ระบบการสืบค้นโดยใช้ประโยคคำถาม

QUESTION-BASED SEARCH ENGINE

กิตติพจน์ เตชะเกิดกมล

ปรีดี สิริสม

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2550

QUESTION-BASED SEARCH ENGINE

KITTIPOD TECHAKARDKAMOL

PREEDEE SIRISOM

A SPECIAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIRMENT FOR DEGREE OF BACHELOR OF SCIENCE

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2007

หัวข้อปัญหาพิเศษ	ระบบการสืบค้นโดยใช้ประโยคคำถาม QUESTION-BASED SEARCH ENGINE
ชื่อนักศึกษา	นายกิตติพจน์ เตชะเกิดกมล 47050316 นายปรีดี สิริสม 47050339
ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ธีระ ฟักอ่อน

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้นับปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ประจำปีการศึกษา 2550

คณะกรรมการสอบ	ลายมือชื่อ
อาจารย์วีระชัย ตันยะสิทธิ์ ประธานกรรมการ	
อาจารย์สันธนะ อู่อุดมยิ่ง กรรมการ	
อาจารย์ธีระ ฟักอ่อน กรรมการและอาจารย์ที่ปรึกษา	

(รองศาสตราจารย์ไพรบูลย์ พันธรักษ์พงษ์)
หัวหน้าภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

แด่ คุณแม่ที่เป็นที่รัก

กิตติพจน์

แด่ พ่อ แม่ และเพื่อนๆทุกคน

ปรีดี

หัวข้อปัญหาพิเศษ	ระบบการสืบค้นข้อมูลโดยใช้ประโยคคำถาม	
ชื่อนักศึกษา	นายกิตติพจน์ เตชะเกิดกมล	47050314
	นายปรีดี สิริสม	47050339
ปริญญา	วิทยาศาสตรบัณฑิต	
ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
ปีการศึกษา	2550	
อาจารย์ที่ปรึกษา	อาจารย์ธีระ ฟักอ่อน	

บทคัดย่อ

ในปัจจุบัน ระบบที่ใช้ในการสืบค้นข้อมูลส่วนใหญ่นั้นเป็นระบบที่เรียกกันว่า ระบบการสืบค้นโดยใช้คำสำคัญ (Keyword-Based Search) ซึ่งระบบในการสืบค้นดังกล่าวเป็นระบบที่ผู้ใช้บริการสืบค้นจำเป็นต้องทำการวิเคราะห์สิ่งที่ตัวเองต้องการค้นหา เพื่อให้ได้มาซึ่งคำสำคัญ (Keyword) เพื่อนำไปใช้ในการสืบค้นต่อไป การสืบค้นในระบบดังกล่าว จะทำได้อย่างไร้ประสิทธิภาพ หากผู้ใช้บริการสืบค้นไม่มีแนวคิดที่ถูกต้องในการใช้งานระบบการสืบค้น ซึ่งกระทำให้เลือกคำสำคัญได้ไม่ดี ทำให้การสืบค้นในระบบดังกล่าวมีประสิทธิภาพลดลง

ระบบการสืบค้นข้อมูลแบบใหม่โดยใช้ประโยคคำถามได้พยายามแก้ไขข้อบกพร่องดังกล่าวของระบบการสืบค้นแบบเดิม เช่นระบบการสืบค้นของเวบไซต์กูเกิ้ล โดยระบบการสืบค้นแบบใหม่โดยใช้ประโยคคำถามนั้น ผู้ใช้บริการสืบค้นไม่จำเป็นต้องนั่งวิเคราะห์สิ่งที่ตัวเองต้องการค้นหา เพื่อให้ได้มาซึ่งคำสำคัญเพื่อใช้ในการค้นหาอีกต่อไป ผู้ใช้บริการสืบค้นสามารถกรอกข้อความประโยคคำถามได้เลยโดยตรง จากนั้นระบบจะทำการคัดเลือกคำค้นที่ช่วยให้การสืบค้นข้อมูลดีขึ้น ให้เองโดยอัตโนมัติ แล้วจึงทำการส่งคำค้นที่ได้จากระบบการสืบค้นข้อมูลโดยใช้ประโยคคำถามส่งไปให้ระบบการสืบค้นโดยใช้คำสำคัญทำการสืบค้นอีกทีหนึ่ง ทำให้ผู้ใช้บริการสืบค้นสามารถใช้บริการสืบค้นได้อย่างมีประสิทธิภาพไม่ว่าผู้ใช้บริการสืบค้นจะมีความรู้ความเข้าใจในแนวคิดของระบบการสืบค้นระดับใดก็ตาม

Title	QUESTION-BASED SEARCH ENGINE
Students	Mr.Kittipod Techakerdkamol 47050314
	Mr.Preedee Sirisom 47050339
Degree	Bachelor of Science
Department	Mathematics and Computer Science
Programme	Computer Science
Academic Year	2007
Advisor	Mr.Teera Fagon

ABSTRACT

Nowadays, commonly used search engines are mostly based on keyword search. One constraint in using this type of search engines is to analyze the context on which you want to search. After some analysis, keywords must be provided to be use as an input query. Because of that, searchers must have some knowledge on how to use keyword based search engine. Otherwise, searching in a keyword based search engine will be lack of efficiency.

Our newly developed question based search engine try to fix those problem. Any individuals who performing the search don't have to come up with keywords anymore. First, searchers query using sentences. Next, the question based search engine will automatically generate keywords by analyzing these sentence. Finally, it will transfer the generated keywords to a keyword based search engine. In summary, the searchers don't need any knowledge of the keyword based search engine to search in our question based search engine.

กิตติกรรมประกาศ

ในการจัดทำปัญหาพิเศษเรื่องระบบการสืบค้นข้อมูลโดยใช้ประโยคคำถามนี้ คณะผู้จัดทำขอขอบคุณอาจารย์ธีระ ฟักอ่อน อาจารย์ที่ปรึกษาปัญหาพิเศษนี้ ที่ได้เสียเวลาให้คำปรึกษา รวมถึงคำแนะนำในการสร้าง ปรับปรุงและแก้ไขปัญหาต่างๆ ที่เกิดขึ้นขณะดำเนินการทำปัญหาพิเศษนี้ อีกทั้งยังเป็นผู้ตรวจสอบความสมบูรณ์ถูกต้องของปัญหาพิเศษนี้อีกด้วย

ขอขอบพระคุณคณะอาจารย์ผู้ประสิทธิ์ประสาทวิชาทุกท่านที่ให้ความรู้ทั้งในด้านทฤษฎีและในด้านการลงมือปฏิบัติจริง นอกจากนี้ยังเป็นตัวอย่างที่ดีทางด้านจริยธรรมและคุณธรรมให้แก่คณะผู้จัดทำทำให้ปัญหาพิเศษนี้สำเร็จลุล่วงไปด้วยดีตามเป้าหมายที่ตั้งไว้

ขอขอบพระคุณเจ้าหน้าที่ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ที่ได้สนับสนุนการทำปัญหาพิเศษนี้ในหลายๆ ด้าน ทั้งในด้านการจัดหาวัสดุอุปกรณ์ในการทำปัญหาพิเศษ ด้านการตรวจสอบเอกสารต่างๆ อีกทั้งยังเป็นผู้ที่คอยช่วยเหลือให้คำแนะนำต่างๆ ได้เป็นอย่างดี

สุดท้ายนี้ข้าพเจ้าขอขอบพระคุณบิดา มารดาของข้าพเจ้าที่ได้ให้การอุปการะและสนับสนุนตลอดมา ไม่ว่าจะเป็นในด้านทุนทรัพย์ต่างๆ ตลอดจนถึงกำลังใจทั้งหลายทั้งปวงที่มีให้เรื่อยมาจนสามารถทำให้ข้าพเจ้าสามารถปฏิบัติงานในปัญหาพิเศษนี้ได้สำเร็จลุล่วงไปด้วยดีอีกด้วย

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญตาราง

ตารางที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ในปัจจุบันระบบเครือข่ายได้กลายเป็นปัจจัยสำคัญที่มีผลกระทบอย่างมากต่อผู้คนทั่วโลก มีข้อมูลมากมายมหาศาลโลดแล่นอยู่ภายในระบบ Internet ทำให้การศึกษาและเทคโนโลยีสารสนเทศมีความก้าวหน้าไปอย่างรวดเร็ว เนื่องจากการศึกษาเรียนรู้สามารถทำได้อย่างรวดเร็วเพราะว่าเราไม่จำเป็นต้องเสียเวลาไปหาข้อมูลจากแหล่งอื่นๆ ยกตัวอย่างเช่น การค้นหาข้อมูลจากห้องสมุด การสอบถามจากผู้รู้หรือผู้ทรงคุณวุฒิ เป็นต้น

ในการที่เราจะค้นคว้าหาข้อมูลนั้น โดยส่วนใหญ่แล้วจะต้องพึ่งพาระบบการสืบค้นข้อมูล (Search Engine) ซึ่งระบบดังกล่าวเป็นที่นิยมมากในปัจจุบันและมีให้เห็นกันโดยทั่วไป ไม่ว่าจะเป็น www.google.com หรือ www.yahoo.com เป็นต้น ซึ่งระบบดังกล่าวจะช่วยสืบค้นข้อมูลที่ผู้ใช้ต้องการได้ในระดับหนึ่ง ทำให้การหาข้อมูลข่าวสารผ่านระบบเครือข่ายสามารถทำได้อย่างรวดเร็วและมีประสิทธิภาพ แต่ทว่าในการทำงานของระบบดังกล่าวก็ยังมีข้อจำกัดในการใช้งานอยู่คือ ผู้ใช้บริการสืบค้นจำเป็นที่จะต้องวิเคราะห์สิ่งที่ตนต้องการสืบค้น เพื่อให้ได้มาซึ่งคำสำคัญ (Keyword) เพื่อที่จะนำไปใช้ในการสืบค้นต่อไป ซึ่งในขั้นตอนนี้นี่เอง หากผู้ใช้บริการสืบค้นไม่มีความรู้ความเข้าใจในระบบการสืบค้นข้อมูลในระดับหนึ่งแล้ว คำสำคัญที่ได้จากการวิเคราะห์อาจมีประสิทธิภาพไม่เพียงพอที่จะใช้ในการสืบค้น ส่งผลให้ผลลัพธ์ที่ได้จากการสืบค้นไม่ตรงกับที่ต้องการ ทำให้ระบบการสืบค้นดังกล่าวจะมีประสิทธิภาพเฉพาะกับบุคคลที่มีความรู้ความเข้าใจทางด้านนี้ในระดับหนึ่งเท่านั้น

ดังนั้นระบบการสืบค้นโดยใช้ประโยคคำถามเป็นระบบที่พยายามที่จะแก้ไขข้อจำกัดดังกล่าวของระบบการสืบค้นข้อมูลแบบเก่า ซึ่งผู้ใช้บริการค้นหาไม่จำเป็นต้องวิเคราะห์เพื่อหาคำสำคัญที่จะนำไปใช้การค้นหาอีกต่อไป เพราะว่าระบบการสืบค้นแบบใหม่นั้นจะใช้ประโยคคำถามในการสืบค้นข้อมูล ส่งผลให้ผู้ใช้บริการสืบค้นข้อมูลไม่จำเป็นต้องมีความรู้ความเข้าใจในระบบการสืบค้นอีกต่อไป ทำให้ระบบการสืบค้นสามารถใช้ได้อย่างมีประสิทธิภาพกับผู้ใช้บริการสืบค้นทุกระดับความรู้

1.2 วัตถุประสงค์

1) เพื่อขจัดความยุ่งยากในการคิดคำค้นด้วยตัวเองซึ่งอาจไม่มีประสิทธิภาพเพียงพอในการ

นำไปใช้สืบค้นข้อมูล

2) เพื่อดึงดูดให้ผู้ที่ไม่อยากใช้บริการระบบสืบค้นข้อมูล หันมาใช้ระบบสืบค้นข้อมูลกันมากขึ้น

1.3 ข้อจำกัดและขอบเขต

1) ปัญหาพิเศษนี้จะทำการวิเคราะห์ภาษาอย่างง่ายเพื่อทำการตัดทอนคำ และเลือกคำค้นที่ได้

ไปใช้ในระบบสืบค้นข้อมูล

2) ในการสืบค้นข้อมูล จะใช้ฐานข้อมูลเวบไซต์จาก www.google.com

1.4 ขั้นตอนในการดำเนินงาน

1) ทำการวางแผนและศึกษาข้อมูลเกี่ยวกับวิธีการทำปัญหาพิเศษ
2) ทำการค้นคว้าวิจัยและพัฒนาฐานข้อมูลพจนานุกรม
3) ทำการค้นคว้าวิจัยและพัฒนาวิธีการตรวจสอบความหมายของประโยคและคำในภาษาไทย
4) เริ่มทำการพัฒนาระบบโปรแกรมเบื้องต้น
5) ทำการศึกษาข้อบกพร่องและแนวทางในการพัฒนาระบบโปรแกรม
6) ศึกษาหางานวิจัยและค้นคว้าข้อมูลเพิ่มเติม เพื่อเลือกงานวิจัยและทฤษฎีที่เหมาะสมกับการ
7) นำมาใช้เป็นแนวทางในการพัฒนา
8) ทดลองนำทฤษฎีที่ได้จากงานในข้อ 6 มาใช้เขียนโปรแกรมจริงและทดลองผลกับกลุ่ม
9) ตัวอย่างจำนวนหนึ่ง
10) ตรวจสอบ ปรับปรุง แก้ไข โปรแกรมและรายงานคู่มือทำปัญหาพิเศษ
11) จัดทำเอกสารประกอบการทำปัญหาพิเศษ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) ผู้ใช้งานที่มีความรู้เกี่ยวกับระบบการสืบค้นข้อมูลไม่มาก สามารถทำการสืบค้นข้อมูลได้อย่างมีประสิทธิภาพใกล้เคียงกับผู้ที่มีความรู้ความเข้าใจในระบบการสืบค้น

2) ได้ระบบวิเคราะห์ภาษาเบื้องต้นที่สามารถวิเคราะห์และตัดทอนคำในภาษาไทยซึ่งมีประสิทธิภาพนำมาใช้ในการสืบค้นได้

