บทที่ 1

บทนำ

1.1 ความเป็นมา

ในปัจจุบันคอมพิวเตอร์มีบทบาทต่างๆ ในชีวิตมนุษย์เป็นอย่างมากในฐานะเป็นเครื่องมือช่วยอำนวยความสะดวกให้แก่มนุษย์ เนื่องจากคอมพิวเตอร์เป็นเครื่องจักรสมองกลที่มีสมรรถนะสามารถทำงานได้รวดเร็ว แม่นยำ ถูกต้อง สามารถจัดการกับปัญหาที่ซับซ้อนได้ และยังสามารถทำงานติดต่อกันได้เป็นเวลานาน จึงทำให้มนุษย์สนใจพัฒนาสมรรถนะของคอมพิวเตอร์ให้สามารถช่วยงานมนุษย์ในด้านต่างๆ โปรแกรมในด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing - NLP) ต่างๆ ก็เป็นรูปแบบหนึ่งในการนำคอมพิวเตอร์มาช่วยอำนวยความสะดวกให้แก่มนุษย์ เช่น การประมวลผลคำ (Word Processing), การแปลภาษาด้วยเครื่อง (Machine Translation) การสังเคราะห์เสียงจากข้อความ (Text-To-Speech) การรู้จำเสียง (Speech Recognition) เป็นต้น ซึ่งเรื่องดังกล่าวเกิดจากความมุ่งหวังของมนุษย์ที่จะทำให้คอมพิวเตอร์สามารถรู้จำและเข้าใจภาษามนุษย์ได้ เพื่อที่มนุษย์จะสามารถติดต่อสื่อสารกับคอมพิวเตอร์ได้สะดวกขึ้นโดยใช้ภาษาธรรมชาติของมนุษย์เอง ดังนั้นการพัฒนาระบบการประมวลผลภาษาธรรมชาติด้วยคอมพิวเตอร์จึงนับได้ว่าเป็นความก้าวหน้าทางเทคโนโลยีที่จะช่วยให้เกิดความสะดวกในการติดต่อสื่อสารระหว่างมนุษย์กับคอมพิวเตอร์ได้ (นัฐวุฒิ ไชยเจริญ,2544, น.1) อันจะส่งผลให้คอมพิวเตอร์สามารถทำงานได้อย่างมีประสิทธิภาพ

การทำให้คอมพิวเตอร์สามารถเรียนรู้และเข้าใจภาษามนุษย์ได้นั้นไม่ใช่ เรื่องง่าย การที่มนุษย์สามารถเรียนรู้ภาษาธรรมชาติได้ก็เพราะมนุษย์มีระบบประสาท ระบบสมอง ที่ซับซ้อนและมีกลไกการเรียนรู้ที่ลึกซึ้ง แต่คอมพิวเตอร์ซึ่งเป็นเครื่องจักรไม่ได้มีเครื่องมือต่างๆ ดังกล่าว (นัฐวุฒิ ไชยเจริญ,2544,น.1) ดังนั้นปัญหาในการทำให้คอมพิวเตอร์เข้าใจภาษาในงานประมวลผลภาษาธรรมชาติจึงต้องอาศัยมนุษย์เป็นผู้กำหนดการทำงานให้แก่คอมพิวเตอร์เพื่อให้คอมพิวเตอร์สามารถเรียนรู้ภาษาธรรมชาติของมนุษย์ได้ ในด้านภาษา ภาษาเป็นเครื่องมือสื่อสารอันมีประสิทธิภาพสูงของมนุษย์ มนุษย์ใช้ภาษา ในการสืบทอด ส่งผ่าน และแลกเปลี่ยนความรู้ความคิดกัน ภาษาที่มนุษย์ใช้สื่อสารมีทั้งภาษาพูดที่อยู่ในรูปเสียงพูด และภาษาเขียนที่อยู่ในรูปอักขระในภาษา สำหรับภาษาไทย ภาษาไทยสามารถสื่อสารได้ทั้งทางภาษาพูดและภาษาเขียน

