**
ซึ่งงานนี้ได้**จัค**ขึ้นมาแล้วเพื่อเผยแพร่ประชา**สัน**พันธ์สัญลักษณ์ตรา**สิม**ค้าไทย
ให้เป็นที่รู้จัก**แสะ**ยอ**นรับ**แก่ผู้บริโภค
อีกทั้งเป็น**ถา**รกระตุ้**มให้**ผู้ส่งออ**ถ** ที่ยังไม่ได้เข้า**ร่วน**โค**ธง**การสัญลักษณ์ตรา**สิม**ค้าไทย
ได้**เห็ม**ถึงควา**นสำคัณ**
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงดู**ค**ผู้เข้า**ร่วน**งานชาวต่างประ**เหศ**
เพื่อสร้างค่านิย**น**ในการบริโภค**ด**สินค้าไทยของชาวไทย
ให้มา**ภ**ยิ่งขึ้น งานแส**คง**ต**ธา**สินค้าไทยนี้จะ**จัค**ขึ้นระหว่างพฤศจิกายน
จะเป็น**วัม**เปิดให้เฉพาะ**นัภ**ธุรกิจชาวต่างประเทศจา**ภ**ทั่วโล**ภ**
ได้มีโอกาส**เฐร**จา**ถา**รค้า**กัษ**ผู้ส่งออ**ภ**ไทย สำ**หรัษ**วันหลังคือจะเป็น**วัม**ขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลา**ภ**หลายราคา
แข่ง**ขัม**ได้คุณภา**ผ** โดยจะมีบริษัทผู้ส่งอ**ฮ**กเข้า**ร่วม**งานในพื้นที่
สินค้านำมาแสดง**ฐ**จะเป็นสินค้า ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

Wrong = 60(**จำนวนตัวอักษรที่ผิด** 60 **ตัว**)

ถรมส่งเสรินภารส่งอ**ฮก**ถระทรวงพาณิชย์**จัค**งานแสดง**สิม**ค้าสัญลักษณ์ต**ธา**สินค้า**ไหย**
ซึ่งงานนี้ได้**จัค**ขึ้นมาแล้วเพื่อเผย**แผร่**ประ**ซา**สัน**พันธ์สัญลักษณ์ตรา**สิม**ค้าไทย
ให้เป็**มที่**รู้**จัภแสะ**ยอ**นรับ**แก่ผู้บริโภค
อี**ภ**ทั้งเป็น**ถา**รกระตุ้**มให้**ผู้ส่งออ**ถ** ที่ยังไม่ได้เข้า**ร่วน**โค**ธง**การ**สัณ**ลักษณ์ต**ธา**สิม**ค้า**ไทย
ได้**เห็ม**ถึงควา**นสำคัณ**
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามาร**กดึง**ดู**ค**ผู้เข้า**ร่วน**งานชาวต่างประ**เหศ**
เพื่อสร้างค่านิย**น**ในการบริโภค**ด**สินค้า**ไหย**ของชาวไทย
ให้มา**ภ**ยิ่ง**ขึ้ม** งานแส**คง**ต**ธา**สินค้าไทยนี้จะ**จัค**ขึ้นระหว่างพฤศจิกาย**ม**
จะเป็น**วัม**เปิดให้เฉพาะ**นัภ**ธุรกิจชาวต่างประเทศจา**ภ**ทั่วโล**ภ**
ได้มี**โฮ**กาสเฐรจา**ถา**รค้า**กัษ**ผู้ส่งออ**ภ**ไทย สำ**หรัษ**วันหลังคือจะเป็น**วัม**ขายปลีกสินค้า
แก่ประ**ซา**ชนทั่วไปสินค้าจะมีหลา**ภ**หลายรา**ดา**
แข่ง**ขัม**ได้คุณภา**ผ** โดยจะมีบริษัทผู้ส่งอ**ฮก**เข้า**ร่วน**งานในพื้**มที่**
สินค้านำมาแสดง**ฐ**จะเป็นสินค้า ที่ใช้สัญ**ลัภ**ษณ์ตราสินค้า**ไหย**มาขาย