ส่วนประกอบของปัญหาพิเศษนี้ ได้แบ่งออกเป็นบทต่างๆ ดังนี้

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยต่างๆ ที่เกี่ยวข้องในการทำปัญหาพิเศษครั้งนี้คือ ทฤษฎีและหลักการความกำกวมของภาษาที่ใช้ในการสืบค้นข้อมูล ปัญหาในการทำการวิเคราะห์และตัดทอนคำในภาษาไทย การประมวลผลภาษาธรรมชาติสำหรับภาษาไทย วิธีการตัดทอนคำในภาษาไทยโดยใช้พจนานุกรม การนำกฎทางอักขรวิธีมาใช้ประกอบในการตัดทอนคำในภาษาไทย การตัดทอนคำในภาษาไทยโดยวิธี Longest matching วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด การจัดการกับคำที่ไม่สามารถหาความหมายได้

บทที่ 3 การวิเคราะห์และออกแบบระบบ ในบทนี้จะกล่าวถึงวิธีการนำทฤษฎีต่างๆ ที่ได้กล่าวถึงในบทที่ 2 มาประยุกต์ใช้ในการออกแบบและพัฒนาระบบ วิธีการออกแบบระบบงาน และแผนภาพตัวอย่างที่ได้จากขั้นตอนการทำงานแต่ละขั้นตอนในระบบ

บทที่ 4 ผลการทดลอง ในบทนี้จะกล่าวถึงผลการทดลองขั้นตอนวิธีที่ได้จากการพัฒนาระบบ การทำงานของระบบ ผลการทดลองเชิงสถิติต่างๆ

บทที่ 5 สรุปผลการวิจัย การอภิปราย และข้อเสนอแนะ ในบทนี้จะกล่าวถึงข้อสรุปที่ได้จากการทดลองพัฒนาระบบที่ได้จากบทที่ 4 พร้อมทั้งสรุปข้อดี ข้อเสีย รวมถึงปัญหาที่พบในการพัฒนาระบบ และข้อเสนอแนะในอนาคตสำหรับการนำเอาระบบไปพัฒนาต่อ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการจัดทำระบบการสืบค้นโดยใช้ประโยคคำถามนั้นมีทฤษฎีเกี่ยวข้องเป็นจำนวนมากในการจัดทำ ซึ่งเราได้เลือกมาเฉพาะทฤษฎีและงานวิจัยที่ได้นำมาใช้จริงตามความเหมาะสมเท่านั้น

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ทฤษฎีและหลักการความกำกวมของภาษาที่ใช้ในการสืบค้นข้อมูล

ในด้านการวิเคราะห์องค์ความรู้ที่ใช้ในการสืบค้นข้อมูลในระบบสืบค้นข้อมูลต่างๆ นั้นมักจะเกิดปัญหาก็คือ ไม่สามารถเข้าใจถึงเจตนารมย์ของผู้ทำการสืบค้นได้ เนื่องจากในระบบสืบค้นนั้นมักจะใช้คำค้นที่สั้นซึ่งมักทำให้เกิดความกำกวมในด้านความหมายของคำ ยกตัวอย่างเช่น ผู้ทำการสืบค้นป้อนคำสืบค้นคำว่า “java” เข้ามา ซึ่งหากสังเกตุให้ดีแล้วคำว่า “java”นั้นสามารถตีความหมายได้หลายอย่าง เนื่องจากผู้ใช้ไม่ได้บ่งบอกถึงรายละเอียดการสืบค้นที่มากเพียงพอ ซึ่งทำให้เกิดความกำกวมและสามารถตีความหมายได้หลายอย่าง ผู้สืบค้นอาจต้องการซื้อลิขสิทธิ์ของภาษาจาว่าไปใช้ในเชิงพาณิชย์ หรือผู้ใช้อาจจะต้องการศึกษาเรียนรู้ในตัวภาษาโปรแกรมจาว่า นอกจากนี้อาจเป็นไปได้ว่าผู้สืบค้นต้องการแหล่งกำเนิดของกาแฟชวาก็ได้ เป็นต้น แม้แต่ประเภทของผู้สืบค้นก็มีส่วนสำคัญที่จะทำให้เกิดความกำกวมในการสืบค้น ผู้ที่เชี่ยวชาญในการสืบค้นจะทำให้เกิดความกำกวมน้อยกว่าผู้ที่เพิ่งเริ่มใช้ระบบการวืบค้น เป็นต้น นอกจากนี้อาจจะเป็นเพราะรูปแบบตัวคำค้นที่ผู้สืบค้นทำการป้อนเข้ามาอาจมีรูปแบบที่ไม่ละเอียดพอ ทำให้ได้ข้อมูลไม่ตรงตามที่ต้องการในการสืบค้น เราสามารถวิเคราะห์หาความกำกวมของคำสืบค้นที่ผู้สืบค้นป้อนเข้าได้ โดยการวิเคราะห์คำในประโยคและความสัมพันธ์ของคำในประโยค ซึ่งจะทำเราวิเคราะห์ออกมาได้ว่า ผู้สืบค้นอาจต้องการหรือสนใจคำนั้น ในด้านใดบ้าง เช่นยกตัวอย่างคำว่าเรือ เราอาจจะตีความหมายที่ผู้สืบค้นต้องการได้หลายแบบ ยกตัวอย่างเช่น

1) เรือเอาไปใช้ทำอะไรได้บ้าง (เช่น ลอยเรือ แล่นเรือ ทิ้งสมอ)
2) ชนิดต่างๆ ของเรือ (เช่น เรือยอร์ช เรือหางยาว)
3) วิธีการพายเรือ (เช่น พายช้า พายเร็ว พายอย่างไรให้ปลอดภัย)
4) จะทำอะไรกับเรือได้บ้าง (เช่น การดัดแปลงเรือ การทาสีเรือ)

ซึ่งหลังจากที่เราทำการวิเคราะห์จุดประสงค์ของผู้ทำการสืบค้นแล้ว เราก็จะนำผลลัพธ์ดังกล่าวมาแสดงเป็นรายการให้ผู้ทำการสืบค้นเลือกใช้จุดประสงค์ที่ถูกต้อง โดยจะแสดงเฉพาะรายการที่เกี่ยวข้องและเป็นไปได้เท่านั้น

2.1.2 ปัญหาในการทำการวิเคราะห์และตัดทอนคำในภาษาไทย (Thai Lexical Analysis)

ในขั้นตอนการตัดคำภาษาไทยนั้น จัดว่าเป็นงานที่ทำได้ยากยิ่ง เนื่องจากภาษาไทยนั้นมีรูปแบบเฉพาะตัวที่ก่อให้เกิดปัญหาในการวิเคราะห์ภาษา ดังนี้

2.1.2.1 ปัญหาในการตัดทอนและแยกคำในภาษาไทย

คำในภาษาไทยนั้นไม่มีการกำหนดขอบเขตในตัวมันเองที่ชัดเจนแบบในภาษาอื่นๆ เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส เป็นต้น ยกตัวอย่างเช่นคำว่า ตากลม ซึ่งหากเราพิจารณาให้ดีแล้วจะเห็นได้ว่าคำว่าตากลมนั้นมีความกำกวมในการตัดทอนคำ เพราะเนื่องจากสามารถตัดทอนคำได้ 2 แบบคือ ตา-กลม และ ตาก-ลม ซึ่งทั้ง 2 แบบนั้นมีความหมายในตัวของมันเอง การที่จะเลือกรูปแบบที่ถูกต้องได้จะต้องมีการพิจารณาบริบทแวดล้อมประกอบด้วยจึงจะสามารถทราบรูปแบบใรการตัดทอนที่แน่ชัดได้ ซึ่งเป็นเรื่องที่ทำได้ยาก การตัดทอนคำในภาษาไทยนั้น ปัจจุบันสามารถทำได้ผลถูกต้อง 95 – 99 % โดยประมาณ

2.1.2.2 ปัญหาในการระบุขอบเขตของประโยคในภาษาไทย

เนื่องจากในภาษาไทยนั้นไม่มีการกำหนดขอบเขตของประโยคที่แน่นอน ทำให้การจัดการแบ่งประโยคในภาษาไทยนั้นกลายเป็นอีกหนึ่งปัญหาสำคัญ ในการจัดการแบ่งประโยคนั้นเราสามารถนำช่องว่างมาช่วยในการแบ่งประโยคได้ แต่ก็ยังมีรูปแบบที่เหมือนกันมากเกินไปและเนื่องจากยังไม่ค่อยมีผู้ที่ทำการวิจัยเรื่องขอบเขตของประโยชน์ในภาษาไทยมากนัก ทำให้ในปัจจุบันความแม่นยำในการตัดทอนประโยคในภาษาไทยนั้นมีความแม่นยำอยู่ที่ประมาณ 89%

2.1.2.3 ปัญหาความกำกวมของประเภทของคำในภาษาไทย

เนื่องจากคำในภาษาไทยนั้นสามารถที่จะมีความหมายได้มากกว่า 1 ความหมาย และสามารถมีประเภทของคำได้มากกว่า 1 ประเภท ซึ่งประเภทของคำประกอบไปด้วย 6 ประเภทด้วยกันคือ คำสรรพนาม คำเชื่อม คำลักษณะนาม คำขึ้นต้น และคำบุพบท ซึ่งเป็นการยากที่จะทำการวิเคราะห์หาประเภทของคำ ยกตัวอย่างประโยคเช่น “คุณลุงมาจากบ้าน” หากเราทำการวิเคราะห์ให้ดีแล้วจะเห็นว่า คำว่า “คุณลุง” นั้นสามารถเป็นได้ทั้ง คำนามและคำขึ้นต้น คำว่ามานั้นเป็นได้ทั้งคำกริยาและคำกริยาวิเศษณ์ ส่วนคำว่าจากนั้นสามารถเป็นได้ทั้งคำกริยาและคำบุพบทของกริยาได้

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 การประมวลผลภาษาธรรมชาติสำหรับภาษาไทย

การตัดคำ (Word Segmentation) นั้นเป็นการแบ่งความยาวของตัวอักษรเพื่อหาขอบเขตของคำในแต่ละคำ เนื่องจากคำต่างๆในภาษาไทยมีลักษณะการเขียนที่ติดต่อกันไปตลอดโดยไม่มีการเว้นวรรคแบ่งระหว่างคำที่แน่นอน ซึ่งจะมีก็เพียงแต่การรเว้นวรรคเป็นระยะ เพื่อให้ผู้อ่านได้

พัก หรือทำความเข้าใจความหมายเป็นตอนๆ ซึ่งไม่ได้มีกฎเกณฑ์ในการเว้นวรรคที่ชัดเจนที่ชัดเจน การเว้นวรรคที่ดีนั้นอาจจะช่วยลดความคลุมเครือของคำค้นได้ แต่ไม่ว่าด้วยวิธีการใดๆ หากสามารถบ่งบอกขอบเขตของคำได้แล้ว การจัดการกับข้อความต่างๆ ก็จะมีความสะดวกสบายมากขึ้น ซึ่งในระบบคอมพิวเตอร์นั้นการที่จะหาขอบเขตของคำให้ได้นั้นเป็นสิ่งที่จำเป็นมาก เพราะสามารถที่จะลดภาระของผู้ทำการสืบค้นและสามารถช่วยในการทำงานในระดับที่ลึกลงไปได้ เช่น ฟังก์ชันการจัดฟอร์มให้ชิดขอบขวา การตรวจคำที่สะกดผิดต่างๆ เป็นต้น

2.2.2 วิธีการตัดทอนคำในภาษาไทยโดยใช้พจนานุกรม

ในขั้นตอนการตัดทอนคำนั้นมีหน้าที่หลักๆคือการหาขอบเขตของคำแต่ละคำในประโยค หากมีการเก็บคำทั้งหมดลงไว้ในพจนานุกรมก็จะสามารถหาขอบเขตของคำได้อย่างมีประสิทธิภาพ แต่ในความเป็นจริงนั้น เราไม่สามารถที่จะทำการตัดทอนคำอย่างถูกต้องสมบูรณ์ได้ เนื่องจากเป็นไปไม่ได้ที่จะเก็บคำทุกคำลงในพจนานุกรม เพราะคำทั้งหมดในภาษาไทยนั้นมีจำนวนมากมายมหาศาล และยังมีปัญหาในส่วนของ วิสามานยนาม(คำที่เป็นชื่อเฉพาะ,พระยาอุปกิตศิลปสาร,1990) ตัวเลข หรือคำที่เกิดจากการบัญญัติขึ้นมาใหม่ ทำให้เราไม่สามารถคาดเดาล่วงหน้าได้ว่าจะมีคำว่าอะไรเกิดขึ้นบ้าง เพราะฉะนั้นแม้ว่าการตัดคำจะอาศัยการเปรียบเทียบคำจากพจนานุกรมที่มีฐานข้อมูลคำศัพท์ใหญ่แค่ไหนก็ตาม ก็จำเป็นต้องยอมให้มีคำที่ไม่มีอยู่ในพจนานุกรมเช่นกัน คำที่บรรจุในพจนานุกรมนั้นไม่จำเป็นต้องเป็นหน่วยย่อยสุดของคำ แต่จะเป็นวลีหรือคำประสมก็ได้ เพราะการเก็บคำที่เป็นหน่วยใหญ่ไว้จะมีส่วนช่วยในการกำหนดความหมายของคำในข้อความได้ชัดเจนกว่าในระดับหนึ่ง

2.2.3 การนำกฎทางอักขรวิธีมาใช้ประกอบในการตัดทอนคำในภาษาไทย

แม้ว่าในภาษาไทยจะไม่มีการเว้นวรรคระหว่างคำ ไม่มีเครื่องหมายวรรคตอนที่ชัดเจน ไม่มีตัวบ่งชี้ทางไวยากรณ์ ไม่มีการเปลี่ยนรูปคำ แต่ก็มีกฎทางอักขรวิธีที่กำหนดลักษณะการประสมอักษร การเว้นวรรคและการขึ้นย่อหน้า ซึ่งทั้ง 3 ลักษณะนี้จะเป็นตัวช่วยในการบ่งบอกขอบเขตการพิจารณาเปรียบเทียบคำในพจนานุกรมได้ดังนี้

1) การขึ้นย่อหน้าเป็นการบ่งบอกถึงการสิ้นสุดข้อความ
2) การเว้นวรรคเป็นตัวบ่งชี้ถึงการจบคำหรือประโยค
3) กฎทางอักขรวิธีเป็นตัวบ่งชี้ถึงความเป็นไปได้ที่จะพิจารณาการที่จะแยกสายอักขระ ออกมา

