

การแจ่งส่วนประโยชน์นิยายกรณในภาษาไทยด้วยตัวแจ่งส่วนแบบแอลอาร์



นาย ประกาศิต กายะสิทธิ์

ศูนย์วิทยพักร

จุฬาลงกรณ์มหาวิทยาลัย

วิทยานพณ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2538

ISBN 974-632-650-3

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

I 166593 45

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

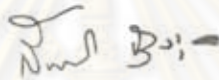
1995

ISBN 974-632-650-3



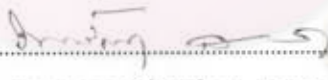
หัวข้อวิทยานิพนธ์ การแจ่งส่วนประโยคคิดไวยากรณ์ในภาษาไทยด้วยตัวแจ่งส่วนแบบแอลอาร์
โดย นาย ประภาสิต กายะสิทธิ์
ภาควิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา อาจารย์ ดร. บุญเสริม กิจศิริกุล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยรับนี้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาโทมหาบัณฑิต

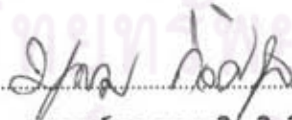


..... คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร. สันติ งามสุวรรณ)

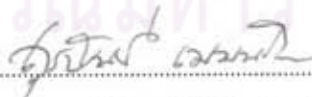
คณะกรรมการสอบวิทยานิพนธ์



..... ประธานกรรมการ
(ศาสตราจารย์ ทักษิณา สวานานนท์)



..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร. บุญเสริม กิจศิริกุล)



..... อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร. สุรพันธ์ เมฆนาวิน)



..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว



ประกาศิต ภาวะสิทธิ์ : การแจกส่วนประโยคผิดไวยากรณ์ในภาษาไทยด้วยตัวแจกส่วนแบบแอลอาร์ (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 โดยการปรับปรุงตัวแจกส่วนแบบจีแอลอาร์ใน 3 รูปแบบด้วยกัน คือ 1. การแก้ไขการผิดไวยากรณ์ ณ จุดที่ตรวจพบการผิดไวยากรณ์เท่านั้น 2. การแก้ไขการผิดไวยากรณ์ด้วยการแจกส่วนแบบจีแอลอาร์-ไอจีแอลอาร์ และ 3. การแก้ไขการผิดไวยากรณ์ด้วยการแจกส่วนแบบย้อนรอยของซ้อนโครงสร้างกราฟ

ผลการวิจัยสามารถสรุปได้ว่า วิธีแรก การแก้ไขการผิดไวยากรณ์ ณ จุดที่ตรวจพบการผิดไวยากรณ์เท่านั้น ยังไม่เพียงพอกับการใช้งานจริง มีข้อผิดพลาดในประโยคหลายจุดที่ถูกมองข้ามไป ส่วนวิธีที่สอง การแจกส่วนแบบจีแอลอาร์-ไอจีแอลอาร์ ก็สามารถครอบคลุมข้อผิดพลาดได้มากขึ้น แต่เนื่องจากไม่มีการย้อนรอยกลับมาแก้ไขประโยคในส่วนที่ทำการแจกส่วนไปแล้ว ดังนั้นข้อผิดพลาดในกรณีที่ตัวแจกส่วนตรวจพบการผิดไวยากรณ์ภายหลังจากที่ได้แจกส่วนผ่านจุดที่เป็นสาเหตุแท้จริงของการผิดไวยากรณ์ไปแล้ว จะถูกมองข้ามไปไม่ทำการแก้ไข

วิธีสุดท้าย การแก้ไขการผิดไวยากรณ์ด้วยการแจกส่วนแบบย้อนรอยของซ้อนโครงสร้างกราฟ สามารถครอบคลุมการแก้ไขข้อผิดพลาดได้ดีกว่าสองวิธีแรก แต่ก็ใช้เวลาในการประมวลผลมากกว่า อย่างไรก็ตามการจำกัดขอบเขตของการแก้ไขปัญหาให้เหมาะสมกับการใช้งาน จะทำให้วิธีนี้เป็นวิธีที่แก้ไขข้อผิดพลาดได้ดี และใช้เวลาในการประมวลผลไม่มากนัก จึงเป็นวิธีที่เหมาะสมกับการใช้งานจริง