ผลการทดสอบ

Generation = 10 Wrong 0

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสตงสินค้าสัญลักษณ์ตราสินค้าโทษ
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดิงดุดบู่เบ้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสีบค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยบี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโสภ
ได้มีโอกาสเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 14 ตัวอักษร)

Generation = 30 Wrong 0

กรมส่งเสริมการส่งฉอกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงดูดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้าภัพผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 1 ตัวอักษร)

Generation = 60 Wrong 0

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงดูดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 0 ตัวอักษร)

Generation = 10 Wrong 20

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแ**ล**ดงสินค้าสัญลักษณ์ตราสิน**ค่าโห**ย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
โท่เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระ**ตุ่น**ให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสิน**ค่า**ไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถ**ตึง**ดูดผู้เข้าร่วมงา**ม**ชาวต่างประเทศ
เพื่อสร้างค่านิ**บน**ในการ**ขรี**โภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งาน**แลค**งตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิด**โท่**เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้า**ถับบู่**ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเ**บ้**าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้า**โห**ยมาขาย

(ผิด 28 ตัวอักษร)

Generation = 30 Wrong 20

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน **โท่**สามารถดึง**ดุ**ดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิด**ไห่**เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอ**ภา**สเจรจาการค้ากับ**บู้**ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 7 ตัวอักษร)

Generation = 60 Wrong 20

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึง**ดุ**ดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

Generation = 10 Wrong 40

กรมส่งเสริม**ภา**รส่ง**ฉอ**กกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าโหย
ซึ่งงานนี้ได้จัดขึ้นมาแล้ว**เผื่อ**เผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้า**โทษ**
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้น**ไท้**ผู้ส่งออก
ที่ยังไม่ได้เข้าร่ว**น**โครงการสัญลักษณ์ตราสินค้า**โหย** ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน **โท่**สามารถดึง**ดุ**ดผู้เข้าร่วมงานชาว**ด้**างประเทศ
เพื่อสร้างค่า**บิบนโบ**การบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิด**ไท้**เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
โด่มีโอกาสเจรจาการค้า**ถับ**ผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลา**บธา**ดา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขา**บ**

(ผิด 32 ตัวอักษร)

Generation = 30 Wrong 40

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่**ได้**เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงดูดผู้เข้าร่วมงาน**ชา**วต่างประเทศ
เพื่อสร้างค่า**มียม**ในการบริโภค**ลืบ**ค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสคงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้า**ถับ**ผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 8 ตัวอักษร)

Generation = 60 Wrong 40

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงคุดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสดงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 2 ตัวอักษร)

Generation = 10 Wrong 60

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้ว**เผื่อ**เผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค **ฮึ**กทั้งเป็นการกระ**ตุ่น**ให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโค**ธง**การสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจาก**นี้เพื่อ**สร้างภาพการจัดงาน ให้สามาร**ถตึง**คุดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการ**ขรี**โภคสินค้า**โหย**ของชาว**โหย** ให้มากยิ่งขึ้น
งานแสคงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิด**โท่**เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอ**ภา**สเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลาย**ธาดา** แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 20 ตัวอักษร)

Generation = 30 Wrong 60

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักแสะยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ได้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงคุดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสคงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้าถับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายธาดา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 7 ตัวอักษร)