นอกจากนี้ในการพิจารณาว่าเป็นคำหรือไม่เราจะใช้กฎทางอักขรวิธี

อักขระกลุ่มที่ 1 กลุ่มของรูปสระ วรรณยุกต์และเครื่องหมายพิเศษประกอบการเขียนที่ประสมกับพยัญชนะใดๆแล้วไม่เกิดการเคลื่อนขวาของตำแหน่งที่จะเขียนต่อไป ซึ่งอักขระกลุ่มนี้จะไม่สามารถอยู่ตัวเดียวได้ ประกอบไปด้วยอักขระดังนี้คือ ' , ้ , ็ , + , ิ , ี , ึ , ื , ุ , ู , ่ , ้ , ๊ , ๋ , ์ , ั , ๆ , ๆ

อักขระกลุ่มที่ 2 กลุ่มอักขระที่จำเป็นต้องมีรูปพยัญชนะตามมาเสมอ ประกอบไปด้วยอักขรดังนี้คือ เ , แ , โ , ใ , ไ

อักขระกลุ่มที่ 3 กลุ่มอักขระที่จำเป็นต้องมีพยัญชนะอยู่หน้าเสมอ ประกอบไปด้วยอักขรดังนี้คือ ะ , า , ำ , ๅ

อักขระกลุ่มที่ 4 กลุ่มอักขระที่เป็นตัวการันต์ มีไม้ทัณฑฆาตบังคับข้างบน เนื่องจากตัวการันต์จะไม่มีการออกเสียงก็จะไม่มีการพิจารณาให้เป็นตัวอักษรแรกของคำ

อักขระกลุ่มที่ 5 คือกลุ่มอักขระที่เหลือทั้งหมด

ซึ่งอักขรวิธีของอักขระในกลุ่มที่ 1 ถึง 4 สามารถนำมาใช้ประกอบในการวิเคราะห์พิจารณาขอบเขตของคำ เพื่อช่วยในการตัดคำให้ถูกต้องยิ่งขึ้นไปอีกได้

2.2.4 การตัดทอนคำในภาษาไทยโดยวิธี Longest matching

เป็นวิธีการตัดคำแบบหนึ่ง โดยจะทำการตรวจสอบและค้นหาคำจากในพจนานุกรม แล้วจึงทำการแยกคำออกมาหากพบในพจนานุกรม โดยพิจารณาจากซ้ายไปขวา การตรวจสอบจะเริ่มตั้งแต่หน่วยของข้อความที่สั้นที่สุดที่สามารถตัดได้หรือ 1 ประโยคนั่นเอง ซึ่งจะใช้เกณฑ์ทางอักขระเป็นตัวบอกขอบเขตของข้อความ โดยเมื่อตรวจสอบคำในพจนานุกรมไม่พบก็จะลดความยาวของข้อความลงไปตามเกณฑ์ทางอักขรวิธี เช่นข้อความ "รายงานมีผลต่อเกรดปลายภาค" ถ้าไม่พบในพจนานุกรมก็จะลดลงเหลือ "รายงานมีผลต่อเกรดปลายภา" แล้วเป็น "รายงานมีผลต่อเกรดปลายภา" จนสุดท้ายจะได้คำว่า "รายงาน" ซึ่งผลจากการตัดคำจะเป็นดังในตารางด้านล่าง

ส่วนของคำที่ยาวที่สุด	ส่วนที่เหลือ
รายงาน	มีผลต่อเกรดปลายภาค
มี	ผลต่อเกรดปลายภาค
ผล	ต่อเกรดปลายภาค
ต่อ	เกรดปลายภาค
เกรด	ปลายภาค
ปลายภาค	

ตารางที่ 2.1 แสดงผลของการตัดคำโดยใช้วิธี Longest Matching ภายใต้เงื่อนไขของอักขรวิธี

จากงานวิจัยที่ผ่านมาพบว่า การใช้วิธี Longest Matching นั้นจะให้ผลที่ถูกต้องประมาณ 80% ซึ่งจากวิธีดังกล่าวมีข้อบกพร่องที่เห็นได้ชัดคือ การเลือกขอบเขตของคำนั้นยาวเกินไป ทำให้คำที่ตามมาผิดเพี้ยนหรือมีความหมายที่ไม่ตรงตามบริบทของภาษา ยกตัวอย่างเช่น "เดินเป็นการออกกำลังกายชนิดหนึ่ง"จะถูกแบ่งเป็น "เดิน/เป็นการ/ออกกำลังกาย/ชนิดหนึ่ง" เมื่อพิจารณาให้ดีแล้วจะสังเกตุเห็นว่า ในคำที่ 2 คำว่า "เป็นการ" นั้นควรจะแยกออกเป็นคำว่า "เป็น" กับ "การออกกำลังกาย" เพราะมีความหมายที่สมเหตุสมผลกว่า

2.2.5 วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรม(Unknown Word) น้อยที่สุด

วิธีนี้สามารถเรียกได้อีกอย่างหนึ่งว่า Backtracking เป็นวิธีการตัดทอนคำโดยจะพิจารณาจากคำที่ได้ในขั้นตอนการตัดทอนคำโดยใช้วิธี Longest Matching อีกทีหนึ่ง โดยจะเลือกมาทีละคำจากซ้ายไปขวา ยกตัวอย่างเช่น จากผลลัพธ์ในขั้นตอนการทำ Longest Matching คือ "เดิน/เป็นการ/ออกกำลังกาย/ชนิดหนึ่ง" ซึ่งเราจะเริ่มทำการ Backtracking จากคำว่า "เดิน" ซึ่งเป็นคำที่อยู่ทางด้านซ้ายสุดหรือคำแรก ซึ่งต่อไปก็จะเป็นคำว่า "เป็นการ" เพื่อตรวจหากรณีที่เป็นไปได้ทั้งหมดในการตัดทอนคำ ซึ่งจะส่งผลให้ได้ผลลัพธ์คือ "เป็น/การ" และจะสิ้นสุดการทำ backtracking ของคำดังกล่าวที่คำว่า "เป็น" เมื่อทำการ Backtracking เสร็จก็จะทำการคำนวณหาแต่ละกรณีที่เป็นไปได้ โดยจะบังคับให้มีการเกิดคำที่ไม่มีในพจนานุกรมน้อยที่สุด จากนั้นจึงทำการเรียงผลการตัดคำได้ใหม่ โดยจะเลือกผลการทดลองที่มีคำที่ไม่มีในพจนานุกรมน้อยที่สุ่ก่อน จากนั้นจึงเลือกคำที่มีจำนวนคำน้อยที่สุดมาใช้ โดยวิธีการในแต่ละวิธีจะไม่สามารถตัดคำได้ถูกต้อง 100% แต่เมื่อใช้หลักการทั้ง 3 อย่างประกอบกันก็สามารถตัดคำได้มีความถูกต้องเกิน 90% และผลที่ได้ก็เร็วมากพอที่จะใช้งานในขั้นต่อไปได้

2.2.6 การจัดการกับคำที่ไม่สามารถหาความหมายได้

รูปแบบของคำที่ไม่สามารถหาความหมายได้นั้นมักจะมีสาเหตุมาจากการผสมกันของพยางค์ที่มีและไม่มีความหมายในคำที่ไม่มีความหมายได้ซึ่งรูปแบบของคำที่ไม่มีความหมายได้ยกตัวอย่างรูปแบบได้ดังนี้

UKU , UKK , KUKK , UK , KU , KKK

โดยให้ U จะเป็นพยางค์ที่ไม่มีความหมาย และK แทนพยางค์ที่มีความหมาย

2.2.6.1 ประเภทของคำที่ไม่สามารถหาความหมายได้ ซึ่งจากรูปแบบที่ยกตัวอย่างไปดังกล่าวนั้นจะสามารถแบ่งรูปแบบของคำที่ไม่มีความหมายออกมาได้ 2 ประเภทด้วยกันคือ

1) คำที่ไม่สามารถหาความหมายได้แบบเห็นได้ชัดเจน เป็นคำที่อยู่ในรูปแบบที่ชัดเจนและไม่ผิดเพี้ยนไปจากความหมายจริง เพียงแต่ไม่พบในพจนานุกรม ยกตัวอย่างเช่นคำว่า กทม.,โลตัส,สุนีย์

2) คำที่ไม่สามารถหาความหมายได้แบบซ่อนเร้น เป็นคำที่ประกอบไปด้วยที่มีและไม่มีความหมายอยู่ภายในคำ ทำให้การวิเคราะห์และค้นหาตัวคำที่แท้จริงทำได้ยากขึ้น คำที่ไม่สามารถหาความหมายได้ในประเภทนี้จะแบ่งออกเป็น2ประเภทย่อยๆ ด้วยกัน คือ

ก. คำที่ไม่สามารถหาความหมายได้แบบซ่อนเร้นเพียงบางส่วน เป็นคำที่มีพยางค์ที่ไม่มีความหมายและมีความหมายประกอบอยู่ด้วยกัน ซึ่งมีได้หลายรูปแบบ ยกตัวอย่างเช่น

คำว่า "สุมานี" คำว่า "มา" มีความหมาย จึงถูกตัดทอนเป็น "สุ/มา/นี"

คำว่า “ไมโครซอฟต์” คำว่า “โค” และคำว่า “ซอ” มีความหมาย จึงถูกตัดทอนเป็น “ไม โค ร ซอ ฟท์”

โดย ตัวอักษรหนาจะเป็นคำที่มีความหมายในดิกชันนารี

ข. คำที่ไม่สามารถหาความหมายได้แบบซ่อนเร้นโดยสมบูรณ์ เป็นคำที่ประกอบไปด้วยพยางค์ที่มีความหมายเพียงอย่างเดียว เพียงแต่การตัดทอนคำไม่ถูกต้องจนทำให้ความหมายของคำที่แท้จริงผิดเพี้ยนและไม่ใช่สิ่งที่ต้องการ คำในประเภทนี้ยกตัวอย่างเช่น

คำว่า “สมชาย” คำว่า “สม” และคำว่า “ชาย” มีความหมาย

คำว่า “กนกพร” คำว่า “กนก” และคำว่า “พร” มีความหมาย

ซึ่งจากรูปแบบของคำที่ไม่สามารถหาความหมายได้เหล่านี้ เราจะนำไปสร้างคำที่ไม่สามารถหาความหมายทั้ง 2 รูปแบบ

2.2.6.2 การสร้างตัวแทนของคำที่ไม่สามารถหาความหมายได้

ในการสร้างตัวแทนของคำที่ไม่สามารถหาความหมายได้นั้น มีจุดประสงค์โดยรวมเพื่อทำการคำที่ไม่สามารถหาความหมายได้ที่ดีที่สุด โดยจะมีการจัดการทั้งกับคำที่ไม่สามารถหาความหมายได้แบบซ่อนเร้นเพียงบางส่วนและคำที่ไม่สามารถหาความหมายได้แบบซ่อนเร้นโดยสมบูรณ์ โดยในการสร้างคำที่ไม่สามารถหาความหมายนั้น เมื่อพบชุดตัวอักษรใดที่ไม่มีความหมายในพจนานุกรมก็ทำการสร้างชุดของคำขึ้นมาโดยทำการรวมเอาคำที่อยู่รอบๆ ตัวชุดอักษรนั้นเข้ากับตัวชุดอักษร ทำให้เกิดชุดอักษรชุดใหม่ โดยจะทำการรวมคำเป็นจำนวน +/- k คำ ซึ่ง k เป็นค่าคงที่ที่กำหนดไว้ ในที่นี้จะใช้ 3 แทนค่าคงที่

หากมีกรณีที่มีชุดอักษรหลายๆ ชุดอักษรด้วยกันจำนวนหลายตัว อยู่ระหว่างชุดอักษรที่มีความหมายแล้ว ถ้าชุดอักษรที่มีความหมายนั้นมีขนาดน้อยกว่า 3 ตัวอักษรให้ทำการรวมชุดตัวอักษรที่มีความหมายกับชุดตัวอักษรที่ไม่มีความหมายเหล่านั้นเข้าด้วยกันเสีย

$$\text{Sentence} = w_1 w_2 \ldots w_a U w_b \ldots w_n$$

where $w_i \in \text{Dictionary}, U \notin \text{Dictionary}$

n = number of words in the sentence.

$$UNK = \{\, \alpha U \beta \mid \alpha \in A, \beta \in B \,\}$$

where UNK = set of unknown word candidates.

$$A = \{\, w_{a-i,a}, i \in [0, K] \,\} \cup \{\varepsilon\}$$

$$B = \{\, w_{b,b+i}, i \in [0, K] \,\} \cup \{\varepsilon\}$$

$$w_{i,j} = w_i \ldots w_j \quad i < j$$

ε = null string., K = constant value

ภาพที่ 2.1 แสดงสมการที่ใช้ในการสร้างตัวแทนคำที่ไม่สามารถหาความหมายได้

กำหนดให้ $w_1 w_2 \ldots w_n$ แทนประโยคต้นแบบ

ให้ w_i แทนส่วนหนึ่งของประโยค

ให้ t_i แทนลำดับประเภทของคำในประโยค

เช่น ให้ประโยค “ฉันไปเที่ยวน้ำตกทีลอซูกับเพื่อน” ซึ่งผลที่ได้จากวิธีการแยกคำจะเป็น

1) ฉัน/t_1 ไป/t_2 เที่ยว/t_3 น้ำ/t_4 ตก/t_5 ที/t_6 ลอซู/t_7 กับ/t_8 เพื่อน/t_9
2) ฉัน/t_1 ไป/t_2 เที่ยว/t_3 น้ำตก/t_4 ตก ที/t_6 ลอซู/t_7 กับ/t_8 เพื่อน/t_9

เราจะได้คำทั้งหมด 7 คำและคำว่า “ลอซู” เป็นคำที่ไม่สามารถหาความหมายได้ ซึ่งเราจะทำตามทฤษฎีข้างต้นซึ่งจะได้คำที่ไม่สามารถหาความหมายได้ ออกมาเป็น

- ลอซู
- ทีลอซู
- น้ำตกทีลอซู
- ลอซูกับ
- ลอซูกับเพื่อน
- ทีลอซูกับ
- ทีลอซูกับเพื่อน
- น้ำตกทีลอซูกับ
- น้ำตกทีลอซูกับเพื่อน

