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

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.