Generation = 60 Wrong 60

กรมส่งเสริมการส่งออกกระทรวงพาณิชย์จัดงานแสดงสินค้าสัญลักษณ์ตราสินค้าไทย
ซึ่งงานนี้ได้จัดขึ้นมาแล้วเพื่อเผยแพร่ประชาสัมพันธ์สัญลักษณ์ตราสินค้าไทย
ให้เป็นที่รู้จักและยอมรับแก่ผู้บริโภค อีกทั้งเป็นการกระตุ้นให้ผู้ส่งออก
ที่ยังไม่ไค้เข้าร่วมโครงการสัญลักษณ์ตราสินค้าไทย ได้เห็นถึงความสำคัญ
นอกจากนี้เพื่อสร้างภาพการจัดงาน ให้สามารถดึงตูดผู้เข้าร่วมงานชาวต่างประเทศ
เพื่อสร้างค่านิยมในการบริโภคสินค้าไทยของชาวไทย ให้มากยิ่งขึ้น
งานแสคงตราสินค้าไทยนี้จะจัดขึ้นระหว่างพฤศจิกายน
จะเป็นวันเปิดให้เฉพาะนักธุรกิจชาวต่างประเทศจากทั่วโลก
ได้มีโอกาสเจรจาการค้ากับผู้ส่งออกไทย สำหรับวันหลังคือจะเป็นวันขายปลีกสินค้า
แก่ประชาชนทั่วไปสินค้าจะมีหลากหลายราคา แข่งขันได้คุณภาพ
โดยจะมีบริษัทผู้ส่งออกเข้าร่วมงานในพื้นที่ สินค้านำมาแสดงจะเป็นสินค้า
ที่ใช้สัญลักษณ์ตราสินค้าไทยมาขาย

(ผิด 3 ตัวอักษร)

ภาคผนวก ค

ผลงานวิจัย

Proceedings

First International Symposium on Cyber Worlds

6-8 November 2002 • Tokyo, Japan

Sponsored by

Hosei University

Edited by

Shietung Peng

Vladimir V. Savchenko

Shuichi Yukita

Los Alamitos, California

Washington • Brussels • Tokyo

Thai OCR Error Correction Using Genetic Algorithm

Boontee Kruatrachue Krich Somguntar Kritawan Siriboon
Research Center for Communication and Information Technology (ReCCIT)
Computer Engineering Department Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang
boontee@ce.kmitl.ac.th s3061624@kmitl.ac.th kritawan@ce.kmitl.ac.th

Abstract

This paper presents an efficient method for Thai OCR error correction based on genetic algorithm (GA). The correction process start with word graph construction from spell checking with dictionary, then a graph is search for a corrected sentence with the highest perplexity (using language model, bi-gram and tri-gram) and word probability from OCR. For a long sentence, a search space is huge and can be resolved using GA. A list of nodes is used for chromosome encoding to represent all possible paths in a graph instead of standard binary string. The performance of the suggested technique is evaluated and compared to the full search for tested sentences of different size constructed from 10 nodes to 200 nodes word graphs.

1. Introduction

The use of dictionary, and language model (bigram, trigram) has been widely used for error correction of English text generated from OCR. In Thai language, there is no space between words. The spell checking has to go through all possible ambiguity characters and word boundaries as shown in Table 1. A complete search using token parsing was proposed to this problem [1], where the input text string generated from OCR was a list of up to five most probable characters (Table 2). Tokens are generated for each characters and pass to the next position in the input string (Table 3). Token with characters that is not part of the word or word in dictionary is discarded, other tokens are passed to the next characters positions in the string. Tokens contained complete word is save along with the word position in the string, which are used to constructed word graph (Figure 1).

The corrected sentence is retrieved by a full search of the word graph for a sentence with maximum perplexity using language model and word probability generated from OCR. This search can be implemented by token parsing along graph node with pruning [1]. But for some long sentences full search is impractical. For example, a word graph with 8 levels with 7 nodes in each level can generate 7^8 (5,764,801) possible sentences. In this paper, GA is used to search the corrected sentence.

Table 1. Comparison of spell checking process between language with and without word boundary.

Space between words	No word boundaries
ๆะ ต้อง Has to	ๆะต้อง Hasto
2 words need to do spell checking: Has, to.	First word checked: Ha*, Has, Hast, Hasto Possible second words: (start from Ha) st, sto (start from Has) to
	Possible correction phrase Has to

* Shows a word or beginning of a word in a dictionary

Table 2. Lists of at most 5 characters from OCR with their recognition probabilities.