โดยเซ็ต K,U,A และ B ในรูปที่ 2.1 จะได้เป็น 2 , ลอซู , {ε,ที,น้ำตกที} และ {ε,กับ,กับเพื่อน} ตามลำดับ โดยทุกรูปประโยคจะต้องผ่านกระบวนการทำงานเดียวกันซึ่งจะทำให้เราได้คำที่ไม่สามารถหาความหมายได้ที่เป็นไปได้จากทุกรูปประโยค ซึ่งสร้างขึ้นจากส่วนของชุดอักษรบางส่วน

บทที่ 3
การวิเคราะห์และออกแบบระบบ

จากข้อด้อยของระบบสืบค้นแบบเก่าที่ผู้ใช้จำเป็นต้องวิเคราะห์หาคำสำคัญของสิ่งที่ต้องการค้นหาด้วยตนเอง ซึ่งอาจทำให้คำสำคัญที่ได้มีประสิทธิภาพไม่เพียงพอที่จะได้ผลลัพธ์ที่ต้องการ ระบบสืบค้นข้อมูลแบบใหม่โดยใช้ประโยคคำถามได้พยายามที่จะแก้ไขข้อด้อยนี้ โดยมีการจัดการเลือกคำสำคัญจากสิ่งที่ผู้ใช้บริการต้องการค้นหาให้ ทำให้ผลลัพธ์ที่ได้มีผลที่น่าพึงพอใจมากขึ้น และผู้ใช้ไม่ต้องเสียเวลาในการวิเคราะห์เพื่อหาคำสำคัญที่จะนำไปใช้ในการสืบค้น

3.1 การวิเคราะห์และออกแบบระบบ

การทำงานของระบบการสืบค้นข้อมูลแบบใหม่โดยใช้ประโยคคำถามนั้นมีวิธีการทำงานคร่าวๆ ดังนี้

1) ผู้ใช้บริการกรอกข้อมูลที่ต้องการค้นหาคำตอบให้อยู่ในรูปของประโยคคำถามเชิงเดี่ยว ยกตัวอย่างเช่น "ทำไมเราต้องรับประทานอาหาร" "เราจะหุงข้าวได้อย่างไร" เป็นต้น
2) ระบบจะทำการนำข้อมูลประโยคคำถามเชิงเดี่ยวที่ผู้ใช้กรอกมา ไปทำการประมวลผลเพื่อหาคำค้นที่ดีที่สุด แล้วจึงนำคำสำคัญที่ได้เหล่านั้นไปทำการสืบค้นข้อมูลด้วยระบบสืบค้นข้อมูลแบบเก่าต่อไป

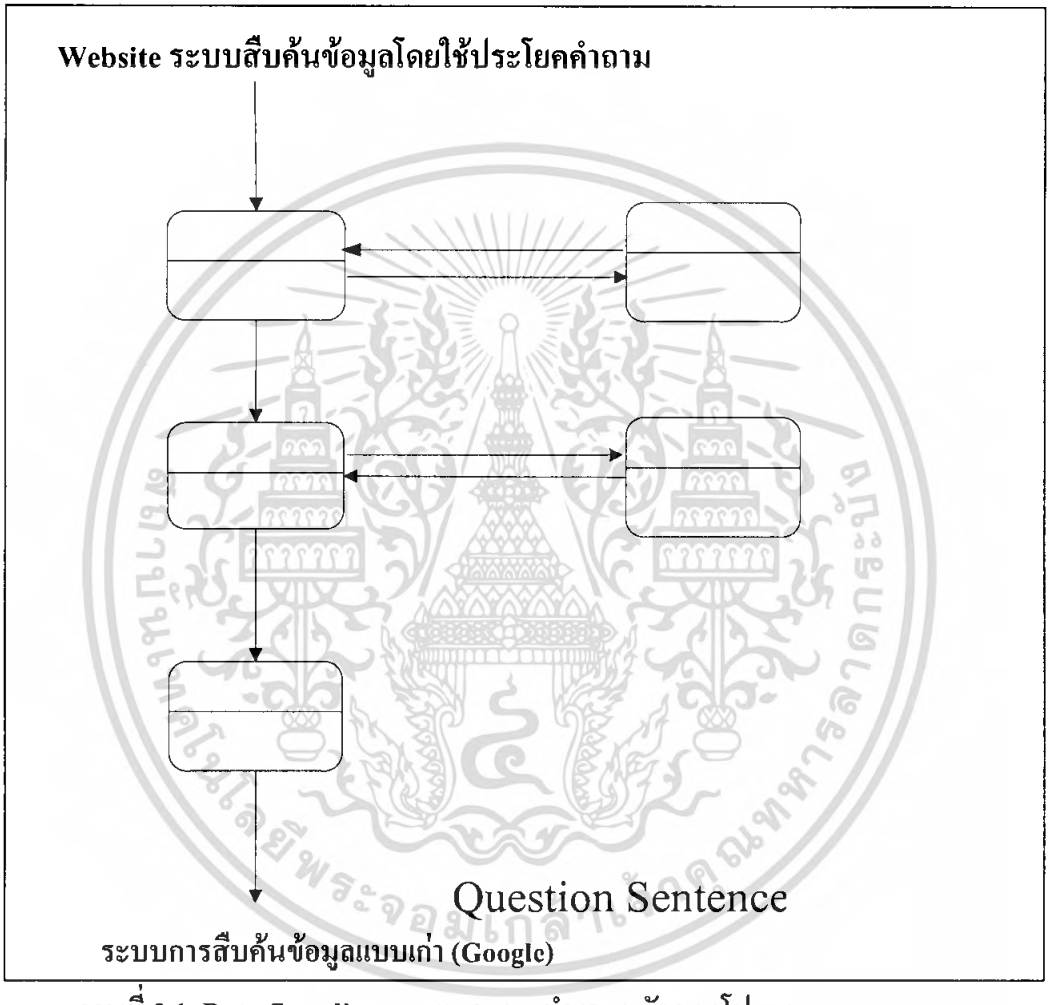
จะเห็นได้ว่าการใช้งานระบบสืบค้นแบบใหม่นั้นขจัดความยุ่งยากที่ผู้ใช้บริการต้องเป็นคนจัดการให้น้อยลงไป และจะส่งผลให้ผลลัพธ์ที่ได้มีประสิทธิภาพดีขึ้นกว่าการที่ผู้ทำการสืบค้นที่ไม่มีความรู้ความเข้าใจในการสืบค้นมากนัก

3.1.2 การออกแบบระบบงาน

ในขั้นตอนการทำงานของโปรแกรมนั้นโดยรวมแล้วจะเป็นการนำเอาประโยคคำถามเชิงเดี่ยวที่ได้จากการป้อนข้อมูลของผู้ใช้บริการ ไปทำการตัดทอน และคัดเลือกคำค้นที่ดีที่สุดส่งต่อไปให้ระบบการสืบค้นข้อมูลแบบเก่าใช้งานต่อ ซึ่งหากเราพิจารณาลำดับขั้นตอนการทำงานดังกล่าวแล้ว เราสามารถแบ่งขั้นตอนการทำงานได้ดังนี้

1) เมื่อผู้ใช้บริการสืบค้นข้อมูลป้อนข้อมูลเข้ามา ระบบจะทำการรับข้อมูลประโยคมาทำการตัดทอน ให้ออกมาอยู่ในรูปของคำหรือกลุ่มคำ ซึ่งเราจะเรียกระบบย่อยที่ทำหน้าที่ในการประมวลผลการทำงานในส่วนนี้ว่าขั้นตอนการวิเคราะห์ภาษา (Lexical Analysis)
2) หลังจากที่ผ่านขั้นตอนการวิเคราะห์ภาษาเรียบร้อยแล้ว ระบบจะนำสิ่งที่ได้จากขั้นตอนการวิเคราะห์ภาษามาใช้เพื่อค้นหาคำสำคัญ (Keyword) ที่แสดงให้เห็นถึงความต้องการของผู้ทำการสืบค้น เราเรียกการทำงานในขั้นตอนนี้ว่าขั้นตอนการวิเคราะห์ความหมายของภาษา (Semantic Analyzer)

3) ทำการนำผลที่ได้จากในข้อ 1 และข้อ 2 มาใช้เพื่อเลือกคำค้นที่มีประสิทธิภาพในการค้นหา โดยมีการถามกลับไปยังผู้ทำการสืบค้นว่ามีเจตนาที่แท้จริงอะไรกันแน่ โดยจะแสดงเฉพาะรายการที่เกี่ยวข้องกับสิ่งที่ผู้ใช้บริการทำการป้อนข้อมูลเข้ามาเท่านั้น ซึ่งได้จากในขั้นตอนในข้อ 2 จากนั้นเมื่อผู้ทำการสืบค้นเลือกเจตนาที่แท้จริงแล้วจริงทำการส่งคำค้นที่ได้ไปยัง Google เพื่อให้ระบบทำการสืบค้นข้อมูลให้ต่อไป ในขั้นตอนนี้เราเรียกว่าขั้นติดต่อกับผู้ทำการสืบค้น (Interface)



ภาพที่ 3.1 Data flow diagram แสดงการทำงานหลักของโปรแกรม

LexicalAnalyzer

Kir

3.1.3 ขั้นตอนการวิเคราะห์ภาษา (Lexical Analyzer)

ในการทำขั้นตอนการวิเคราะห์ภาษานั้นประกอบไปด้วยขั้นตอนย่อยทั้งหมดดังนี้

Sequenc

3.1.3.1 การตัดทอนคำโดยใช้พจนานุกรม

ในการตัดทอนคำโดยใช้พจนานุกรมนั้นเราจะใช้ฐานข้อมูลคำศัพท์ (Text Corpus) ซึ่งในที่นี้เราได้นำฐานข้อมูลคำศัพท์มาจาก Lexitron ซึ่งประกอบไปด้วยข้อมูลคำศัพท์ทั้งหมดราว 40000 กว่าคำ แต่เนื่องจากฐานข้อมูลคำศัพท์ของ Lexitron เองมีคำศัพท์ไม่มากพอ ทำ

Symbol Table

Semantic Analyzer

P

ให้เราต้องทำการเพิ่มคำศัพท์บางประเภทเข้าไปเองบางส่วน เช่น คำแสลง ศัพท์วัยรุ่น หรือนามเฉพาะบางตัวที่นิยมใช้ จากนั้นจึงทำการแยกคำประเภทคำถามออกมาเป็นอีกฐานข้อมูลหนึ่งเพื่อให้สะดวกในการทำงาน

ในการตัดทอนคำโดยใช้ฐานข้อมูลคำศัพท์นั้นมีขั้นตอนในการทำงานดังนี้

1) โหลดข้อมูลพจนานุกรมเข้ามาในระบบ

2) ทำการตัดทอนโดยใช้วิธีการ Longest Matching ก่อน เพื่อเป็นการลดภาระก่อนที่จะเข้าสู่ขั้นตอนการทำ Back Tracking ซึ่งในขั้นตอนนี้จะมีการใช้ขั้นตอนทางอักขรวิธีในภาษาไทยมาประกอบการทำการตัดทอนคำด้วย

3) เข้าสู่ขั้นตอนการทำ Backtracking โดยนำผลลัพธ์ที่ได้จากการทำในขั้นตอนที่ 1 มาทำการ Backtracking ทีละคำ จนครบหมดทุกคำ ซึ่งมีการนำขั้นตอนทางอักขรวิธีในภาษาไทยมาใช้เช่นกัน

4) ทำการหารูปแบบที่ดีที่สุดโดยพิจารณาจากการที่มีจำนวนตัวอักษรที่พบน้อยที่สุด หลังจากนั้นทำการเลือกรูปแบบที่มีจำนวนคำน้อยที่สุด

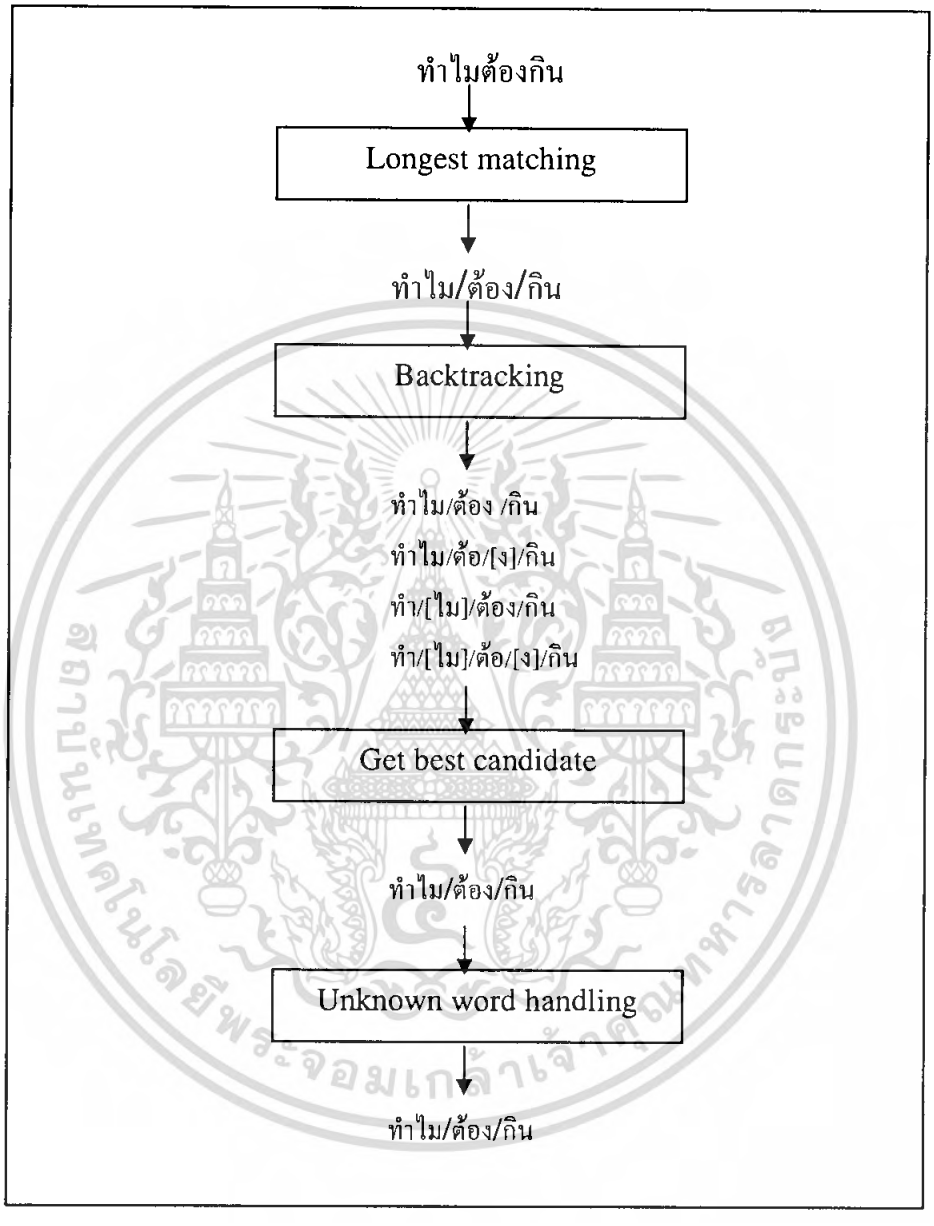
5) ทำการแก้ไขคำที่ไม่ถูกต้องโดยใช้วิธีการรวมคำที่ไม่พบในพจนานุกรมที่อยู่ระหว่างคำที่พบที่มีขนาดไม่เกิน 2 ตัวอักษร ซึ่งหากรูปแบบที่เลือกมาในขั้นตอนการเลือกรูปแบบไม่มีคำที่ไม่พบในพจนานุกรมอยู่เลยก็จะข้ามขั้นตอนนี้ไป