นอกเหนือจากการปรับปรุงตัวแจกส่วนเพื่อให้สามารถแจกส่วนประโยคที่ผิดรูปแบบได้แล้ว ผลลัพธ์ที่ได้จากการแก้ไขประโยค ยังสามารถนำกลับมาใช้เก็บเป็นค่าสถิติเพื่อใช้ในการปรับปรุงพจนานุกรมชนิดคำของคำศัพท์ต่างๆ และใช้ในการปรับไวยากรณ์ที่มีอยู่ให้เหมาะสมกับการใช้งานแต่ละอย่างด้วย

ภาควิชา..... วิศวกรรมคอมพิวเตอร์
สาขาวิชา..... วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา..... 2538

ลายมือชื่อนิติบัตร ๒๖:๓๕๓๓ ภาวะสิทธิ์
ลายมือชื่ออาจารย์ที่ปรึกษา Jha reat
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

จุฬาลงกรณ์มหาวิทยาลัย

C517968 : MAJOR COMPUTER SCIENCE

KEYWORD: 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.



ศูนย์วิจัยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา - วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2538

ลายมือชื่อนิสิต ธวัชศักดิ์ ทยะสิทธิ์
ลายมือชื่ออาจารย์ที่ปรึกษา ดร. กิ่งกมล
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม ดร. ธีรพงศ์ วัฒนศิริ



กิตติกรรมประกาศ

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

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

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณ บิดา มารดา ซึ่งเป็นผู้มีพระคุณสูงสุดหาที่เปรียบมิได้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย



สารบัญ

หน้า

บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ข
สารบัญตาราง.....	ฅ
สารบัญภาพ.....	ญ

บทที่

1. บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 ขอบเขตการทำวิจัย.....	3
2. ไวยากรณ์ไม่พึงบริบทและการแจกส่วนประโยค.....	5
2.1 ไวยากรณ์ไม่พึงบริบท.....	5
2.2 การแจกส่วนประโยค.....	6
3. ตัวแจกส่วนแบบจีแอลอาร์.....	9
3.1 กองซ้อนโครงสร้างกราฟ.....	9
3.2 ตารางแอลอาร์.....	11
4. การตรวจพบการผิดไวยากรณ์.....	22
4.1 การตรวจพบการผิดไวยากรณ์.....	22
4.2 การจำแนกประเภทของข้อผิดพลาดที่ตรวจพบ.....	22
4.3 การแก้ไขการผิดไวยากรณ์ที่เกิดขึ้น.....	23
5. การแก้ไขการผิดไวยากรณ์ ณ ตำแหน่งที่ตรวจพบข้อผิดพลาดเท่านั้น.....	26
6. การแก้ไขการผิดไวยากรณ์โดยวิธีการแจกส่วนแบบจีแอลอาร์-ไอจีแอลอาร์.....	30
7. การแก้ไขการผิดไวยากรณ์โดยวิธีการแจกส่วนแบบย้อนรอยกองซ้อนโครงสร้างกราฟ.....	35
7.1 การแก้ไขการผิดไวยากรณ์โดยวิธีการแจกส่วนแบบย้อนรอยกองซ้อนโครงสร้างกราฟ.....	35
7.2 ข้อควรระวังในการพัฒนาโปรแกรม.....	41
8. การจำกัดขอบเขตของการแก้ไขปัญหา.....	43
9. การเลือกประโยคที่เหมาะสมจากผลลัพธ์ที่ได้จากการแก้ไขประโยค.....	46

10. การเก็บค่าสถิติการเบี่ยงเบนทางไวยากรณ์.....	48
11. ผลการทดลองการแก้ไขประโยคผิดไวยากรณ์.....	51
12. บทสรุป.....	58
เอกสารอ้างอิง.....	61
ภาคผนวก	
ภาคผนวก ก. ผลการทดลองการแก้ไขการผิดไวยากรณ์.....	64
ภาคผนวก ข. คำศัพท์และชนิดคำของคำศัพท์ที่ใช้.....	74
ภาคผนวก ค. คำศัพท์ที่ใช้.....	77
ประวัติผู้เขียน.....	79