H .87	a .88	s .86	t .85	o .89
	o .77	5 .79	f .79	a .78
	e .75		l .79	e .76
	d .74		l .74	d .73
	0 .72			0 .73

Table 3. Token passing examples without passing word boundary token.

1	2	3	4	5(char. Position)
H H	a Ha*	s s, Has, Hos,Hes	t t, st, Hast, Host, Hest	o to, lo
	o Ho	5	f ff,sf	a a, la, ta
	e He		l ll, Hasl	e Haste, fe
	d		l	d
	0			0
	Ha,ho,he	Has	Hast, Host,st	to,lo,a,la,ta,fe ,Haste
1	3	4	9	7(num. of token)

Proceedings of the First International Symposium on Cyber Worlds (CW'02)
0-7695-1862-1/02 $17.00 © 2002 IEEE

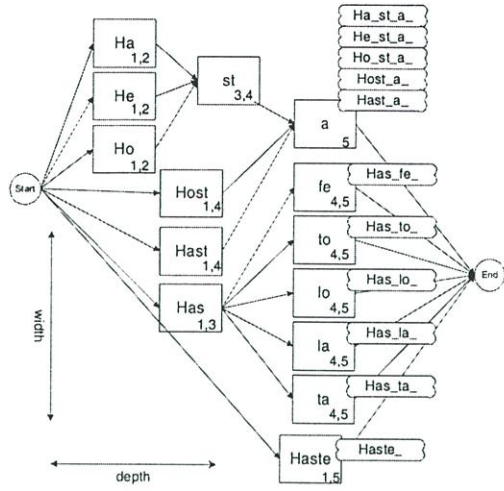


Figure 1. Word graph generated from character level token passing with 11 tokens at the end node.

2. The Genetic algorithm

The concept of genetic algorithm [2], [3] is to simulate the natural evolution process. In nature, the individuals constituting a population adapt to the environment in which they live. The fittest individuals have the highest probability of survival and tend to increase in number, while the less fit individuals tend to die out. This survival of the fittest is the basic idea behind the genetic algorithm.

The algorithm maintains a population of individuals, each of which corresponds to a specific solution. A measure of fitness defines the quality of an individual. Starting with a set of random individuals, a process of evolution is simulated. The main components of this process are crossover and mutation, which mimic propagation and random changes occurring in nature. After a number of generations, highly fit individuals will emerge corresponding to good solution to the given optimization problem. Figure 2 outlines a typical genetic algorithm.

```
Procedure GA{
    t := 0;  Initialize population P(t);  Evaluate P(t);
    While not terminated (do){
     Pp(t) := Parents_selection  P(t);
     P'(t) := Crossover  Pp(t);
     P"(t) := Mutation  P'(t);
     Evaluate P"(t);
     P(t+1) := Select(P"(t) U Q); //select the survival
      t := t+1;   } }
```

Figure 2. Structure of the genetic algorithm

In this algorithm, P(t) denotes a population at generation t. Q is a set of population to be consider for selection, e.g. Q = P(t).

After initialization parents are selected according to a probabilistic function based on relative fitness. In other word, those individuals with higher relative fitness are more likely to be selected as parents. An offspring population P'(t) and P"(t) are generated by means of crossover, which exchanges information between parents, and mutation, which further perturbs the offspring. Then the offspring are evaluated. Finally, select(P"(t) U Q) is select the survivors from actural fitness.

2.1. Chromosome Representation

The first step in designing a genetic algorithm for particular problem is to devise a suitable representation scheme. A chromosome representation that stores current solution state is necessary to make available the crossover and mutation operation, and enhance the performance of the algorithm. Since the length of search paths of a graph can vary in length the list representation is used instead of normal binary string or other encoding [4]. As shown in Figure 4, a chromosome is represented by a list of nodes of word in the word graph (Figure 3). A node contains word and its position in the input string. The valid chromosome is the one with nodes that has the start and end position in order without skipping. Since the chromosome is not encoded using binary string, the crossover and mutation are modified accordingly.