3.1.4 ขั้นตอนการวิเคราะห์ความหมายของภาษา

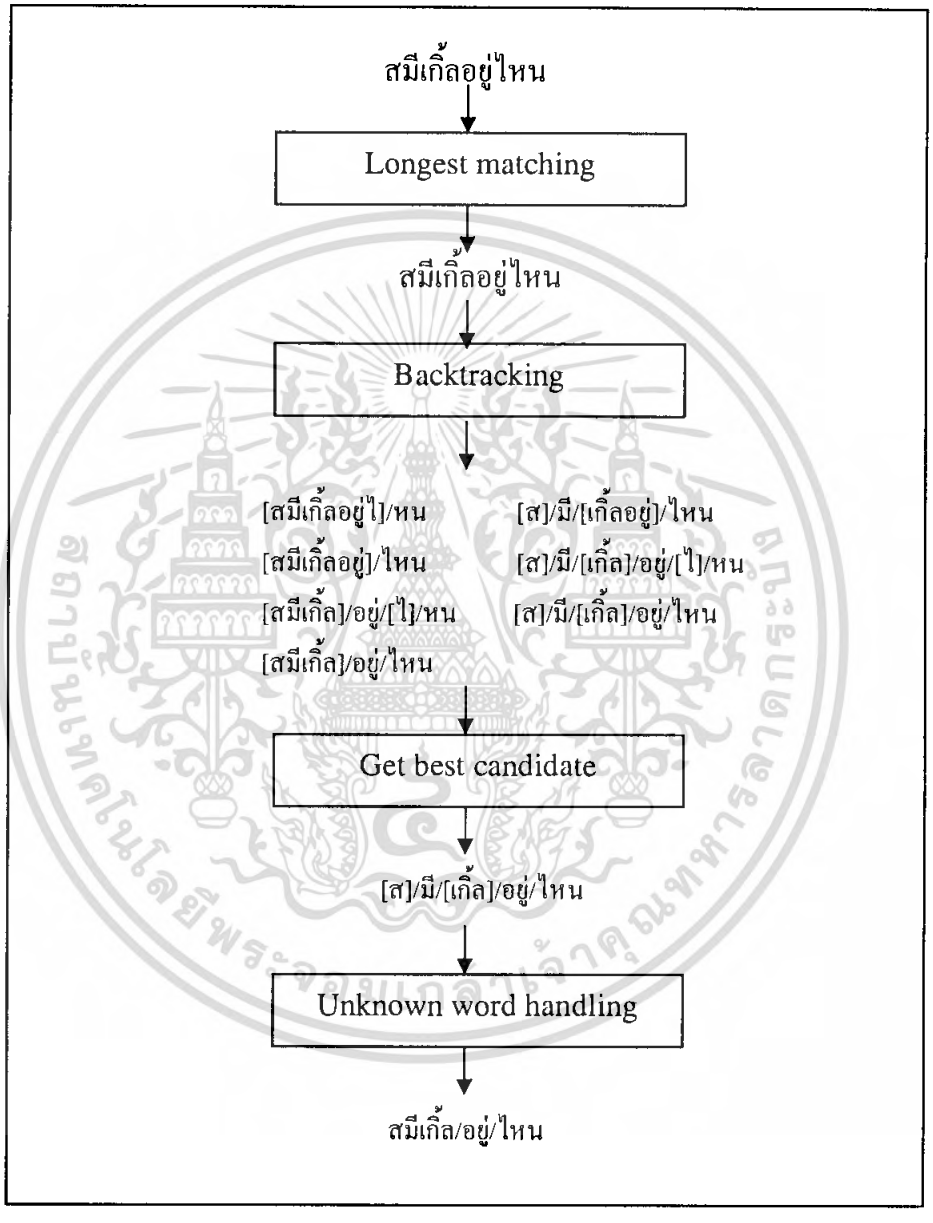
ในขั้นตอนนี้จะมีค้นหาคำที่มีความหมายต่อการวิเคราะห์เจตนานั้นทำไปเพื่อวิเคราะห์ถึงจุดประสงค์ของผู้ทำการสืบค้น และทำการดึงเอาคำถามที่เกี่ยวข้องจากการวิเคราะห์คำค้นที่ได้จากในขั้นตอนการวิเคราะห์ภาษา มาสอบถามผู้ใช้ นอกจากนี้ยังมีการเตรียมคำค้นพิเศษที่มีไว้ใช้ในการปรับปรุงคำค้นอีกด้วย หลังจากที่ทำการปรับปรุงแล้วจะมีการเรียงเอาคำที่มีประเภทของคำเป็นคำนามไปเรียงไว้หน้าสุดของคำค้น เพื่อผลการค้นหาที่ดีขึ้นอีกด้วย

3.1.5 ขั้นตอนการสอบถามผู้ใช้

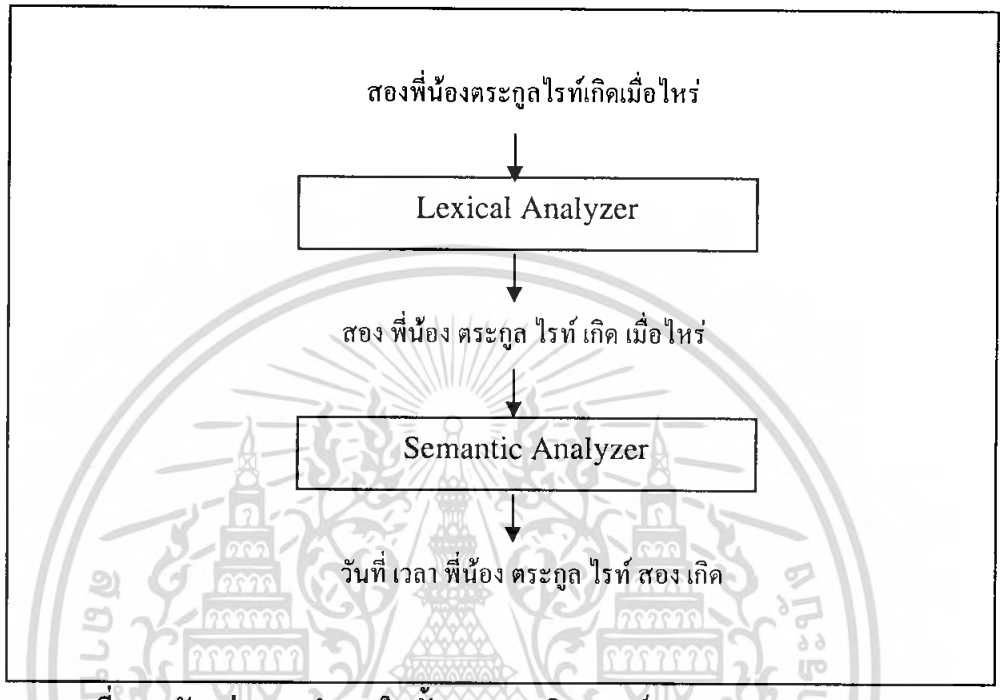
ในขั้นตอนนี้ไม่มีขั้นตอนที่ยุ่งยากมากนัก เพียงแค่นำข้อมูลผลลัพธ์ที่ได้ในขั้นตอนการวิเคราะห์ความหมายของภาษามาสอบถามความต้องการของผู้ทำการสืบค้น จากนั้นจึงส่งคำค้นที่มีการปรับแล้วไปยัง Google



ภาพที่ 3.2 ตัวอย่างการทำงานในขั้นตอนการวิเคราะห์ภาษากรณีที่มีคำค้นที่ไม่มีคำที่ไม่พบในพจนานุกรม



ภาพที่ 3.3 ตัวอย่างการทำงานในขั้นตอนการวิเคราะห์ภาษากรณีที่มีคำค้นที่ไม่พบในพจนานุกรม โดยมีคำที่มีจำนวนตัวอักษรน้อยกว่า 3 ตัวขั้นกลาง



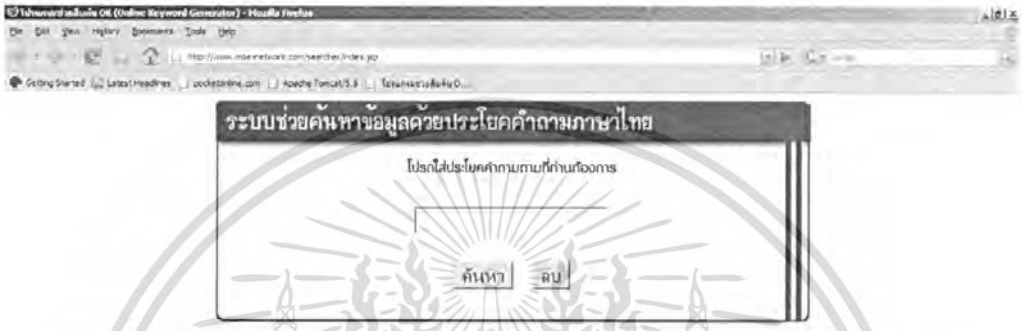
ภาพที่ 3.4 ตัวอย่างการทำงานในขั้นตอนการวิเคราะห์ความหมายของภาษา

บทที่ 4

ผลการทดลอง

4.1 การทำงานของระบบ

ระบบของเราจะมีการทำงานโดยการรับประโยคที่จะใช้ในการค้นหาข้อมูลจากผู้ใช้ผ่านทางป้อนข้อมูลผ่านแป้นพิมพ์

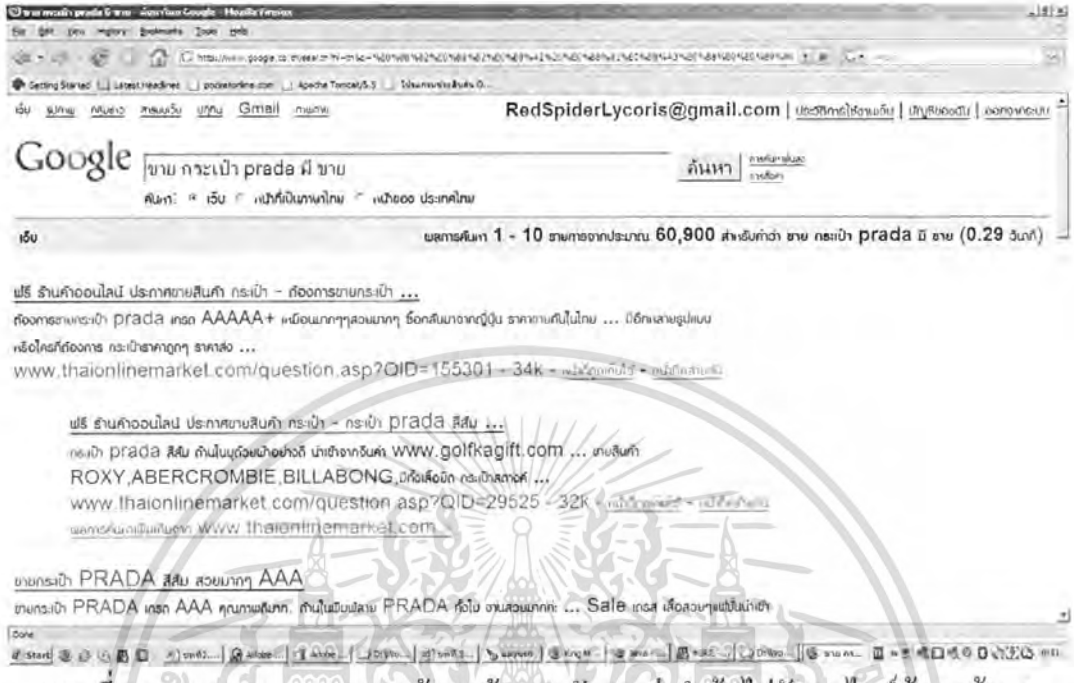


ภาพที่4.1 ภาพแสดงหน้าตาโปรแกรม

โดยเมื่อผู้ใช้ทำการป้อนข้อมูลประโยคคำถามแล้วทำการกดค้นหา ตัวโปรแกรมจะทำการผ่านวิธีการตัดคำและทำการคัดเลือกคำค้นที่จะใช้ในการค้นหาข้อมูล และแสดงให้ผู้ใช้เห็นถึงผลลัพธ์ที่ได้จากการตัดคำ และแสดงลิสต์ให้ผู้ใช้สามารถเลือกว่าต้องการจะค้นหาข้อมูลเพื่อหาข้อมูลแบบใด เช่น ต้องการหาสถานที่ ต้องการหาสินค้าเป็นต้น

ภาพที่ 4.2 แสดงผลลัพธ์ที่ได้จากการใส่ประโยคคำถามเข้าไป

เมื่อทำการกดเลือกรูปแบบการค้นหาตัวโปรแกรมจะทำการส่งคำค้นและคีย์เวิร์ดที่ได้ไปยังเวบไซต์ที่ใช้ค้นหาข้อมูลโดยในที่นี้จะอ้างอิงเวบของ กูเกิ้ล(www.google.com)



