การแจงส่วนประโยคผิดไวยากรณ์ในภาษาไทยด้วยตัวแจงส่วนแบบแอลอาร์

นาย ประกาศิต กายะสิทธิ์

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
พ.ศ. 2538

ISBN 974-632-650-3

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

I16659375

PARSING ILL-FORMED THAI SENTENCES WITH THE LR PARSER

Mr. Prakasit Kayasit

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science
Department of Computer Engineering
Graduate School
Chulalongkorn University
1995

ISBN 974-632-650-3

หัวข้อวิทยานิพนธ์ การแจงส่วนประโยคผิดไวยากรณ์ในภาษาไทยด้วยตัวแจงส่วนแบบแอลอาร์

โดย นาย ประกาศิต กายะสิทธิ์

ภาควิชา วิศวกรรมคอมพิวเตอร์

อาจารย์ที่ปรึกษา อาจารย์ ดร. บุญเสริม กิจศิริกุล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ ดร. สันติ ถุงสุวรรณ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ศาสตราจารย์ ทักษิณา สวนานนท์)

..... อาจารย์ที่ปรึกษา

(อาจารย์ ดร. บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาร่วม

(อาจารย์ ดร. สุรพันธ์ เมฆนาวิน)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

ประกาศิต กายะสิทธิ์ : การแจงส่วนประโยคผิดไวยากรณ์ในภาษาไทยด้วยตัวแจงส่วนแบบแอลอาร์ (PARSING ILL-FORMED THAI SENTENCES WITH THE LR PARSER) อ. ที่ปรึกษา : อาจารย์ ดร. บุญเสริม กิจศิริกุล, 79 หน้า. ISBN 974-632-650-3

งานวิจัยที่ทำในวิทยานิพนธ์ฉบับนี้ ได้เสนอแนะการแก้ไขประโยคที่ผิดรูปแบบไวยากรณ์ อันเนื่องมาจากข้อผิดพลาด 3 ประการคือ ข้อผิดพลาดจากการแทรกองค์ประกอบใดๆลงในประโยค ข้อผิดพลาดจากการขาดองค์ประกอบบางตัวในประโยค และข้อผิดพลาดที่เกิดจากการแทนที่องค์ประกอบใดๆในประโยค โดยการปรับปรุงตัวแจงส่วนแบบจีแอลอาร์ใน 3 รูปแบบด้วยกัน คือ 1. การแก้ไขการผิดไวยากรณ์ ณ จุดที่ตรวจพบการผิดไวยากรณ์เท่านั้น 2. การแก้ไขการผิดไวยากรณ์ด้วยการแจงส่วนแบบจีแอลอาร์-ไอจีแอลอาร์ และ 3. การแก้ไขการผิดไวยากรณ์ด้วยการแจงส่วนแบบย้อนรอยกองซ้อนโครงสร้างกราฟ

ผลการวิจัยสามารถสรุปได้ว่า วิธีแรก การแก้ไขการผิดไวยากรณ์ ณ จุดที่ตรวจพบการผิดไวยากรณ์เท่านั้น ยังไม่เพียงพอกับการใช้งานจริง มีข้อผิดพลาดในประโยคหลายจุดที่ถูกมองข้ามไป ส่วนวิธีที่สอง การแจงส่วนแบบจีแอลอาร์-ไอจีแอลอาร์ ก็สามารถครอบคลุมข้อผิดพลาดได้มากขึ้น แต่เนื่องจากไม่มีการย้อนรอยกลับมาแก้ไขประโยคในส่วนที่ทำการแจงส่วนไปแล้ว ดังนั้นข้อผิดพลาดในกรณีที่ตัวแจงส่วนตรวจพบการผิดไวยากรณ์ภายหลังจากที่ได้แจงส่วนผ่านจุดที่เป็นสาเหตุแท้จริงของการผิดไวยากรณ์ไปแล้ว จะถูกมองข้ามไปไม่ทำการแก้ไข