Table 4. Corrected words and their position saved from Token Passing Algorithm

Start	End	Data	Probability
0	3	น้ำ	1
0	3	บ่ำ	0.8835
3	8	เพื่อ	1
3	8	เผื่อ	0.97
8	11	การ	0.89
8	11	ภาร	1
8	11	การ	0.98
11	13	เพ	1
11	13	เผ	0.99
11	15	เพาะ	1
11	15	เผาะ	0.99
13	15	วะ	0.91
15	19	ปลูก	0.98
15	19	ปลุก	0.9114

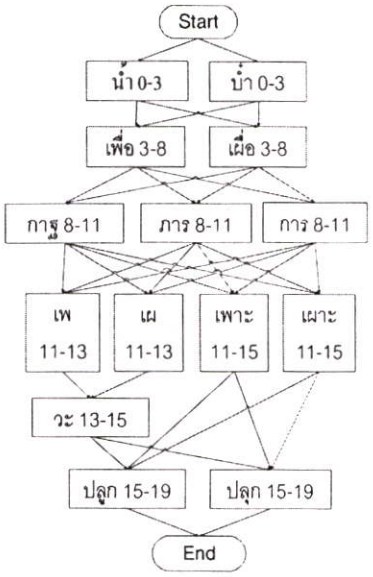


Figure 3. Word graph of Table 4

2.2. Initial Population

The initial populations (search path) are generated randomly from a word graph as shown below.

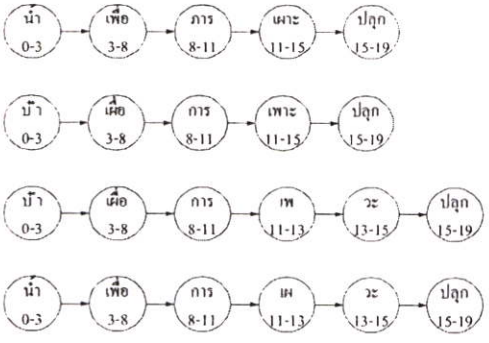


Figure 4. Example of initial population from word graph

2.3. Evaluate

Fitness function of a chromosome is based on each node (word) probability and the perplexity of the whole sentence. The word probability is the average probability of characters in a word provided by OCR. The perplexity is calculated from all words in a search path using language model [5] and normalized to the maximum of 1. The fitness value is weighted with more emphasis on perplexity value (70%) rather than word probability.

$$Fitness_1 = \frac{\sum_{i=1}^{n} Probability(node\ i)}{n} \quad (1)$$

$$Fitness_2 = Perplexity(sentence) \quad (2)$$

$$Fitness_T = \left(\frac{Fitness_1 * 30}{100}\right) + \left(\frac{11 - \log_{10}(Fitness_2)}{10} * \frac{70}{100}\right) \quad (3)$$

Table 5. Example results from computing fitness.

Chromosome	*Fitness1*	*Fitness2*	*FitnessT*
1	0.98028	25946	0.755
2	0.948	9761	0.775
3	0.942	45854	0.726
4	0.965	6183	0.794

2.4. Parent Selection Methods

In this paper we uses the roulette wheel selection scheme [6] as its selection mechanism. The selection probability is the ratio of the individual fitness value to the sum of fitness value of all individuals. The roulette is divided into multiple fans, where the angle of each fan is proportional to the selection probability of the corresponding individual.

2.5. Crossover Methods

In this experiment, we use one-point crossover, where parts of two parent chromosome are swapped at a randomly selected point to create two children chromosomes. Figure 5 shows detail crossover procedure. Parent chromosomes are selected by Roulette wheel. The common character positions of each word between two parent chromosome are located and use as the crossing points. In figure 5 the crossing points are 3, 8, 11, 15 but not 13. Hence, some parents may not be crossable and new parents are re-selected. The number of crossover in each generation is controlled by crossover probability (Pc).

```
Procedure crossover;{
    Crossover_count := population = 0;
     total _ crossover = Pc * (Pop_size/ 2)
     while Crossover_count  < total _ crossover {
      Parent1 := Random using Roulette wheel P(t);
      Parent2 := Random using Roulette wheel P(t);
      If(parent1 and parent2 have crossing point){
        Position  :=  Random from all crossing point;
        Children:=Crossover(Parent1,Parent2,Position);
        Crossover_count++; population +=2; } }
      while (population < Pop_size){
         Children := Random using Roulette wheel P(t);
          population ++; }
}
```

