

บทที่ 1

บทนำ

เอกสารนี้จัดทำขึ้นเพื่อใช้ประกอบวิทยานิพนธ์เรื่อง “การคัดแยกข้อความที่มีความหมายจากโครงสร้างต้นไม้ของไวยากรณ์” โดยมีรายละเอียดของที่มาของวิทยานิพนธ์ วัตถุประสงค์ ขอบเขต และข้อมูลอื่น ๆ ที่เกี่ยวข้องดังจะกล่าวในหัวข้อต่าง ๆ ต่อไปนี้

1.1 หลักการ ทฤษฎี เหตุผล และ/หรือสมมติฐาน

การคัดแยกข้อความที่ต้องการจากภาษาธรรมชาติ (Natural Language) เป็นงานวิจัยหนึ่งที่น่าสนใจในระบบการคัดแยกข้อมูล (Information Extraction) ซึ่งระบบการคัดแยกข้อมูลนั้นจำเป็นต้องมีความเข้าใจเกี่ยวกับภาษาธรรมชาตินั้น ๆ ดังนั้นจึงมีหลากหลายวิธีที่นำมาใช้ในกระบวนการของการวิเคราะห์บทความ ซึ่งมีความเกี่ยวข้องกับการระบุส่วนต่าง ๆ ของบทความที่มีความสัมพันธ์กัน การทำงานกับภาษาธรรมชาตินั้นนำมาซึ่งปัญหาเกี่ยวกับการวิเคราะห์ภาษา โดยโครงสร้างของประโยคที่มีความกำกวมจะมีผลต่อความถูกต้องของการคัดแยกข้อความที่ต้องการ โดยการคัดแยกข้อความแบบเดิมจะใช้วิธีการทำแพตเทิร์นแมชชีน (pattern matching) ซึ่งเป็นการค้นหาข้อความจากคำสำคัญและถือว่าข้อความทั้งหมดที่อยู่หน้าและหลังคำสำคัญนั้นเป็นประธานและกรรมของประโยค ข้อความที่ได้จากวิธีการข้างต้นนี้อาจทำให้ได้ข้อความที่เป็นส่วนของคำขยายที่ไม่จำเป็นต้องมาด้วย เช่น ประโยคที่มีลักษณะเป็นประโยคความซ้อนที่ีอนุประโยคทำหน้าที่เป็นส่วนขยายของประธานและกรรมของประโยค ดังนั้นผลจากการคัดแยกประธานและกรรมที่มีความสัมพันธ์กับคำสำคัญจึงได้ส่วนขยายที่เป็นอนุประโยคซึ่งอาจยาวมากติดมาด้วย จึงได้มีการใช้พาสเซอร์ (parser) ซึ่งเป็นส่วนหนึ่งของระบบลิงก์แกรมมาพาสเซอร์ (Link Grammar Parser) ในการวิเคราะห์โครงสร้างของประโยคก่อนที่จะทำการคัดแยกข้อความที่ต้องการ

สำหรับหลักการของระบบลิงก์แกรมมา นั้นจะทำการแบ่งประโยคตามโครงสร้างของประโยคในบทความ โดยที่การเรียงลำดับของคำในประโยคนั้นจะต้องมีการเชื่อมความสัมพันธ์กันระหว่างคำสองคำ ซึ่งเราจะเรียกว่า “ลิงค์” ลิงค์ของคำแต่ละคู่จะต้องไม่มีการไขว้กัน และความสัมพันธ์ของคำในประโยคจะมีลักษณะเป็นกราฟที่มีการเชื่อมกัน (connected graph) เราสามารถอธิบายลิงก์แกรมมาได้โดยให้คำแต่ละคำในประโยคมีการเชื่อมต่อกันโดยลิงค์ที่มีหลากหลายชนิดตามหน้าที่ของคำในประโยคนั้น ลิงค์เกิดจากคอนเนคเตอร์ (connectors) ชนิดเดียวกันของคำสองคำที่มีการเชื่อมกัน คอนเนคเตอร์จะมีการระบุทิศทาง ซึ่งอาจชี้ไปทางซ้ายหรือทางขวาของคำ โดยคอนเนคเตอร์ที่ชี้ไปทางซ้ายจะต้องเชื่อมต่อกับคอนเนคเตอร์ ชนิดเดียวกันของคำอีกคำ

หนึ่งที่จะไปทางขวา เพื่อทำหน้าที่ในการแสดงความสัมพันธ์ของคำแต่ละคำในประโยคตาม ความสัมพันธ์ที่มีต่อกัน คอนเนคเตอร์ที่จะไปทางขวาจะแสดงด้วยเครื่องหมาย “+” และคอนเนค-เตอร์ที่จะไปทางซ้ายจะแสดงด้วยเครื่องหมาย “-” ลิงก์แกรมมาจะประกอบไปด้วยเซตของคำและ คอนเนคเตอร์ของแต่ละคำ โดยจะมีดิคชันนารีทำหน้าที่ในการเก็บรวบรวมคำและคอนเนคเตอร์ ของแต่ละคำไว้ การเรียงลำดับของคำแต่ละคำในประโยคนั้นจะถูกกำหนดโดยหลักไวยากรณ์ของ ภาษา ในการเชื่อมต่อกันของคำแต่ละคำนั้นก็จะต้องเป็นไปตามกฎที่กำหนด ประโยคที่ถูกต้องก็จะ เกิดจากการเชื่อมโยงกันอย่างถูกต้องของคำในประโยค