ภาพที่ 4.3 ภาพแสดงผลจากการค้นหาข้อมูลหลังจากส่งคำค้นไปยังเวบไซต์ค้นหาข้อมูล

4.2 ผลการทดลอง

ในการทดลองเราได้แบ่งออกเป็น 2 ส่วนคือ

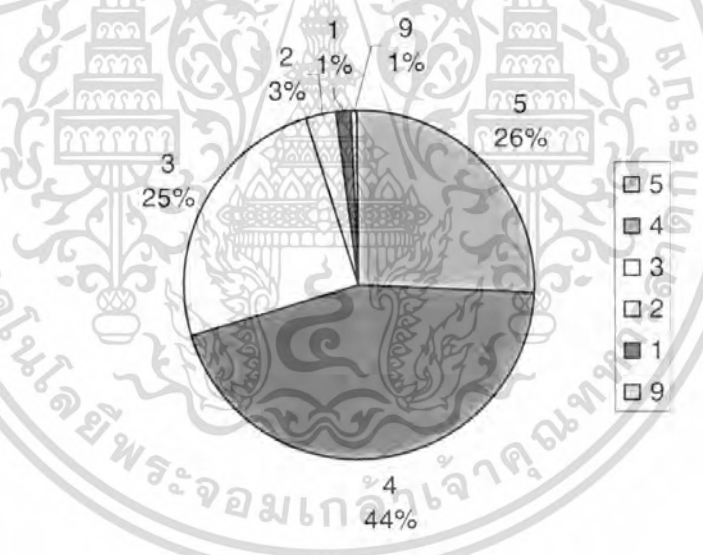
- ส่วนของการสำรวจกลุ่มตัวอย่าง
- ส่วนของการทำการทดลองวัดประสิทธิภาพโปรแกรม

4.2.1 การสำรวจกลุ่มตัวอย่าง

โดยในการทดลองการใช้งานตัวโปรแกรมนั้นทางเราได้ทำการสำรวจข้อมูลจากกลุ่มตัวอย่างโดยที่แบ่งแยกเป็นกลุ่มตัวอย่างที่เป็นนักศึกษาและกลุ่มของบุคคลในวัยทำงานโดยเราได้อธิบายถึงหลักการทำงานของตัวโปรแกรมและให้กลุ่มตัวอย่างทำการทดลองใช้งานตัวโปรแกรมแล้วทำแบบสอบถามโดยจากการสำรวจเราได้ทำการสำรวจกลุ่มตัวอย่างจำนวน41คนซึ่งเป็นนักศึกษาภาควิชาคอมพิวเตอร์จำนว16คนและบุคคลทั่วไปอีก25คนโดยจากการสำรวจเราได้ผลการสำรวจดังนี้

ค่าถาม/ความรู้สึก	5	4	3	2	1	9
1	4	9	3	0	0	0
2	4	9	1	2	0	0
3	3	7	4	1	0	1
4	4	6	6	0	0	0
5	4	8	4	0	0	0
6	4	7	5	0	0	0
7	4	8	3	0	1	0
8	8	5	2	1	0	0
9	2	5	8	0	1	0
ผลรวม	37	64	36	4	2	1
ค่าเฉลี่ย	4.111111	7.111111	4	0.444444	0.222222	0.111111

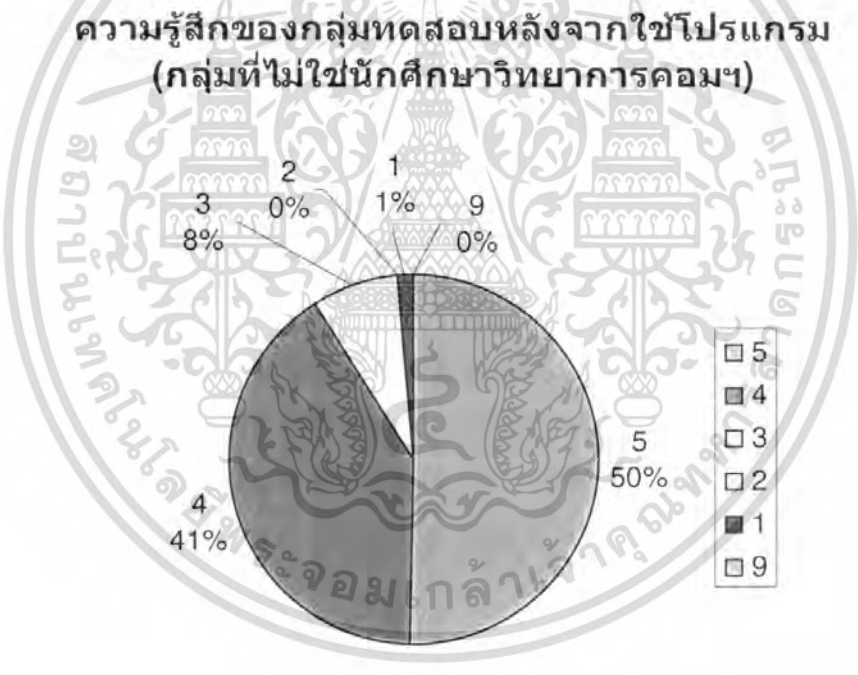
ตารางที่ 4.1 แสดงความรู้สึกของกลุ่มทดสอบหลังจากใช้โปรแกรม(กลุ่มนักศึกษาวิทยาการคอม)



ภาพที่ 4.4 กราฟแสดงความรู้สึกของกลุ่มทดสอบหลังใช้โปรแกรม(กลุ่มนักศึกษาวิทยาการคอมฯ)

ค่าถาม/ความรู้สึก	5	4	3	2	1	9
1	12	12	1	0	0	0
2	11	13	1	0	0	0
3	9	13	2	0	1	0
4	13	9	2	0	1	0
5	10	15	0	0	0	0
6	15	9	1	0	0	0
7	15	5	4	0	1	0
8	15	8	2	0	0	0
9	13	8	4	0	0	0
ผลรวม	113	92	17	0	3	0
ค่าเฉลี่ย	12.55556	10.22222	1.888889	0	0.333333	0

ตารางที่ 4.2 แสดงความรู้สึกของกลุ่มทดสอบหลังจากใช้โปรแกรม(กลุ่มบุคคลทั่วไป)



ภาพที่ 4.5 กราฟแสดงความรู้สึกของกลุ่มทดสอบหลังใช้โปรแกรม (กลุ่มบุคคลทั่วไป)

ค่าถาม/ความรู้สึก	5	4	3	2	1	9
1	16	21	4	0	0	0
2	15	22	2	2	0	0
3	12	20	6	1	1	1
4	17	15	8	0	1	0
5	14	23	4	0	0	0
6	19	16	6	0	0	0
7	19	13	7	0	2	0
8	23	13	4	1	0	0
9	15	13	12	0	1	0
ผลรวม	150	156	53	4	5	1
ค่าเฉลี่ย	16.66667	17.33333	5.888889	0.444444	0.555556	0.111111

ตารางที่ 4.3 แสดงความรู้สึกของกลุ่มทดสอบหลังจากใช้โปรแกรม(ทั้ง2กลุ่มรวมกัน)

ภาพที่ 4.5 กราฟแสดงความรู้สึกของกลุ่มทดสอบหลังใช้โปรแกรม (ทั้ง 2 กลุ่มรวมกัน)

ซึ่งจากการสำรวจจะเห็นได้ว่ากลุ่มทดลองที่เราได้ทำการสำรวจนั้นค่อนข้างพอใจกับการทำงานของตัวโปรแกรม โดยจากการสำรวจเห็นได้ว่าผู้ใช้ส่วนใหญ่พอใจกับการใช้งานโปรแกรมและเห็นด้วยว่าผลลัพธ์ที่ได้นั้นตรงกับความต้องการของผู้ใช้

4.2.2 การทดลองวัดประสิทธิภาพโปรแกรม

เพื่อวัดประสิทธิภาพของโปรแกรมเราได้ทดสอบโปรแกรมโดยการค้นหาข้อมูลเปรียบเทียบกับระบบค้นหาข้อมูลปกติ โดยทำการค้นหาข้อมูลตัวอย่าง จำนวน 40 ตัวอย่าง โดยจะทำการค้นหาข้อมูลประเภทเดียวกันเปรียบเทียบกันซึ่งได้ผลออกมาตามตารางนี้

	จำนวนผลลัพธ์ที่เกี่ยวข้อง (ผ่านระบบ)	**จำนวนผลลัพธ์ที่เกี่ยวข้อง (ไม่ผ่านระบบ)**
ผลรวม	171	84
ค่าเฉลี่ย	4.275	2.1
ค่ามาตรฐานความเบี่ยงเบน	3.194446	3.295685

ตารางที่ 4.4 แสดงผลลัพธ์จากการค้นหาข้อมูลแบบผ่านและไม่ผ่านระบบช่วยการค้นหา

จากการทดสอบค่าผลลัพธ์ที่ได้นั้นจะวัดจากความเกี่ยวข้องของข้อมูลที่ค้นหาได้ว่าตรงกับความต้องการของผู้ใช้ระบบหรือไม่โดยจะมีค่าสูงสุดเป็น10 (จำนวนข้อมูลในหน้าการแสดงของระบบการค้นหาข้อมูล)และต่ำสุดเป็น 0 โดย จากการทดสอบ เมื่อทำการค้นหาข้อมูลผ่านระบบช่วยค้นหาข้อมูล จะได้ค่าเฉลี่ยค่าความเกี่ยวข้องของข้อมูลเป็น 4.725 ที่ค่ามาตรฐานความเบี่ยงเบนเท่ากับ 3.14446 ส่วนการหาข้อมูลโดยไม่ผ่านระบบช่วยค้นหาข้อมูลจะได้ ค่าเฉลี่ยของค่าความเกี่ยวข้องของข้อมูลเป็น 2.1 โดยมีค่าความเบี่ยงเบนมาตรฐานอยู่ที่ 3.295685 นั่นแสดงให้เห็นว่าการค้นหาข้อมูลผ่านระบบช่วยค้นหาข้อมูลนั้นช่วยเพิ่มประสิทธิภาพการค้นหาข้อมูลได้มากขึ้น

หลังจากนั้นเราได้ทำการทดสอบข้อมูลชุดเดิมแต่เปรียบเทียบกับการค้นหาแบบมีการใช้คีย์เวิร์ดเข้ามาช่วยซึ่งแสดงดังตารางที่ 4.5

	จำนวนผลลัพธ์ที่เกี่ยวข้อง (ผ่านระบบ)	**จำนวนผลลัพธ์ที่เกี่ยวข้อง (ไม่ผ่านระบบ)**
ผลรวม	171	168
ค่าเฉลี่ย	4.275	4.2
ค่าเบี่ยงเบนมาตรฐาน	3.194446	3.314305

ตารางที่ 4.5 แสดงผลลัพธ์จากการค้นหาข้อมูลแบบผ่านและค้นหาข้อมูลโดยใช้คำค้นแบบไม่ผ่านระบบช่วยการค้นหา

ซึ่งจากการทดสอบจะเห็นได้ว่าประสิทธิภาพในการค้นหาข้อมูลนั้นแทบไม่แตกต่างกันมากนักโดยการค้นหาข้อมูลผ่านระบบจะหาข้อมูลได้มีประสิทธิภาพที่ดีกว่าเล็กน้อยซึ่งจากการทดลองทั้งหมดกล่าวได้ว่า โปรแกรมนี้จะช่วยให้ผู้ใช้สามารถค้นหาขอมูลต่างๆได้ง่ายขึ้นโดยเฉพาะสำหรับผู้ใช้งานที่ไม่ค่อยมีประสบการณ์ในด้านการค้นหาข้อมูลเท่าไรนัก ถึงแม้ว่าประสิทธิภาพในการค้นหาข้อมูลจะไม่ได้เพิ่มมากขึ้นเมื่อเทียบกับการค้นหาข้อมูลกับวิธีการค้นหาข้อมูลโดยใช้คำค้นในการค้นหาข้อมูล แบบที่ใช้กันทั่วไปในระบบค้นหาข้อมูลก็ตาม

บทที่ 5
สรุปผลการดำเนินงาน การอภิปราย และข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน ในการสร้างระบบการสืบค้นข้อมูลด้วยประโยคคำถามนั้นจัดทำขึ้นเพื่อให้ผู้ที่ไม่มีความรู้ในการใช้ระบบการสืบค้นข้อมูลหรือผู้ที่มีประสบการณ์ในการสืบค้นข้อมูลน้อยใช้ระบบการสืบค้นได้อย่างมีประสิทธิภาพ เนื่องจากการสืบค้นด้วยระบบสืบค้นข้อมูลประเภทที่ต้องใช้คำค้นนั้น ผู้ทำการสืบค้นจะต้องทำการคิดค้นคำค้นเหล่านั้นด้วยตัวเอง ซึ่งอาจส่งผลให้ผู้ทำการสืบค้นไม่อยากที่จะใช้งานอีกเนื่องจากมีความยุ่งยาก หรืออาจคิดค้นคำค้นได้ไม่มีประสิทธิภาพพอ ทำให้ได้ผลลัพธ์การสืบค้นที่ไม่น่าพึงพอใจ ระบบการสืบค้นข้อมูลด้วยประโยคคำถามนั้น มีการทำงานหลักๆ ก็คือ ทำการตัดทอนคำโดยใช้พจนานุกรม แก้ไขคำที่ไม่พบในพจนานุกรม จากนั้นจึงทำการวิเคราะห์สิ่งที่ผู้ทำการสืบค้นต้องการโดยถามกลับไปยังผู้ทำการสืบค้น เมื่อทราบถึงเจตนาของผู้ที่ทำการสืบค้นแล้วจึงทำการส่งคำค้นทั้งหลายเหล่านั้นไปยังระบบสืบค้นแบบใช้คำค้น หลังจากที่ได้พัฒนาระบบการสืบค้นแล้ว เราได้นำวิธีการวัดผลหลายๆ วิธีมาทดสอบระบบเข้าด้วยกัน ทั้งนี้เพื่อวัดทั้งในด้านความพึงพอใจของผู้ใช้งานระบบและทั้งประสิทธิภาพในการสืบค้นของตัวระบบการสืบค้นเอง ผู้ที่ทดลองใช้ระบบการสืบค้นโดยใช้ประโยคคำถามแต่ละคนชื่นชอบในความสะดวกสบายที่ระบบจัดทำให้เป็นอย่างมาก และขอติดต่อนำไปลองใช้จริงเป็นจำนวนมาก