โดยมีระบบเสียงและระบบอักขระเป็นของตนเอง(นัฐวุฒิ ไชยเจริญ,2544, น.2) ลักษณะภาษาเขียนของภาษาไทยที่ไม่เอื้ออำนวยต่อการพัฒนา โปรแกรมได้โดยง่ายอันได้แก่ การที่ภาษาเขียนสามารถเขียนคำเรียงติดต่อกันไปได้โดยไม่มีตัวบ่งขอบเขตของคำที่แน่ชัด เหมือนอย่างเช่นการเว้นช่องว่างระหว่างคำในภาษาอังกฤษ จึงมีความจำเป็นอย่างยิ่งที่จะต้องหาวิธีหรือกระบวนการตัดคำในภาษาไทย (Thai Word Segmentation) เพื่อให้มีผลที่ดีและมีประสิทธิภาพมากที่สุด

ดังนั้นจึงมีรูปแบบเทคนิคที่ใช้ในการตัดคำ และเทคนิคที่นิยมนำมาประยุกต์ใช้มี 3 วิธีด้วยกันคือ

1. การใช้กฎไวยากรณ์ทางภาษา (Rule - Based)

2. การอ้างอิงคำจากพจนานุกรม (Dictionary - Based)

3. การสร้างโมเดลเรียนรู้จากฐานข้อความขนาดใหญ่ (Machine Learning or Corpus Based)

ซึ่งทุกเทคนิคที่ใช้ในการตัดคำจะมีข้อดีและข้อเสียที่แตกต่างกันแต่มีปัญหาลักษณะหนึ่งในการประมวลผลหลังจากการตัดคำ คือ ปัญหาของคำที่คอมพิวเตอร์ไม่รู้จักคำ (Unknown Word) โดยปกติแล้วในการประมวลผลมักจะใช้วิธีให้คอมพิวเตอร์รู้จักคำต่างๆ ผ่านทางคลังคำศัพท์หรือพจนานุกรม มักจะพบได้ว่านิพจน์ระบุนาม (Named Entity) เป็นคำที่คอมพิวเตอร์ไม่รู้จักในการตัดคำภาษาไทย ยังเป็นประเด็นที่ยังมีปัญหาและมีความสำคัญต่อการประมวลผลทางภาษาในการพัฒนางานด้านต่าง ๆ เกี่ยวกับภาษาที่ปรากฏในคอมพิวเตอร์ โดยเฉพาะงานด้านการตัดคำในภาษาไทย (Thai Word Segmentation) ซึ่งเข้ามามีบทบาทในการสนองความต้องการบริโภคข้อมูลข่าวสารของมนุษย์มากขึ้นเรื่อยๆ

ปัญหาที่กล่าวถึงนั้นเนื่องมาจากโดยธรรมชาติของนิพจน์ระบุนาม (Named Entity) แล้วมีแนวโน้มที่จะเกิดขึ้นใหม่อยู่ตลอดเวลา และไม่มีตัวอักษรพิมพ์ใหญ่ที่ระบุชื่อเฉพาะที่ใช้ในภาษาอังกฤษ ทำให้การหาขอบเขตของนิพจน์ระบุนามภาษาไทยยากกว่าภาษาอื่น การใช้วิธีการสอนให้คอมพิวเตอร์รู้จักนิพจน์ระบุนามโดยการใช้คลังคำศัพท์หรือพจนานุกรม จึงไม่สามารถครอบคลุมถึงนิพจน์ระบุนามที่เกิดขึ้นใหม่ ปัญหาที่เกี่ยวเนื่องอีกประการหนึ่งคือปัญหาจากลักษณะโครงสร้างของภาษาไทยและลักษณะโครงสร้างของนิพจน์ระบุนามเองที่มีรูปแบบในลักษณะเดียวกับคำนามทั่วไป เป็นผลให้ระบบคอมพิวเตอร์ไม่สามารถแยกแยะนิพจน์ระบุนามออกจากคำนามทั่วไปได้ ผลจากการที่คอมพิวเตอร์ไม่รู้จักและไม่สามารถแยกแยะนิพจน์ระบุนามได้ ทำให้ความสามารถในการประมวลผลภาษามีประสิทธิภาพลดลง การศึกษาหาวิธีการที่จะทำ