สำหรับพาสเซอร์ (parser) นั้นจะเป็นส่วนประกอบหนึ่งของลิงก์แกรมมาพาสเซอร์ โดย พาสเซอร์จะนำลิงก์ที่ได้จากลิงก์แกรมมาใช้ในการพาสเซอร์เพื่อแยกแต่ละประโยคให้อยู่ในรูป โครงสร้างของประโยคที่ประกอบไปด้วยวลีต่าง ๆ เมื่อลิงก์ในประโยคถูกสร้างขึ้นเรียบร้อยแล้ว จะเข้าสู่กระบวนการในการพาสเซอร์ต่อไป โดยทำการแบ่งประโยคออกเป็นโดเมนชนิดต่าง ๆ ตาม ชนิดของลิงก์ที่แสดงในประโยค หรืออาจกล่าวได้ว่าโดเมนนั้นเป็นสับเซตของลิงก์ที่ประกอบกัน เป็นประโยค กลุ่มของคำที่ได้จากการแบ่งตามโดเมนนั้นจะมีความสัมพันธ์กับองค์ประกอบของ โครงสร้างของประโยคที่เป็นวลีต่าง ๆ ทำให้สามารถสร้างวลีเหล่านั้นได้ตามกฎที่กำหนด

ดังนั้นงานวิจัยนี้จึงได้นำเสนอการคัดแยกข้อความจากประโยคในภาษาอังกฤษ โดยนำ หลักการของลิงก์แกรมมาพาสเซอร์มาใช้ในการวิเคราะห์โครงสร้างของประโยคซึ่งอาจประกอบไป ด้วย นามวลี บุพบทวลี กริยาวลี เป็นต้น ซึ่งองค์ประกอบต่าง ๆ ของประโยคนั้นจะได้จากลิงก์ของ ลิงก์แกรมมาพาสเซอร์และองค์ประกอบของประโยคที่จะถูกจัดให้อยู่ในรูปของโครงสร้างแบบ ต้นไม้ เราจึงสามารถจะท่องไปในโครงสร้างของประโยคที่มีลักษณะเป็นโครงสร้างต้นไม้เพื่อทำ การคัดแยกข้อความที่ต้องการได้ โดยข้อความที่คัดแยกได้นั้นจะต้องเป็นนามวลีที่มีความสัมพันธ์ กับคำสำคัญและคำสำคัญนั้นจะต้องเป็นคำกริยาในกริยาวลี

1.2 วัตถุประสงค์ของการศึกษา

เพื่อพัฒนาระบบการคัดแยกข้อความจากบทความภาษาอังกฤษให้มีความหมายสมบูรณ์มากขึ้น โดยการวิเคราะห์โครงสร้างของประโยคด้วยหลักการลิงก์แกรมมาพาสเซอร์

1.3 ประโยชน์ที่ได้รับจากการศึกษา

1.3.1. ทำให้ได้กระบวนการในการคัดแยกข้อความจากบทความภาษาอังกฤษให้มีความหมาย สมบูรณ์มากขึ้นได้

1.3.2. ทำให้เกิดการประยุกต์ใช้โครงสร้างต้นไม้กับกระบวนการในการคัดแยกข้อความได้

1.4 แผนการดำเนินการ ขอบเขตและวิธีการวิจัย

1.4.1 แผนการดำเนินการ

- (1) ศึกษาหลักการของระบบการคัดแยกข้อความ
- (2) ศึกษาหลักการทำงานของลิงค์แกรมมาพาสเซอร์
- (3) จัดทำระบบการคัดแยกข้อความ
- (4) ทดสอบ ปรับปรุงระบบและประเมินผล
- (5) จัดทำเอกสารประกอบ

1.4.2 ขอบเขต

- (1) คัดแยกข้อความจากบทความภาษาอังกฤษให้มีความหมายสมบูรณ์ โดยการวิเคราะห์โครงสร้างของประโยคด้วยหลักการลิงค์แกรมมาพาสเซอร์
- (2) ประเมินผลด้วยค่ารีคอลและค่าพรีซิชัน ซึ่งเป็นค่ามาตรฐานที่ใช้ในการประเมินระบบการคัดแยกข้อความ

1.4.3 วิธีการวิจัย

- (1) ศึกษาหลักการของระบบการคัดแยกข้อความ
- (2) ศึกษาหลักการทำงานของลิงค์แกรมมาพาสเซอร์ เพื่อใช้ในการวิเคราะห์โครงสร้างของประโยค
- (3) จัดทำระบบการคัดแยกข้อความ
- (4) ทดสอบ ปรับปรุงระบบและประเมินผลด้วยค่ารีคอลและค่าพรีซิชัน
- (5) จัดทำเอกสารประกอบ

1.5 สถานที่ในการดำเนินการวิจัยและรวบรวมข้อมูล

1.5.1 สถานที่

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

1.5.2 อุปกรณ์

- (1) ฮาร์ดแวร์
เครื่องคอมพิวเตอร์ส่วนบุคคล
- (2) ซอฟต์แวร์
 - โปรแกรมระบบปฏิบัติการไมโครซอฟต์วินโดวส์ 2000 เซิร์ฟเวอร์ (Microsoft Windows 2000 Server)
 - โปรแกรมภาษาซี พลัส พลัส บิวเดอร์ (C++ Builder)

- โปรแกรมภาษาซี (C Language)

1.6 ระยะเวลาในการดำเนินการวิจัย

ตั้งแต่เดือน มกราคม ถึงเดือน กันยายน พ.ศ. 2548 เป็นระยะเวลา 9 เดือน



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright © by Chiang Mai University
All rights reserved