5.2 การวิจารณ์ผลการดำเนินงานของระบบ ข้อสังเกตุที่เห็นได้ชัดจากการดำเนินงานก็คือ สามารถทำการวัดผลได้ยาก เนื่องจากระบบสืบค้นข้อมูลด้วยประโยคคำถามนั้นไม่มีฐานข้อมูลเวบไซต์เป็นของตัวเอง ทำให้การวัดผลเชิงการสืบค้นสารสนเทศ (Information retrieval) นั้นทำได้ยาก จึงจำเป็นต้องพึ่งผู้ประเมินในการวัดผลประสิทธิภาพของระบบ ซึ่งตัวผู้ประเมินเองนั้นก็อาจมีความหลากหลายในด้านต่างๆ ทำให้การวัดผลด้วยวิธีแบบนี้ดูไม่ค่อยน่าเชื่อถือนัก ในการเลือกทฤษฎีต่างๆ ที่นำมาใช้ในการทำโครงงานนี้ได้พิจารณาเลือกโดยดูจากผลการทดลองที่ได้ผลดีในระดับหนึ่ง เหตุที่เลือกเช่นนี้เพราะว่างานด้านการวิเคราะห์ไวยกรณ์ภาษาไทยนั้นทำได้ยาก นอกจากนี้ยังต้องใช้ฐานข้อมูลต่างๆ ที่ได้จากการทดลองและปฏิบัติงาน ซึ่งข้อมูลเหล่านี้มักไม่เปิดเผยให้คนภายนอกจึงสามารถทำได้แค่วิธีง่ายๆ เท่านั้น

5.3 ข้อเสนอแนะ ในการทำการวิเคราะห์ไวยกรณ์ภาษาไทยนั้น ยังถือว่าอยู่ในระดับขั้นพัฒนา ซึ่งมีผู้ที่ทำการ ค้นคว้าและวิจัยงานในด้านนี้ไม่มากนัก ทำให้การหาข้อมูลและวัสดุอุปกรณ์ต่างๆ เป็นไปด้วยความยากลำบาก เนื่องจากงานทางด้าน

วิเคราะห์ไวยกรณ์ภาษาไทยที่ทำแล้วได้ผลสูงหรือประสบผลสำเร็จมักเก็บไว้ส่วนตัว ไม่เปิดเผยต่อคนนอก ประกอบกับการที่การวิเคราะห์ไวยกรณ์ภาษาไทยนั้นทำได้ยากและมีความกำกวมมาก ทำให้การทำงานด้านนี้เน้นไปที่การใช้ฐานข้อมูลขนาดใหญ่เป็นส่วนมาก ถึงแม้ว่าจะทำให้ผลการทำงานของระบบออกมาดี แต่ก็ต้องแลกกับเวลาที่เสียมากขึ้นในการประมวลผล

รายการอ้างอิง

[1] Michael negnevitsky, Artificial intelligence : A Guide to Intelligent Systems , McGrawHill, 2004 .

[2] Y. Daniel Liang , Introduction to Java Programming , Prentice Hall . 2004 .

[3] Linz Peter, An Introduction to Formal Languages and Automata 4th ed. Jones and Bartlett. 2006.

[4] Alfred V.Aho, Ravi Sethi, Jeffrey D.Ullman, Compilers Technicals, Principles and Tools. 2003.

[5] Paisarn Charoenpornsawat., Boonserm Kijsirikul, and Surapant Meknavin, "Feature-Based Proper Name Identification in Thai" , In Proceeding of the National Computer Science and Engineering Conference'98 (NCSEC'98). Bangkok, Thailand, 1998.

[6] Paisarn Charoenpornsawat., Boonserm Kijsirikul, and Surapant Meknavin , "Feature-based Thai Unknown Word Boundary Identification Using Winnow" , In Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS'98), Chieng Mai, Thailand , 1998.

[7] Peter Bruza , Robert McAthur , Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval , Athens , Greece , 2000.

[8] Virach Sornlertlamvanich , Thatsanee Charoenporn , Hitoshi Isahara , "ORCHID: Thai Part-Of-Speech Tagged Corpus", Proceedings of the Natural Language Processing Pacific Rim Symposium , 1997

[9] James Allan , Hema Raghavan , "Using Part-of-speech Patterns to Reduce Query Ambiguity" , Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval , 2002.

ภาคผนวก ก.

ฐานข้อมูลประเภทของคำในภาษาไทย

การออกแบบฐานข้อมูลคำศัพท์

การออกแบบเรายังใช้งานกับรูปแบบการออกแบบเดิมซึ่งเน้นการเก็บข้อมูลที่จำเป็นให้น้อยที่สุดเท่าที่จะเป็นไปได้ เพราะว่ามาตรฐานการทำเครื่องหมายในภาษาอย่าง เอสจีเอมแอล(SGML) นั้นแพงมากและยังต้องการการวิจัยก่อนที่จะนำไปใช้งาน

โครงสร้างของฐานข้อมูลคำศัพท์

โครงสร้างของฐานข้อมูลคำศัพท์จะแบ่งออกเป็น2ส่วน คือ ส่วนที่เป็นบรรทัดของข้อมูลของคำศัพท์ ซึ่งจะขึ้นต้นด้วย "%" และส่วนที่เป็นลำดับบรรทัดซึ่งจะขึ้นต้นด้วย"#" โดยทั้ง2คลาสจะไม่ได้อ้างอิงเป็นส่วนหนึ่งของคำศัพท์ โดยในส่วนของบรรทัดของข้อมูลคำศัพท์จะขึ้นต้นบรรทัดด้วย % ใช้ในการให้ข้อมูลเกี่ยวกับคำศัพท์นั้นๆเหมือนกับที่โชว์ในตารางที่1 โดยข้อมูลของคำศัพท์นั้นจะมีทั้งภาษาไทยและภาษาอังกฤษ ทำให้เข้าถึงข้อมูลได้จากทั้งภาษาไทยและอังกฤษ ในส่วนของปีของคำศัพท์ถ้าเป็นภาษาไทยจะใช้ พ.ศ.และสามารถแปลงปีเป็น ค.ศได้เพื่อป้องกันความสับสนในการอ้างอิง หากมีบรรทัดขึ้นด้วย%ที่ไม่ได้เป็นคำที่กำหนดไว้ในตาราง จะเป็นส่วนของข้อมูลเพิ่มเติมโดยผู้ใช้ ซึ่งจะเรียกว่า บรรทัดแสดงความคิดเห็น

Mark-up	Description
%TTitle:	Title of the document written in Thai.
%ETitle:	Title of the document written in English.
%TAuthor:	Author's name written in Thai.
%EAuthor:	Author's name written in English.
%TInbook:	Title of the book where the document exists, written in Thai.
%EInbook:	Title of the book where the document exists, written in English.
%TPublisher:	Publisher of the book, written in Thai.
%EPublisher:	Publisher of the book, written in English.
%Page:	Page number or the range of pages of the document.
%Year:	Published year (A.D.).
%File:	File number of the document. A long document may be separated into a number of files.

ตารางที่ 1 Mark-up for Text Infromation Line

ในส่วนของลำดับบรรทัด จะขึ้นต้นบรรทัดด้วย # ใช้เป็นตัวบอกลำดับของบรรทัดของเนื้อหา โดยจะมีอยู่ 2 รูปแบบที่แสดงให้เห็นในตารางที่ 2 ตัวเลขจะเป็นดรรชนีบอกหมายเลขในแต่ละย่อหน้า โดยตัวเลขจะถูกสร้างขึ้นโดยอัตโนมัติเมื่อมีการทำสัญลักษณ์บนข้อมูล และเมื่อมีการแก้ไขข้อมูลต่างๆตัวเลขจะมีการปรับให้ข้อมูลสอดคล้องกันเสมอ

Mark-up	Description
#P[number]	Paragraph number of a text. The number in the bracket is shown in a sequence within a text.
#[number]	Sentence number of a paragraph. The number in the bracket is shown in a sequence within a paragraph.

ตารางที่ 2 Mark-up for Numbering Line

นอกจากอักขระที่กล่าวมาข้างต้นแล้วยังมีอักขระพิเศษที่จะแสดงอยู่ในตารางที่ 3 คือ "\\" "//" "/POS"

ซึ่ง "\\" จะทำให้มีการแตกข้อมูลในย่อหน้าออกมาเป็นบรรทัดๆเพื่อหลีกเลี่ยงปัญหาความซ้ำซ้อนโดยจะแยกออกเฉพาะตรงที่มีการเว้นวรรค

"//" ใช้เป็นตัวบอกจุดสิ้นสุดประโยค

และ "/"ที่ตามด้วย POS ใช้ในการบ่งบอกประเภทของคำ

Mark-up	Description
\\	Line break symbol.
//	Sentence break symbol.
/[POS]	Tag marker for the appropriate POS of a word.

ตารางที่ 3 Special Characters for Mark-up

นอกจากนี้ยังมีอักขระอื่นๆนอกจากที่กล่าวมา เราจะใช้การใส่<>ครอบคำเอาไว้เพื่อเป็นการกำหนดตัวอักขระซึ่งจะมีรูปแบบตามตารางที่4

Special characters	Defined strings	Special characters	Defined strings
	<space>	/	<slash>
!	<exclamation>	:	<colon>
"	<quotation>	;	<semi_colon>
#	<number>	<	<less_than>
$	<dollar>	=	<equal>
%	<percent>	>	<greater_than>
&	<ampersand>	?	<question_mark>
'	<apostrophe>	@	<at_mark>
(	<left_parenthesis>	[	<left_square_bracket>
)	<right_parenthesis>	]	<right_square_bracket>
*	<asterisk>	^	<circumflex_accent>
+	<plus>	_	<low_line>
,	<comma>	{	<left_curly_bracket>
-	<minus>	}	<right_curly_bracket>
.	<full_stop>	~	<tilde>

ตารางที่4 Defined Strings for Special Characters

%TTitle: คาร์บอนไดออกไซด์เลเซอร์กำลังสูงแบบไหลเวียนตามแนวแกน

%ETitle: High-Power Compact Axial Flow CO2 Laser

%TAuthor: ผศ.พิพัฒน์ โชคสุวัฒนสกุล

%EAuthor: [Asst. Prof. Pipat Choksuwatanasakul]

%EInbook: The 6th NECTEC Annual Conference

%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม

%EPublisher: National Electronics and Computer Center, Ministry of Science Technology and Environment

:

#P5

:

#5

ในการวิจัยครั้งนี้เราได้ลองศึกษาการเกิดดิสชาร์จจากลักษณะของรูปทรงของแคโถดที่ใช้ต่างๆ กัน\\

พบว่าการใช้แคโถดเป็นรูปทรงกระบอกกลวงทำให้เกิดกระแสในการดิสชาร์จ//

ใน/RPRE

การ/FIXN

วิจัย/VACT

:

การ/FIXN

ดิสชาร์จ/VACT

//

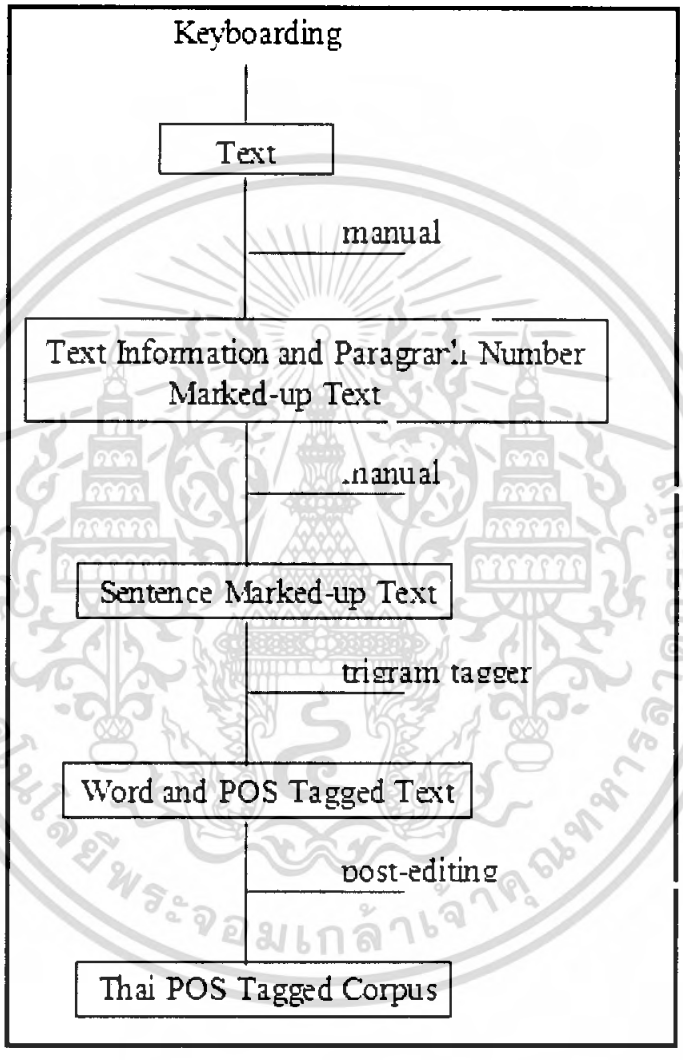
:

รูปที่ 1 A Sample of Thai POS Tagged Corpus

ขั้นตอนการสร้างORCHID

แม้ว่าขนาดของไฟล์คำศัพท์จะเพิ่มขึ้นจากการใช้งานคอมพิวเตอร์ในด้านการพิมพ์ที่กว้างขวางขึ้น ซึ่งภายใต้ทรัพยากรที่จำกัดและมีเนื้อหาภาษาไทยมากมายหลายรูปแบบ การสร้างORCHIDจะทำตามขั้นตอนในรูปที่ 2 ข้อมูลส่วนใหญ่จะรับมาทางแป้นพิมพ์เพราะตัวORCHIDนั้นยังอยู่ในขั้นตอนการพัฒนาและยังขึ้นกับคุณภาพของข้อมูลที่รับเข้ามาอยู่ ซึ่งการดำเนินการอื่นๆนอกจากส่วนของ POS tag ยังทำงานโดยใช้การสนับสนุนจากโปรแกรมอื่นๆ