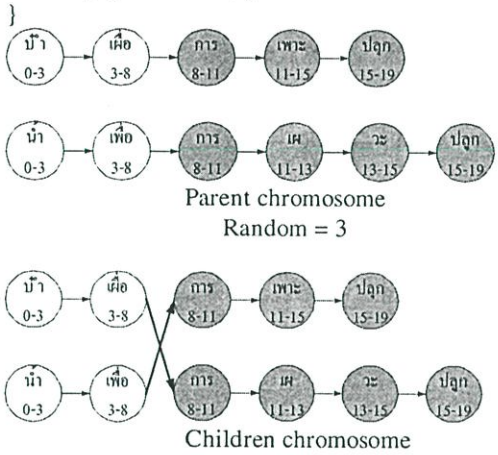


Figure 5. Procedure crossover and example of one-point crossover

2.6. Mutation Methods

Mutation is a background operator, which produces random changes in various chromosomes. In this experimented we use a simple way to achieve mutation by altering one or more genes. We random change word(s) in randomly selected chromosome with alternative word(s) in the word graph. In figure 6, shows example of mutation, where word at position (8,11) are change to other word located at the same position. The Number of mutation perform in each generation is defined by mutation probability (Pm).

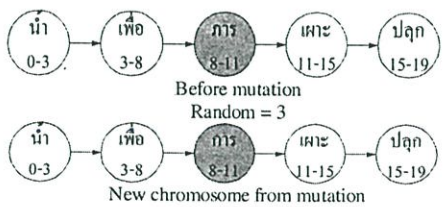


Figure 6. Example of mutation.

2.7. Survive

The next generation populations are selected from the new population generated from crossover and mutation and some of the parent population. The reason that we include the parent in to survival step so that the best population in the previous generation is not lost.

3. Statistical test results

Statistical test results were conducted from strings of five probable characters of various length generated from OCR. First, word graph is extracted using tokens parsing and dictionary. Then genetic is applied to find the correct sentence from all possible sentences. In this experimented we fixed population size to 30, number of generation to 20, crossover ratio to 0.6, mutation ratio to 0.05. The statistical results are average over 10 runs per sentence. The results are summarized in table 6, where node is the number of node in word graph and possible sentence are the number of all possible sentence generated from word graph using full search. From the table 6, we can see that the corrected (optimal) sentence can be located more often for sentence with smaller number of nodes.

Table 6. Results from 10 runs of various length sentences

Node	Possible sentence	Optima	GA		Frequency*
			best	worst	
14	48	0.817	0.817	0.817	100%
25	234	0.724	0.724	0.724	100%
52	2,037	0.870	0.870	0.806	90%
58	2,298	0.917	0.917	0.845	90%
55	1,560	0.782	0.782	0.733	90%
74	323,680	0.947	0.947	0.82	60%
80	27,272	0.852	0.852	0.80	70%
86	370,320	0.775	0.775	0.705	40%
104	1,741,824	0.847	0.847	0.80	20%
117	1,330,428	0.847	0.847	0.80	20%
217	>10,000,000	0.846	0.846	0.75	10%

* Frequency is the percentage for getting the optima calculated from 10 runs

We further varied the population size and number of generation, while fixed the remaining parameters as the previous experiment. The results of the sentence with 74 nodes are shown in Table 7, where the average for obtaining the optimal solution will increase in proportionally to the population and the number of generation.