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

ตารางที่	หน้า
3.1. ตารางแอลอาร์สำหรับไวยากรณ์ไม่ทิ้งบริบท จากรูปที่ 3.5	12
5.1 ตารางแอลอาร์ที่ใช้ในการแจกส่วนประโยค (อ้างอิงจากตารางที่ 3.1).....	27
6.1 ตารางแอลอาร์ตารางที่ 1 ที่ได้จากการแปลงกฎการผลิตในรูปที่ 6.1.....	30
6.2 ตารางแอลอาร์ตารางที่ 2 ที่ได้จากการแปลงกฎการผลิตในรูปที่ 6.2.....	31
7.1 ตารางแอลอาร์ที่ใช้ในการแจกส่วนประโยค.....	37
10.1 แสดงค่าสถิติที่ได้จากการแก้ไขประโยค “ วาง หนังสือ สีแดง โต๊ะ สีดำ ”.....	49
10.2 แสดงค่าสถิติที่ได้จากการแก้ไขประโยค “ Study makes hope ”.....	50
10.3 แสดงค่าสถิติของชนิดคำใหม่ที่พบจากการแก้ไขประโยค “ Study makes hope ”.....	50
11.1 แสดงผลการทดสอบความถูกต้องในการแก้ไขประโยค.....	54
11.2 การทดสอบเวลาที่ใช้ในการประมวลผล (ไวยากรณ์ชุดที่ 1).....	55
11.3 การทดสอบเวลาที่ใช้ในการประมวลผล (ไวยากรณ์ชุดที่ 2).....	56

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

ภาพที่	หน้า
1.1 โครงสร้างของระบบประมวลผลภาษาธรรมชาติ.....	2
2.1 ไวยากรณ์ไม่พื้งบริบทแบบง่ายสำหรับภาษาอังกฤษ.....	5
2.2 แสดงโครงสร้างต้นไม้ของประโยค " ฉัน เห็น เขา ".....	6
2.3 แสดงขั้นตอนการวิเคราะห์ประโยค " ฉัน เห็น เขา ".....	7
2.4 การแจงส่วนประโยคจากบนลงล่าง (Top-down parsing).....	8
2.5 การแจงส่วนประโยคจากล่างขึ้นบน (Bottom-up parsing).....	8
3.1 แสดงตัวอย่างข้อมูลในกองซ้อนที่มี A เป็นข้อมูลล่างสุด และ E เป็นข้อมูลบนสุด.....	10
3.2 แสดงการแยกส่วนที่ส่วปลายของกองซ้อน เพื่อทำการลดทอนกฎทั้ง 3 แบบ.....	10
3.3 แสดงการรวมกลุ่มค่าปลายของกองซ้อนทั้ง 3 สาย.....	11
3.4 การจัดความกำกวมของข้อมูลในกองซ้อนที่ค่าปลาย F1 กับ G1.....	11
3.5 ไวยากรณ์ไม่พื้งบริบทแบบง่ายสำหรับภาษาไทย.....	12
3.6 แสดงความหมายของสัญลักษณ์ที่ใช้ในการอธิบายขั้นตอนการแจงส่วนประโยค.....	13
3.7 สถานะพร้อม 3 คำที่เข้าตรวจสอบคือคำบุพบท "ใน" จากตารางแอลอาร์ การกระทำที่ได้คือ re3.....	14
3.8 แสดงขั้นตอนการทำงานหลังจากดึงคำ GSS(3) ออกจากกองซ้อนแล้ว.....	14
3.9 สถานะพร้อมคือ 0 คำที่ตรวจสอบคือคำนาม (n) จากตารางแอลอาร์ ได้การกระทำ sh3.....	15
3.10 ผลที่ได้หลังการทำงานตามขั้นตอน U_0	15
3.11 แสดงการแจงส่วนประโยค " ฉัน เห็น เขา ใน สวน " (ขั้นตอนที่ 1-14).....	20
3.12 ก) แสดงต้นไม้แจงส่วนในแบบที่ 1 จากการกระจายกองซ้อนโครงสร้างกราฟ.....	20
3.12 ข) แสดงต้นไม้แจงส่วนในแบบ 2 จากการกระจายกองซ้อนโครงสร้างกราฟ.....	21
4.1 การตรวจพบข้อผิดพลาดที่คำนาม "สวน" ขณะที่สถานะพร้อมคือสถานะที่ 3.....	22
4.2 แสดงการแยกสายโครงสร้างในขั้นตอนที่ 5 เพื่อแทรกชนิดคำที่เหมาะสมลงในประโยค.....	24
5.1 แสดงตำแหน่งของข้อผิดพลาดจริงกับข้อผิดพลาดที่ตรวจพบ.....	26
5.2 แสดงขั้นตอนการแจงส่วนประโยคและมีการตรวจพบข้อผิดพลาดที่คำ "ดี".....	27
5.3 แสดงการแยกสายโครงสร้างข้อมูลในขั้นตอนที่ (5) หลังพบข้อผิดพลาดที่ "ดี".....	28
5.4 แสดงการแทนที่คำบุพบท [p] ที่คำ "ดี" และการทำงานต่อจนจบประโยค.....	29
6.1 แสดงกฎการผลิตของไวยากรณ์ไม่พื้งบริบทอย่างง่ายในภาษาไทย.....	30
6.2 แสดงกฎการผลิตที่ได้จากการกลับไวยากรณ์ไม่พื้งบริบทในรูปที่ 6.1.....	31