วิธีสุดท้าย การแก้ไขการผิดไวยากรณ์ด้วยการแจงส่วนแบบย้อนรอยกองซ้อนโครงสร้างกราฟ สามารถครอบคลุมการแก้ไขข้อผิดพลาดได้ดีกว่าสองวิธีแรก แต่ก็ใช้เวลาในการประมวลผลมากกว่า อย่างไรก็ตามการจำกัดขอบเขตของการแก้ไขปัญหาให้เหมาะสมกับการใช้งาน จะทำให้วิธีนี้เป็นวิธีที่แก้ไขข้อผิดพลาดได้ดี และใช้เวลาในการประมวลผลไม่มากนัก จึงเป็นวิธีที่เหมาะสมกับการใช้งานจริง

นอกเหนือจากการปรับปรุงตัวแจงส่วนเพื่อให้สามารถแจงส่วนประโยคที่ผิดรูปแบบได้แล้ว ผลลัพธ์ที่ได้จากการแก้ไขประโยค ยังสามารถนำกลับมาใช้เก็บเป็นค่าสถิติเพื่อใช้ในการปรับปรุงพจนานุกรมชนิดคำของคำศัพท์ต่างๆ และใช้ในการปรับไวยากรณ์ที่มีอยู่ให้เหมาะสมกับการใช้งานแต่ละอย่างด้วย

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ประกาศิต กายะสิทธิ์

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา

ปีการศึกษา 2538 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

C517968 : MAJOR COMPUTER SCIENCE
KEY WORD: GLR PARSER / GLR-IGLR / GSS BACKTRACKING / ILL-FORM / PARSER
PRAKASIT KAYASIT : PARSING ILL-FORMED THAI SENTENCES WITH THE LR PARSER. THESIS ADVISOR : BOONSERM KIJSIRIKUL, Ph. D. 79 pp. ISBN 974-632-650-3

This research proposes three strategies to process ill-formed input sentences by means of an extended GLR parser. The first strategy edits error only at the detected position while the other strategies, called GLR-Inverted GLR (GLR-IGLR) parsing and Graph-Structured Stack (GSS) backtracking parsing, not only handle the erroneous element at the detected position but also edit other overlooked positions.

In general, editing error only at the detected position is not enough because it is possible that the detected element may not be the real erroneous element. The second strategy, GLR-IGLR parsing, can cover the errors more than the first strategy; however, some defective points are still overlooked because it has no process for handling the parsed errors before the detected point. The last strategy, GSS backtracking parsing is more effective than others but it also needs more complex implementation and spends longer processing time. However, by limiting the scope of error editing appropriately, this method will be an effective parsing and uses less time. Consequently, this method is the suitable parsing for handling ill-formed input in a real work.

Aside from the ill-formed input parsing module, we include the method for collecting the statistics of deviated data which can guide us to adapt the current grammar to attain better performance.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา - วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ประกาศิต กยะสิทธิ์

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา

ปีการศึกษา 2538 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

หอสมุดกลาง และสถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณท่านอาจารย์ที่ปรึกษา อาจารย์ ดร. บุญเสริม กิจศิริกุล และท่านอาจารย์ที่ปรึกษาร่วม อาจารย์ ดร. สุรพันธ์ เมฆนาวิน ที่คอยผลักดัน และให้คำปรึกษาในการทำวิทยานิพนธ์ฉบับนี้จนสำเร็จ ขอขอบคุณ อาจารย์ วิรัช ศรเลิศล้ำวาณิช หัวหน้าห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติที่ให้ความอนุเคราะห์ในการใช้อุปกรณ์ และโปรแกรมคอมพิวเตอร์ที่มีส่วนช่วยในการทำงานเป็นอย่างมาก

นอกจากนี้ขอขอบคุณ คุณปกรณ์ ชุนหสวัสดิกุล และผู้ร่วมงานในห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ทุกท่านที่ให้ความช่วยเหลือในด้านต่างๆ จนวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี ขอขอบคุณ โครงการ พ.ส.ว.ท ที่ให้ทุนการศึกษากับผู้วิจัยจนสำเร็จการศึกษา

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณ บิดา มารดา ซึ่งเป็นผู้มีพระคุณสูงสุดหาที่เปรียบมิได้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

หน้า

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

ตารางที่ หน้า

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

ภาพที่ หน้า

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย