บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

จากอดีตจนถึงปัจจุบันเทคโนโลยีทางด้านคอมพิวเตอร์ได้มีการพัฒนาไปอย่างรวดเร็ว มีการประยุกต์ใช้คอมพิวเตอร์ร่วมกับเทคโนโลยีอื่นๆ มากมาย เช่น การติดต่อสื่อสาร ระบบควบคุมเครื่องจักรต่างๆ เป็นต้น ในปัจจุบันและอนาคตคอมพิวเตอร์จะถูกพัฒนาให้มีขีดความสามารถมากขึ้นเรื่อยๆ เพื่อให้สามารถทำงานต่างๆ แทนมนุษย์ได้

การโต้ตอบกับผู้ใช้ (User Interface) ระหว่างมนุษย์กับคอมพิวเตอร์ในปัจจุบันนั้นยังมีความแตกต่างจากการสื่อสารด้วยกันระหว่างมนุษย์เองค่อนข้างมาก มนุษย์ยังต้องติดต่อหรือสั่งให้คอมพิวเตอร์ทำงานด้วยคำสั่งในรูปแบบที่คอมพิวเตอร์จะสามารถเข้าใจได้ ทำให้การใช้คอมพิวเตอร์ในงานหลายอย่างถูกจำกัดให้ผู้เชี่ยวชาญในการติดต่อกับระบบใช้ได้เท่านั้น คนทั่วไปหรือคนที่ยังไม่ชินกับระบบไม่สามารถใช้งานระบบดังกล่าวได้ เพื่อให้คนที่มีความเชี่ยวชาญทางเทคนิคน้อยได้มีโอกาสเข้าถึงระบบคอมพิวเตอร์ที่ซับซ้อนได้มากขึ้น จึงเป็นการสมควรอย่างยิ่งที่จะต้องมีการนำเทคนิคการสื่อสารกับคอมพิวเตอร์แบบที่มนุษย์เข้าใจได้ง่ายมาใช้ในการสั่งงานคอมพิวเตอร์

การประมวลผลภาษาธรรมชาติ (Natural Language Processing) เป็นเทคนิคหนึ่งที่ถูกนำมาใช้งานเพื่อการสื่อสารที่ดีขึ้นระหว่างมนุษย์และคอมพิวเตอร์ การประมวลผลภาษาธรรมชาติคือกระบวนการที่ทำให้ผู้ใช้คอมพิวเตอร์สามารถนำข้อมูลเข้าเครื่องในรูปแบบภาษาพูดหรือภาษาเขียนที่ผู้ใช้ใช้สื่อสารกับคนทั่วไป คอมพิวเตอร์จะตีความข้อมูลนั้นโดยจำลองวิธีการตีความของมนุษย์ การใช้งานภาษาธรรมชาติมีตัวอย่างเช่น การสั่งงานเครื่องใช้ในบ้านผ่านระบบภาษาธรรมชาติ [1] และการสร้างหุ่นยนต์สนทนาอัตโนมัติ [2], [3] เป็นต้น

ภาษาไทยที่มีการพิมพ์ผิดจากการไม่ตั้งใจ (เช่น การพิมพ์ตก พิมพ์แทนที่ พิมพ์สลับ พิมพ์เกิน) นั้น ทำให้หุ่นยนต์สนทนาไม่สามารถโต้ตอบได้อย่างมีประสิทธิภาพ เพราะว่าคำจะไม่ตรงกับคำสำคัญในข้อมูลคำถาม-คำตอบของหุ่นยนต์สนทนา ซึ่งการที่หุ่นยนต์สนทนาไม่สามารถคาดเดาคำผิดได้เช่นเดียวกับมนุษย์นั้นอาจก่อให้เกิดความรำคาญกับผู้ใช้งาน แทนที่จะช่วยให้การติดต่อกับคอมพิวเตอร์นั้นเป็นไปตามธรรมชาติมากขึ้น

งานวิทยานิพนธ์ชิ้นนี้จึงออกแบบและพัฒนาวิธีการแก้คำผิดแบบไม่ตั้งใจโดยอัตโนมัติในภาษาไทย เพื่อนำไปใช้ในการสื่อสารกับหุ่นยนต์สนทนา (Chat Robot) โดยเน้นเฉพาะกรณีที่เกิดการพิมพ์แทนที่ พิมพ์เกินและพิมพ์สลับเท่านั้น ซึ่งจะทำให้มีความยืดหยุ่นหากสามารถป้อนอินพุต

ที่มีคำผิดพลาดโดยไม่ตั้งใจได้ โดยวิธีการแก้คำผิดนั้นอาศัยคุณสมบัติของตัวอักษรประชิดบนแผงแป้นอักขระเข้ามาช่วยลดรายการคำใกล้เคียงคำผิด

วิธีการแก้คำผิดแบบไม่ตั้งใจที่ได้จากงานวิทยานิพนธ์นี้ สามารถนำไปประยุกต์ต่อยอดใช้ให้เกิดประโยชน์ได้อย่างมาก เช่น นำไปรวมกับวิธีการแก้คำผิดที่เกิดจากสาเหตุอื่นๆ เพื่อนำไปใช้กับงานด้านโปรแกรมประมวลคำ (Word Processing) ได้

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเพิ่มความสามารถของหุ่นยนต์สนทนาในภาษาไทยโดยการแก้คำผิดแบบไม่ตั้งใจโดยอัตโนมัติ

1.3 ขอบเขตของการวิจัย

1. ในงานวิจัยนี้จะใช้วิธีการแบ่งคำในประโยคภาษาไทยของโปรแกรม SWATH

2. อ้างอิงรูปแบบมาตรฐานของ เอไอเอ็มแอล เวอร์ชัน 1.0.1 ในการสร้างฐานความรู้ให้กับหุ่นยนต์

3. การเขียนแพทเทิร์นของแต่ละแคตากอรีในฐานความรู้เอไอเอ็มแอลภาษาไทย ต้องมีการเว้นวรรคระหว่างคำในประโยคเองเหมือนรูปแบบที่ได้จากการตัดคำด้วยโปรแกรม SWATH

4. ประโยคข้อความที่รับเข้ามา มีขอบเขตดังนี้

- ข้อความอินพุตสามารถมีคำผิดได้ แต่ต้องไม่ใช่คำที่จงใจให้ผิด เช่น "สามารถ" จงใจพิมพ์ผิดเป็น "สามาด"
- คำผิดแบบไม่ตั้งใจ หมายถึง คำผิดที่เกิดจากการพิมพ์พลาดในการใช้คีย์บอร์ดเท่านั้น เช่น กดชิฟต์ไม่ติด กดปุ่มที่อยู่ข้างเคียงไปแทนที่ พิมพ์เกิน เป็นต้น
- รับอินพุตเฉพาะที่เป็นข้อความภาษาไทยคราวละ 1 ประโยค เท่านั้น
- จะต้องเป็นประโยคความเดียวเท่านั้น (ในงานวิทยานิพนธ์นี้จะไม่รองรับประโยคความรวมหรือประโยคความซ้อน)

<u>หมายเหตุ</u> ประโยคความรวมหรือประโยคความซ้อน หมายถึง ประโยคที่สามารถแยกออกเป็นประโยคย่อยได้มากกว่า 1 ประโยค

5. ในงานวิจัยนี้จะทำการแก้ไขคำผิดแบบไม่ตั้งใจโดยอัตโนมัติ ที่มีระยะแก้ไขไม่เกิน 1 ตัวอักษรใน 1 คำ โดยจะทำเฉพาะ 3 กรณีเท่านั้น คือ กรณีพิมพ์ผิดที่เกิดจากลักษณะการพิมพ์แทนที่ พิมพ์เกินและพิมพ์สลับ (ซึ่งทั้ง 3 กรณีสามารถจะมีคำผิดกี่ที่ใน 1 ประโยคก็ได้ แต่คำผิดนั้นต้องไม่อยู่ติดกัน และที่ผิดแต่ละที่ต้องอยู่ภายในแต่ละคำที่จะถูกตัด)

นอกจากนี้จะยกเว้นบางกรณีที่ไม่สามารถตรวจสอบหรือระบุขอบเขตของคำที่ผิดได้ถูกต้อง ได้แก่กรณีต่อไปนี้

- คำผิดที่เกิดจากการสลับอักขระระหว่างคำ เช่น ตัดแขน พิมพ์เป็น ดัแดขน
- คำผิดที่สามารถแบ่งได้เป็นคำย่อยที่ถูกได้ 2 คำ เช่น เที่ยงธรรม พิมพ์ผิด เป็น เที่ยวธรรม

6. ในงานวิจัยนี้จะทำการหารูปแบบการพิมพ์ผิดในภาษาไทย ว่าตำแหน่งอักขระใดในคำที่เป็นสาเหตุของการพิมพ์ผิดมากกว่ากัน (เช่น ตำแหน่งแรก ตำแหน่งกลาง หรือตำแหน่งสุดท้ายของคำ) แล้วจะนำมาใช้ปรับปรุงวิธีการแก้ไขคำผิดที่พัฒนาให้เหมาะสมกับภาษาไทย

7. วัดผลการทดลองโดยการสร้างชุดกฎคำถาม-คำตอบ มาจำนวนอย่างน้อย 100 กฎ (ซึ่งจะนำมาจากบันทึกบทสนทนาของโปรแกรม msn จากผู้ใช้จำนวนอย่างน้อย 30 คน) มาใช้ในการทดสอบผลการจับคู่แพทเทิร์น แล้วทำการวัดเมื่อมีการแก้ไขคำผิดแล้วสามารถจับคู่ได้ประสิทธิภาพดีขึ้นกว่าเดิมมากเท่าใดในรูปของเปอร์เซ็นต์

8. ใช้ภาษาจาวาในการพัฒนาวิธีการแก้คำผิดแบบไม่ตั้งใจโดยอัตโนมัติ สำหรับหุ่นยนต์สนทนาภาษาไทย

1.4 ขั้นตอนและวิธีดำเนินการวิจัย

1. สำรวจเทคนิคที่ใช้ในการทำหุ่นยนต์สนทนา (Chat Robot) ที่มีอยู่ในปัจจุบัน
2. ศึกษาและเลือกวิธีการตัดคำภาษาไทยที่เหมาะสมในการนำมาใช้กับเอไอเอ็มแอล
3. ออกแบบและพัฒนาการแก้ไขคำผิดอัตโนมัติ โดยจะรับอินพุตมาจากผลที่ได้ของโปรแกรมตัดคำ
4. ศึกษาและสร้างฐานความรู้ (Knowledge Base) ให้กับหุ่นยนต์ โดยอ้างอิงตามรูปแบบของมาตรฐาน เอไอเอ็มแอล เวอร์ชัน 1.0.1
5. รวมตัวแปลคำสั่ง (Interpreter) เข้ากับโปรแกรมตัดคำและส่วนการแก้ไขคำผิดอัตโนมัติ
6. ทดสอบและปรับปรุงประสิทธิภาพของวิธีการที่ได้พัฒนา
7. วิเคราะห์และสรุปผลการวิจัย พร้อมข้อเสนอแนะ
8. จัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ได้วิธีการแก้คำผิดแบบไม่ตั้งใจโดยอัตโนมัติในภาษาไทย สำหรับนำไปใช้กับหุ่นยนต์สนทนา

2. สามารถนำวิธีแก้คำผิดที่ได้ไปประยุกต์ใช้ในการช่วยลดรายการคำใกล้เคียงคำผิด และช่วยจัดลำดับ (Ranking) ในงานด้านการตรวจแก้คำผิดสำหรับภาษาไทย เช่น นำไปใช้ในโปรแกรมประมวลผลคำ (Word Processing) ได้

1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตีพิมพ์เป็นผลงานวิชาการ ดังนี้

1. "การจัดลำดับในการตรวจแก้คำผิดอัตโนมัติสำหรับภาษาไทยโดยใช้ตัวอักษรประชิดบนแผงแป้นอักขระ" โดย วนิดา เกษร์สุวรรณ์ วิษณุ โคตรจรัส และณัฐพงษ์ สังขมาลัย ในงานประชุมวิชาการ Thailand Computer Science Conference (ThCSC2004)