บทที่ 1

บทนำ

ความเป็นมาและปัญหา

การเขียนประโยคภาษาไทยนั้นเราจะเขียนคำต่าง ๆ วางเรียงติดต่อกันไปจนจบประโยค จึงจะได้ถ้อยความเป็นที่เข้าใจตามที่ต้องการจะสื่อความหมาย เช่น ฉันไปโรงเรียน ซึ่งแตกต่างจากการเขียนประโยคภาษาอังกฤษที่คำแต่ละคำจะแบ่งแยกด้วยช่องว่างอย่างชัดเจน เช่น I go to school. ซึ่งลักษณะของการเขียนประโยคภาษาไทยนั้นทำให้การประมวลผลภาษาไทยด้วยคอมพิวเตอร์ทั้งในแง่ของ โปรแกรมประมวลผลคำภาษาไทย (Thai Word Processor) การตรวจสอบตัวสะกดภาษาไทย (Spelling Check) การวิเคราะห์ข้อความทางภาษา (Language Parsing) เช่นระบบแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) ฯลฯ เป็นไปด้วยความลำบาก เนื่องจากการแยกคำในการจัดรูปแบบของโปรแกรมประมวลผลคำภาษาไทยที่มีอยู่นั้นจะแยกคำโดยคำนึงถึงความกว้างของหน้ากระดาษและจำนวนคำเท่านั้น ไม่สามารถแยกคำให้ถูกต้องตามหลักภาษาศาสตร์ได้ ซึ่งจะมีผลให้ข้อความไม่ต่อเนื่องและความหมายอาจผิดไปได้

ปัจจุบันเทคโนโลยีทางคอมพิวเตอร์พัฒนาไปอย่างรวดเร็วมาก นักวิจัยนำเอาเทคนิคทางปัญญาประดิษฐ์มาสนับสนุนพัฒนางานประมวลผลภาษาธรรมชาติไทย เพื่อแก้ปัญหาด้านภาษาศาสตร์ นับแต่การประมวลผลคำภาษาไทย (Thai Word Processor) ในเรื่องการตัดคำ (Thai Word Separator) เพื่อช่วยในการปรับแต่งรูปแบบข้อความ เช่นการจัดขอบซ้ายขวาของเอกสารนั้น ข้อความที่ปรากฏจะต้องมีการแยกคำในตำแหน่งที่เหมาะสมโดยที่ไม่ทำให้ความหมายของข้อความผิดไป การตัดคำจะใช้หลักการพื้นฐานทางภาษาศาสตร์ ซึ่งก็คือเลือกผลของการจำแนกแยกแยะคำที่ให้จำนวน 'คำ' มากที่สุดที่จะทำให้ได้คำศัพท์ที่เป็นพื้นฐานที่สุด ขั้นตอนนี้จะอาศัยกฎเกณฑ์หลักภาษาไทยเข้ามาวิเคราะห์ว่าข้อความที่ถูกตัดแบ่งมาแล้วสอดคล้อง หรือว่าขัดแย้งกับหลักภาษาไทยหรือไม่ ซึ่งการนี้จะต้องใช้พจนานุกรมที่รวบรวมไวยากรณ์มาสนับสนุน การตรวจสอบตัวสะกดจะต้องสามารถแยกคำศัพท์ที่ปรากฏในเอกสาร และนำคำศัพท์ดังกล่าวลงไปสืบค้นในพจนานุกรมที่รวบรวม

คำศัพท์และตรวจสอบว่าสะกดถูกต้องหรือไม่ ตลอดจนการตรวจสอบการใช้ภาษาในระดับคำ (Style check at the word level) การตรวจสอบไวยากรณ์ (Syntactic Analysis) การวิเคราะห์เชิงความหมาย (Semantic Analysis) จะต้องอาศัยพจนานุกรมทั้งสิ้น ระบบแปลภาษาด้วยเครื่องคอมพิวเตอร์ โดยใช้ภาษาต้นแบบเป็นภาษาไทยนั้นประโยคที่พิมพ์เข้ามาจะถูกวิเคราะห์เพื่อแยกแยะตัดทอนให้ได้คำออกมาก่อนที่จะทำการวิเคราะห์ความหมายของคำเหล่านั้น นอกจากนี้ยังพัฒนาพจนานุกรมคำพ้อง (Thesaurus) ที่รวบรวมศัพท์ที่มีความหมายเหมือนกันเพื่อใช้สำหรับการวิเคราะห์เชิงความหมายคำศัพท์ เพื่อสนับสนุนงานประมวลผลภาษาไทยด้วยคอมพิวเตอร์ด้านอื่น ๆ จะเห็นว่าพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยมีบทบาทในงานพัฒนาโปรแกรมประมวลภาษาไทยด้วยคอมพิวเตอร์มากมายหลายด้าน

วัตถุประสงค์ของวิทยานิพนธ์

จากที่กล่าวมาแล้วข้างต้น วิทยานิพนธ์ฉบับนี้จึงได้กำหนดวัตถุประสงค์หลัก คือ เพื่อศึกษาและพัฒนารูปแบบโครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่เหมาะสม เพื่อใช้ในการวิจัยและพัฒนาการประมวลผลภาษาธรรมชาติไทย โดยใช้พจนานุกรมสนับสนุนการวิจัยและพัฒนาดังกล่าวทั้งนี้มีองค์ประกอบในการพิจารณาดังนี้

1. เสนอรูปแบบโครงสร้างข้อมูลสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่ใช้เนื้อที่ในการเก็บอย่างมีประสิทธิภาพ โดยมีข้อสมมุติฐานว่าพจนานุกรมนั้นมีขนาดใหญ่ที่ต้องใช้จานแม่เหล็กช่วยในการเก็บ

2. พัฒนาอัลกอริทึมที่ใช้บันทึกคำศัพท์ลงในพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่รวดเร็ว และที่ผู้ใช้สามารถเพิ่มเติมรายละเอียดพจนานุกรมไทยและอัลกอริทึมสำหรับการสืบค้นคำศัพท์ อัลกอริทึมทั้งสองนี้จะเป็นประโยชน์ในงานประมวลผลภาษาไทยด้วยเครื่องคอมพิวเตอร์

ขอบเขตและเงื่อนไขของวิทยานิพนธ์

1. การออกแบบโครงสร้างข้อมูลสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยจะใช้โครงสร้างแบบต้นไม้ที่มีการสืบค้นแบบดิจิตอลประกอบกับแถวลำดับคู่ โดยจะนำแนวทางมาจากการศึกษาและพิจารณาข้อดีข้อด้อยของทฤษฎีต่าง ๆ ที่ใช้ในงานวิจัยที่เกี่ยวข้องกับการออกแบบโครงสร้างข้อมูลการเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่ผ่านมา

2. การออกแบบโครงสร้างข้อมูลและการพัฒนาโปรแกรมเพื่อบันทึกคำศัพท์ดังกล่าวจะอยู่บนพื้นฐานของสมมุติฐานที่ว่ามีพจนานุกรมที่สมบูรณ์อยู่แล้ว

3. การพัฒนาโปรแกรม จะใช้ภาษาระดับสูง (high level language) ที่มีความยืดหยุ่นสูง และพัฒนาโปรแกรมบนเครื่องไมโครคอมพิวเตอร์

4. รหัสภาษาไทยที่ใช้ จะใช้รหัสของสำนักงานมาตรฐานอุตสาหกรรม (ส.ม.อ.)