6.3 ก) ขั้นตอนแรกการแจงส่วนแบบจีแอลอาร์ ตรวจสอบข้อผิดพลาดและทำการแก้ไขที่ "ดี".....	32
ข) ขั้นตอนที่ 2 การแจงส่วนแบบไอจีแอลอาร์ ตรวจสอบข้อผิดพลาดและทำการแก้ไขที่ "กั๊ด".....	32
6.4 แสดงขั้นตอนการแจงส่วนประโยคโดยวิธีไอจีแอลอาร์ ข้อผิดพลาดถูกตรวจพบที่คำ "กั๊ด".....	32
6.5 แสดงการแยกสายโครงสร้างข้อมูลในขั้นตอนที่ (6) หลังพบข้อผิดพลาดที่ "กั๊ด".....	33
6.6 แสดงการแทนที่คำบุพบท [p] ที่คำ "กั๊ด" และการทำงานต่อจนจบประโยค.....	34
7.1 ก) แสดงองค์ประกอบต่างๆ ที่มีในกองซ้อนขณะที่มีการตรวจพบข้อผิดพลาดที่ "ดี".....	36
ข) คำที่เป็นองค์ประกอบอิสระทั้ง 3 จุด แต่ละจุดจะถูกแยกออกเพื่อแก้ไขประโยค.....	36
7.2 แสดงขั้นตอนการทำงานและการตรวจพบข้อผิดพลาดที่ "ดี".....	37
7.3 แสดงการแยกโครงสร้างข้อมูลในขั้นตอนที่ (5) เพื่อแก้ไขประโยคที่คำ "ดี".....	38
7.4 แสดงการแก้ไขประโยคในสายที่ 3 ของการแก้ไขประโยคที่ "ดี".....	38
7.5 แสดงการแยกโครงสร้างข้อมูลในขั้นตอนที่ (2) เพื่อแก้ไขประโยคที่ "กั๊ด".....	39
7.6 แสดงการแก้ไขประโยคในโครงสร้างสายที่ 3 ของตำแหน่งคำ "กั๊ด".....	40
7.7 แสดงตัวอย่างการตรวจพบข้อผิดพลาดจุดที่ 2 ในขั้นตอนที่ (10) จากรูปที่ 7.4.....	41
7.8 แสดงตัวอย่างองค์ประกอบอิสระในแต่ละสายโครงสร้างข้อมูล.....	42
8.1 แสดงจำนวนโครงสร้างข้อมูล T_m สายที่ถูกแยกออกมาเมื่อมีการแก้ไขประโยค.....	43
8.2 การแยกสายข้อมูลของการแก้ไขประโยคแบบย้อนรอยกองซ้อนโครงสร้างกราฟ.....	44
9.1 แสดงไวยากรณ์ PCFG ของกฎการผลิตที่ใช้.....	46
9.2 แสดงโครงสร้างต้นไม้ของประโยค " I saw a man ".....	47
10.1 แสดงไวยากรณ์ไม่พึ่งบริบทอย่างง่ายของภาษาไทยที่ใช้ในการอธิบาย.....	48
10.2 แสดงโครงสร้างต้นไม้ของประโยคตัวอย่างที่กำหนด.....	48
10.3 แสดงโครงสร้างต้นไม้ของประโยค " [n] makes [n] ".....	50
11.1 การเปรียบเทียบผลการทดสอบความถูกต้องของการแก้ไขประโยค.....	55
11.2 กราฟแสดงความสัมพันธ์ของเวลาที่ใช้ในการแก้ไขประโยคเป็นร้อยละของเวลาปกติ.....	56
11.3 กราฟแสดงความสัมพันธ์ของเวลาที่ใช้ในการแก้ไขประโยคเป็นร้อยละของเวลาปกติ.....	57
12.1 แสดงภาพโดยรวมของงานวิจัยที่ทำ.....	58