บทคัดย่อ

การตรวจสอบคำสะกดสามารถนำมาเพิ่มความถูกต้องผลลัพธ์ของ OCR ,ซึ่งจะใช้เวลามาก สำหรับภาษาไทย เนื่องจากไม่มีขอบเขตที่แบ่งระหว่างแยกคำอย่างชัดเจน, การตรวจสอบคำสะกดจึงต้องทำการตรวจสอบทุกตัวอักษรที่ติดกัน และทุกคำที่เป็นไปได้ ซึ่งยังมีความคลุมเครือในการแบ่งแยกขอบเขตคำต่อเนื่องกันไป ในงานวิจัยนี้จะนำเสนอ โทคเกนพาสซิง อัลกอริทึม, ซึ่งมักในปัญหาการรู้จำเสียงคำพูด, เข้ามาช่วยในการแก้ไขปัญหานี้ โดยจะใช้ผลลัพธ์จาก OCR ซึ่งประกอบด้วยตัวอักษรซึ่งมีค่าความน่าจะเป็นสูงที่สุด ห้าตัว โทคเกนจะถูกสร้างขึ้นมา จากตัวอักษรแต่ละตัวและส่งผ่านต่อไปรวมกับตัวอักษรชุดต่อไป ในแต่ละครั้งที่โทคเกนจะถูกส่งไป, จะทำการตรวจสอบตรวจสอบคำสะกดของแต่ละโทคเกนด้วยพจนานุกรม, โทคเกนที่มีคำซึ่งสะกดผิดจะถูกทิ้งไป ส่วนโทคเกนซึ่งเป็นคำที่สมบูรณ์แล้วจะถูกใช้ในการสร้างกราฟของคำ, ซึ่งจะประกอบด้วยคำทั้งหมด ซึ่งมาจากตัวอักษรของแต่ละโทคเกน ที่มีลำดับคะแนนสูงที่สุดห้าตัวอักษร โทคเกนพาสซิง อัลกอริทึมจะถูกนำมาใช้อีกครั้งในระดับของคำในการเลือกประโยคที่มีความความน่าจะเป็นสูงที่สุด

ABSTRACT

Spell checking can be used to improve OCR result, which is quite time consuming for Thai language. Since, there is no explicit word boundary, the spell checking has to go through all possible ambiguity characters and ambiguity word boundary. This paper proposed a Token Passing algorithm, often used in speech recognition, to this problem. The output of the OCR consists string of 5 most probable characters. The letter token are generated for each letters and passed to the next 5 characters. Each time the token is parsed, the dictionary is used to check for the correct spelling. The wrong spelling token are discarded. Tokens with complete word are used to construct words graph, which will be fully constructed when letter token reach the best five characters. Token passing Algorithm is used again in the word level to select the best possible sentence.