ให้คอมพิวเตอร์สามารถแยกแยะและบ่งชี้ได้ว่าหน่วยภาษานั้นเป็นนิพจน์ระบุนามได้จึงเป็นเรื่องสำคัญเรื่องหนึ่งในการประมวลผลภาษาธรรมชาติ (อมรทิพย์ กวินปณิธาน,2546, น.5)

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อเพิ่มประสิทธิภาพในการตัดคำภาษาไทย (Thai Word Segmentation) โดยการแก้ไขปัญหาของคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Word) ที่เป็นนิพจน์ระบุนาม ประเภท ชื่อคน ชื่อองค์กร และชื่อสถานที่

2. เพื่อศึกษาประเภทตัวบ่งบอกนิพจน์ระบุนาม (Named Entity)

3. เพื่อสร้างโมเดลการจดจำนิพจน์ระบุนามในการจดจำ ชื่อคน ชื่อองค์กร และชื่อสถานที่

1.3 ขอบเขตของงานวิจัย

1. ศึกษาคำขึ้นต้นและคำลงท้ายที่บ่งบอกชื่อเฉพาะ (Named Entity) จากคลังข้อมูล เช่น คำว่า นาย นางสาว นายแพทย์ ศาสตราจารย์ เป็นต้น

2. ศึกษาและพัฒนาขั้นตอน วิธีที่เรียกว่า Conditional Random Fields (CRFs) เพื่อสร้าง โมเดล การจดจำนิพจน์ระบุนาม

3. ประเภทของข้อมูลในคลังข้อมูลที่จะใช้ จากคลังข้อมูลของ BEST2010 (Benchmark for Enhancing the Standard of Thai language processing) ขนาด 5 ล้านคำในการฝึกฝน

1.4 ขั้นตอนการวิจัย

1. ศึกษาปัญหาที่เกิดข้อผิดพลาด จากโปรแกรม Thai Lexeme Analyser (TLex)
2. ศึกษาเทคนิคการตัดคำภาษาไทย
3. วางขอบเขตและกำหนดเป้าหมายของงานวิจัย
4. ออกแบบการทดลอง
5. พัฒนาโปรแกรมและทำการทดลอง

6. สรุปผลการวิจัยและจัดทำรายงานวิทยานิพนธ์

7. นำเสนองานวิจัย

1.5 ผลที่ได้รับ

1. นำวิธีการการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) มาประยุกต์เพื่อใช้แก้ไขปัญหาการตัดคำ

2. นำโมเดลการจดจำนิพจน์ระบุนาม เพื่อมาเพิ่มประสิทธิภาพในการตัดคำภาษาไทย (Thai Word Segmentation)

3. โมเดลการจดจำนิพจน์ระบุนาม สามารถจดจำชื่อ คน ชื่อองค์กร และสถานที่ได้อย่างมีประสิทธิภาพ

1.6 รายละเอียดของวิทยานิพนธ์

รายละเอียดของวิทยานิพนธ์ที่สามารถแบ่งออกเป็นส่วนต่างๆ ซึ่งสามารถแยกอธิบายได้ดังหัวข้อต่อไปนี้

บทที่ 1 บทนำ จะกล่าวถึงความเป็นมาและปัญหา วัตถุประสงค์ของงานวิจัย ขอบเขตงานวิจัย และผลที่ได้รับจากการวิจัย

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ประกอบด้วย ทฤษฎีต่างๆ ที่เกี่ยวข้องกับงานวิจัยนี้ งานวิจัยที่แสดงให้เห็นเทคนิคการตัดคำวิธีต่างๆ

บทที่ 3 วิธีการดำเนินการวิจัย ประกอบด้วยรายละเอียดของข้อมูลที่ใช้ในการวิจัย เครื่องมือที่ใช้ ขั้นตอนการวิจัย

บทที่ 4 ผลการทดลอง แสดงให้เห็นผลการทดลองแต่ละกระบวนการ

บทที่ 5 สรุปผลการศึกษาและข้อเสนอแนะ ประกอบด้วย สรุปและอภิปรายผลการศึกษาวิจัย และข้อเสนอแนะต่างๆ