5. การพิจารณาประสิทธิภาพจะเปรียบเทียบจากโครงสร้างข้อมูลที่มีอยู่ในปัจจุบัน โดยมีองค์ประกอบดังนี้

5.1 ความรวดเร็วในการค้นหาคำพจนานุกรม

5.2 เนื้อที่ที่ใช้ในการเก็บข้อมูล

<u>ขั้นตอนการวิจัย</u>

1. ศึกษาการออกแบบโครงสร้างข้อมูลสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยจากงานวิจัยต่าง ๆ โดยละเอียด

2. ออกแบบโครงสร้างข้อมูลสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

3. ออกแบบและพัฒนาอัลกอริทึมเพื่อทำการบันทึกและสืบค้นคำในโครงสร้างพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่ออกแบบไว้

4. ศึกษาทางทฤษฎีถึงประสิทธิภาพในแง่เวลาที่ใช้ในการสืบค้น และเนื้อที่ที่ใช้ในหน่วยความจำของโครงสร้างข้อมูลและอัลกอริทึมที่ออกแบบมา

5. ถ้าเป็นไปได้ทำการทดลองเปรียบเทียบประสิทธิภาพกับโครงสร้างข้อมูลงานวิจัยที่กล่าวไว้ข้างต้น

6. สรุปผลการวิจัยและข้อเสนอแนะ

<u>ประโยชน์ที่คาดว่าจะได้รับ</u>

1. ได้โครงสร้างข้อมูลสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่เหมาะสมและมีประสิทธิภาพอันได้แก่

1.1 สามารถสืบค้นข้อมูลอย่างรวดเร็ว

1.2 ประหยัดเนื้อที่ในหน่วยความจำและหน่วยความจำสำรองของคอมพิวเตอร์

2. ใช้เป็นแนวทางในการพัฒนาโปรแกรมประมวลผลคำภาษาไทย ให้มีประสิทธิภาพมากขึ้น ในแง่ของ

2.1 การจัดรูปแบบเอกสารที่สมบูรณ์ขึ้น สามารถแยกคำที่มีความหมายที่ถูกต้องมากขึ้น

2.2 การตรวจสอบความถูกต้องของตัวสะกดที่ปรากฏในเอกสาร

2.3 การตรวจสอบความหมายระดับคำของประโยค

3. ใช้เป็นแนวทางในการวิเคราะห์ข้อความภาษาไทย ทั้งการตรวจสอบไวยากรณ์และการวิเคราะห์เชิงความหมาย ในงานการประมวลผลภาษาธรรมชาติไทย เช่น การแปลภาษาด้วยเครื่องคอมพิวเตอร์

4. ใช้เป็นแนวทางในการวิจัยเกี่ยวกับการออกแบบโครงสร้างข้อมูลสำหรับเก็บพจนานุกรมภาษาอื่น ๆ ที่มีลักษณะการเขียนประโยคคล้ายภาษาไทย

<u>โครงสร้างของวิทยานิพนธ์</u>

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 7 บท อันได้แก่

บทที่ 1 กล่าวถึงความเป็นมาของปัญหา วัตถุประสงค์ ขอบเขต และเนื้อหาของวิทยานิพนธ์

บทที่ 2 แสดงรายละเอียดพจนานุกรมอิเล็กทรอนิกส์และผลงานวิจัยเรื่องโครงสร้างที่ใช้จัดเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่ผ่านมา

บทที่ 3 กล่าวถึงแนวทางการออกแบบโครงสร้างพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

บทที่ 4 กล่าวถึงรายละเอียดโครงสร้างพจนานุกรมอิเล็กทรอนิกส์ที่ทำการออกแบบตลอดจนขั้นตอนการสืบค้นคำศัพท์และการเพิ่มคำศัพท์พร้อมตัวอย่างประกอบคำอธิบาย

บทที่ 5 กล่าวถึงรายละเอียดของการพัฒนาการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

บทที่ 6 รายงานผลการทดสอบเกี่ยวกับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

บทที่ 7 บทสรุปพร้อมทั้งคำแนะนำสำหรับการวิจัยต่อๆ ไป

นอกจากเนื้อหาดังกล่าวแล้ววิทยานิพนธ์นี้ยังประกอบด้วยภาคผนวกอีก 3 บท ดังนี้

ภาคผนวก ก. แสดงตารางรหัสภาษาไทย

ภาคผนวก ข. ตัวอย่างคำศัพท์ที่จัดในเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

ภาคผนวก ค. รายละเอียดโครงสร้างข้อมูลสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่นำใช้ทดสอบประสิทธิภาพการทำงาน