Table 7.
(a) The population size effect on the frequency of finding optima
(b) The number of generation effect on the frequency of finding optima

(a)

Pop Size	Frequency
6	10%
10	20%
16	40%
26	50%
30	60%
50	80%

(b)

Generation	Frequency
3	10%
5	30%
10	40%
15	50%
20	60%
25	60%
30	60%
40	70%
50	80%

Table 8, compare the number of sentences used in genetic search (GA sentences) and the total number of sentence generated using full search (possible sentences). The data are gathered from the first experiment, where the generation number that locate the optima are recorded and used to calculated GA sentences. The comparison results are also shown in figure 7, where the performance ratio is the ratio of the GA sentences over all possible sentences (full search). From this graph, we can see that genetic search space is very small in comparing to full search especially in medium to large problem. But for small search space problem (48 sentences), GA search space is much higher.

Table 8. Comparing the number of sentences between GA and full search.

Possible sentence	Generation of optimal solution		Mean Generation.	Y
	Min	Max		
48	1	13	4.4	2.75
234	1	3	1.8	0.230
2,037	4	39	18.2	0.268
2,298	9	254	254	0.331
1,560	1	43	14.4	0.276
323,680	72	5912	2385.4	0.221
27,272	6	358	358	0.039
370,320	19	7253	1719	0.1392
1,741,824	13	8210	2268.4	0.390
1,330,428	25	7304	2325	0.0524

Y = (popsize * mean generation)/possible sentence

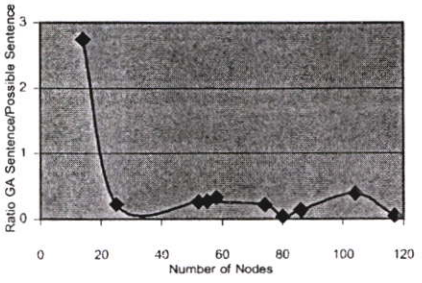


Figure 7. Compare full search and genetic search space on problem size

4. Conclusions

In this paper, we investigated the possibility of using GA to correct sentence generated from the OCR. We present the graph path chromosome representation and genetic operators on the path chromosome. We also compare the GA with full search and experiment to find the suitable size of the problem suitable for GA (number of nodes in word graph > 14 nodes). The experiment results are very encouraging: it can find the best correct sentence very rapidly in comparable to full search especially in medium to large problem.

5. References

[1] B. Kruatrachue, K. Siriboon and M. Rodphone, "Thai OCR Error Correction Using Token Passing Algorithm", IEEE Pacific Rim Conference: Communications, Computer and Signal Processing, 2001., Vol. 2, pp. 599-602

[2] D. B. Fogel. Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press, Piscataway, NJ, 1995.

[3] Th.Back, H.P.Schwefel"Evolutionary Computation An Overview", Proceedings of IEEE International Conference on Evolutionary Computation, 1996., pp. 20-29

[4] M.Gen,R. Cheng and D. Wang "Genetic Algorithm for Solving Shortest Path Problems", Proceedings of IEEE International Conference on Evolutionary Computation, 1997., pp. 401-406

[5] P. Clarson, and R. Rosenfeld. "Statistical Language modeling using the CMU-CAMBRIDGE toolkit", http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html.

[6] D.E Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reding (MA), 1989.

[7] K.F. Man, K.S. Tang and S. Kwong, Genetic Algorithm Concepts and Designs, Springer-Verlag London Limited, 1999.

ประวัติผู้เขียน

นาย กริช สมกันธา ภูมิลำเนา จังหวัดเชียงใหม่ ที่อยู่ปัจจุบัน บ้านเลขที่ 99 ถนนนันทาราม ตำบล หายยา อำเภอ เมือง จังหวัดเชียงใหม่ ปีการศึกษา 2546 สำเร็จการศึกษาปริญญาโท สาขาวิศวกรรมคอมพิวเตอร์ จากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ปีการศึกษา 2541 สำเร็จการศึกษาปริญญาตรีสาขาวิศวกรรมคอมพิวเตอร์ และ ปีการศึกษา 2538 สำเร็จการศึกษาประกาศนียบัตรวิชาชีพชั้นสูงแผนกเทคนิคคอมพิวเตอร์ จากสถาบันเทคโนโลยีราชมงคลวิทยาเขตภาคพายัพจังหวัดเชียงใหม่