Figure 2 Procedure in Constructing ORCHID



รูปที่ 2 ขั้นตอนการทำงานของORCHID

กานแยกคำและการกำหนดประเภทของคำ

เรากำหนดปัญหาของการแยกคำใหม่เหมือนกับการกำหนดประเภทของคำ โดยเราจะใส่ค่าประเภทของคำให้คำๆหนึ่งโดยคำนึงถึงความเป็นไปได้ซึ่งจะคำนวณค่าความเป็นไปได้จากแบบจำลองไตรแกรม ซึ่งแสดงในสมการที่ 1 ซึ่ง T จะเป็นลำดับ ประเภทของคำ($t_1,\ldots,t_n$) และ W เป็นลำดับของคำ($w_1,\ldots,w_n$) โดยเราจะใช้วิธีการของวิเทอบิ(Viterbi algorithm) เพื่อคำนวณหาลำดับความเป็นไปได้ของประเภทของคำ และให้แรงค์กับคำในลำดับตามค่าความเป็นไปได้ที่คำนวณออกมา และเพื่อลดจำนวนของค่าความเป็นไปได้ที่คำนวณออกมาเราจะใช้กฎของอักขรวิธีในเรื่องของการประสมตัวอักษรเพื่อลดจำนวนคำที่สามารถเป็นไปได้ลง

$$P(W,T)=\prod_{i=1}^{n} P(t_i|t_{i-1},t_{i-2})\cdot P(w_i|t_i) \qquad \ldots\ldots\ldots (1)$$

คลาสของคำ

ในตอนเริ่มต้นพัฒนาโปรแกรมนั้น ประเภทของคำของเราจะมี 13 หัวข้อหลักและ 45 หัวข้อย่อยซึ่งจะใช้งานทั้งในส่วนของการวิเคราะห์และส่วนของการสร้างและเราได้แก้ไขใหม่ปรับประเภทของคำออกเป็น 14 หัวข้อหลักและ 47 หัวข้อย่อยรวมถึงตารางที่ 5 ด้วย โดยจะเน้นเปลี่ยนในส่วนของหัวข้อย่อยในส่วนของลักษณนาม(CLAS)และคำนำหน้า(FIXP) โดยเราเพิ่มส่วงส่วนของลักษณนามออกเป็น 5ประเภทย่อยๆและแก้ส่วนของคำนำหน้าออกเป็น2ประเภทย่อย โดยเราทำการเปลี่ยนแปลงส่วนของลักษณะนามเพื่อลดความกำกวมของคำลงเพราะจากการศึกษาพบว่าคำที่เป็นลักษณนามส่งผลกับโครงสร้างของภาษาไทยด้วย
ในส่วนของ คำนำหน้านั้น เราได้ทำการแก้ไขเพื่อช่วยแก้ไขปัญหาความกำกวมที่เกิดจากการขาดการผันคำในการเปลี่ยนบทบาทของประโยคในโครงสร้างของคำนามและคำบุพบท

(1) การ/FIXN ออกกำลังกาย/VACT และ/JCRG การ/FIXN พักผ่อน/VACT ที่/PREL เพียงพอ/VSTA เป็น/VSTA สิ่ง/NCMN จำเป็น/VSTA สำหรับ/RPRE มนุษย์/NCMN ทุก/DDBQ คน/CNIT

(2) การ/FIXN ออกกำลังกาย/VACT และ/JCRG ∅ พักผ่อน/VACT ที่/PREL เพียงพอ/VSTA เป็น/VSTA สิ่ง/NCMN จำเป็น/VSTA สำหรับ/RPRE มนุษย์/NCMN ทุก/DDBQ คน/CNIT
ในประโยคที่(2) นั้นยังเป็นประโยคที่ยอมรับได้และมีความหมายเหมือนกับประโยคที่ (1) แม้ว่าคำว่า การ/FIXN จะหายไปจากประโยคด้านบนเราก็ยังสามารถที่จะกำหนดความหมายของคำว่า “การพักผ่อน” แต่ถ้านำคำที่แยกออกมาเป็น2ส่วนเป็นคำว่า “การ” กับ “พักผ่อน” มาคิดเป็นคำเดียวก็จะเกิดความไม่เหมาะสมที่จะให้คำว่า “พักผ่อน” เป็นคำประเภทเดียวกับคำว่า “การออกกำลังกาย”

เราใช้47หัวข้อย่อยนี้เป็นเซตประโยคของประเภทของคำใน ORCHID corpus โดยตารางที่ 5 จะแสดงเซตประโยค

No.	POS	Description	Example
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator “ไม่”	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator “ไม่”	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator “ไม่”	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น

ตารางที่ 5 Thai Part-of-Speech as the Tagset for ORCHID

No.	POS	Description	Example
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ 2 ตัว
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	Adverb with prefixed form	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง

ตารางที่ 6 Thai Part-of-Speech as the Tagset for ORCHID(ต่อ)

No.	POS	Description	Example
35	CFQC	Frequency classifier	ครั้ง, เที่ยว
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก, ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย,โอ้, เออ, เอ๋, อ๋อ
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	Punctuation	(,), “, „, ;

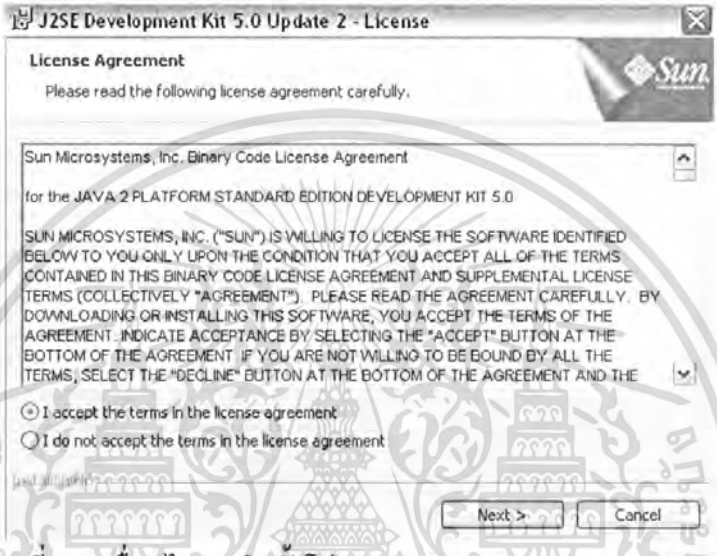
ตารางที่ 7 Thai Part-of-Speech as the Tagset for ORCHID(ต่อ)

ภาคผนวก ข.

การติดตั้งและใช้โปรแกรมระบบการสืบค้นข้อมูลโดยใช้ประโยคคำถาม

ในการติดตั้งโปรแกรมระบบการสืบค้นข้อมูลโดยใช้ประโยคคำถามประกอบไปด้วยวิธีการดังนี้

1) ติดตั้งตัวแปลภาษาจาว่า (jdk-1_5_0_02-windows-i586-p.exe) ซึ่งสามารถหาดาวน์โหลดได้ทั่วไปหรืออาจหาได้จาก http://www.sun.com หลังจากที่ได้ตัวโปรแกรมมาแล้วให้ดับเบิ้ลคลิ๊กที่ไฟล์ jdk-1_5_0_02-windows-i586-p.exe เพื่อให้แกรมทำงานจากนั้นคลิก I accept the terms in the license agreement แล้วคลิก Next ดังรูป

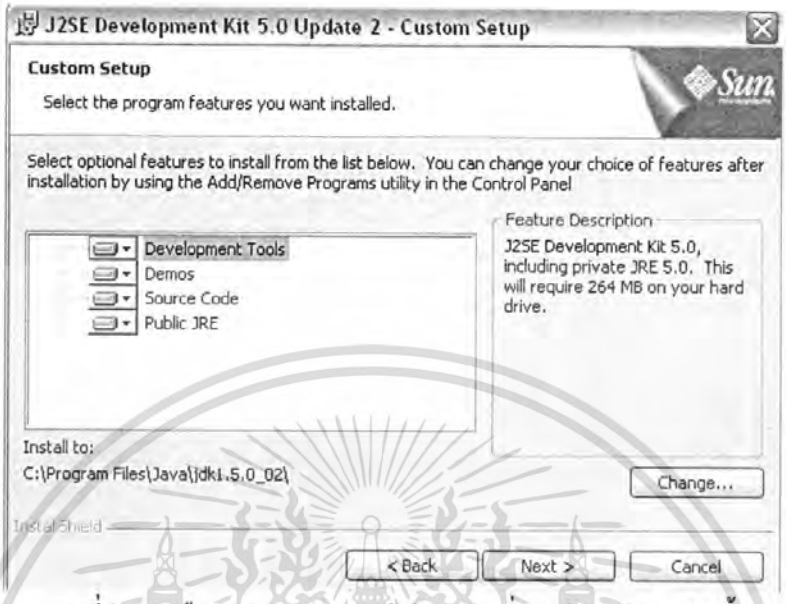


ภาพที่ ข.1 เงื่อนไขการติดตั้งโปรแกรม J2SE Development Kit 5.0

หลังจากนั้นให้เลือกไดเร็กทอรี่ที่ต้องการลงหรือกดปุ่ม OK เพื่อใช้ค่ามาตรฐาน

ภาพที่ ข.2 ระบุตำแหน่งที่จะติดตั้งโปรแกรม

จากนั้นให้กด Next โปรแกรมจะทำการติดตั้งตัวเองให้อัตโนมัติ หลังจากนั้นให้คลิ๊กที่ปุ่ม Finish

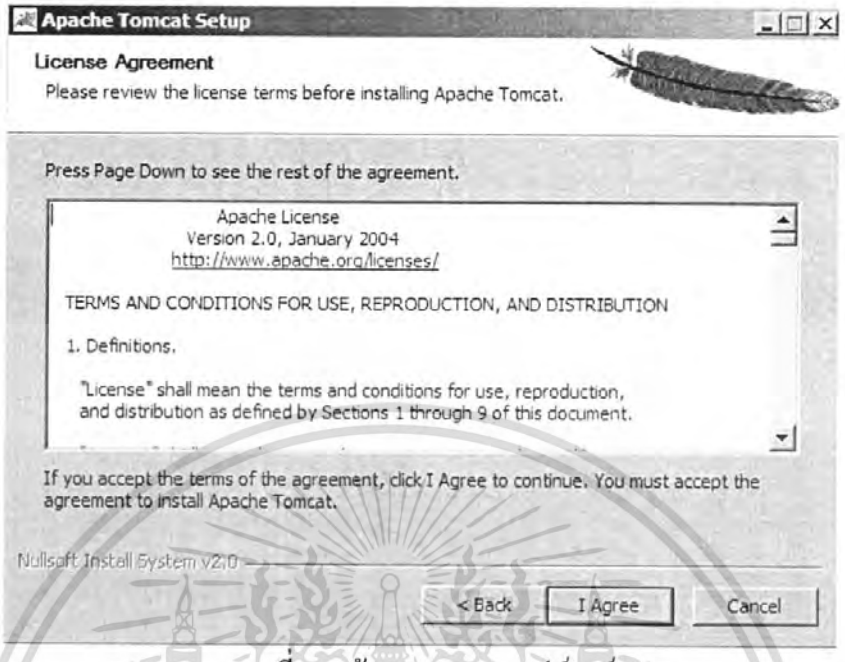


ภาพที่ ข.3 เลือกคุณสมบัติของโปรแกรมที่คุณต้องการจะติดตั้ง

2) ติดตั้ง Apache Tomcat Java Web Server (apache-tomcat-5.5.25.exe) ซึ่งสามารถหาดาวน์โหลดได้จากเวบไซท์ http://tomcat.apache.org หลังจากที่ทำการดาวน์โหลดมาแล้วให้ดับเบิ้ลคลิ๊กเพื่อรันโปรแกมติดตั้งขึ้นมาจะพบหน้าจอดังรูปด้านล่างให้คลิ๊กที่ปุ่ม Next

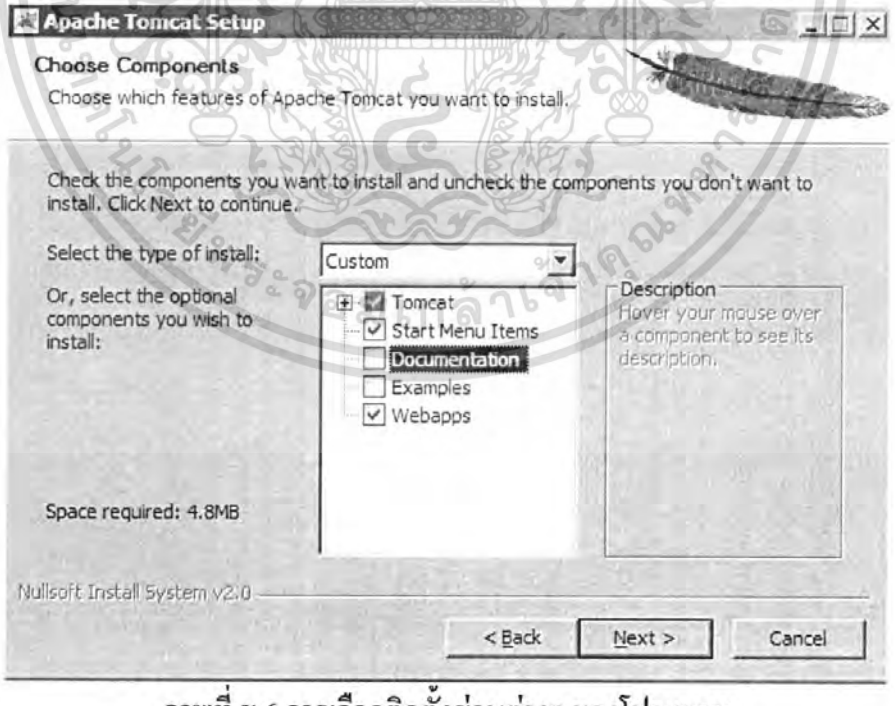
ภาพที่ ข.4 หน้าจอเริ่มการติดตั้ง Apache Tomcat

จากนั้นจะพบหน้าจอดังภาพด้านล่างให้คลิ๊ก I Agree เพื่อยอมรับข้อตกลงของซอฟต์แวร์



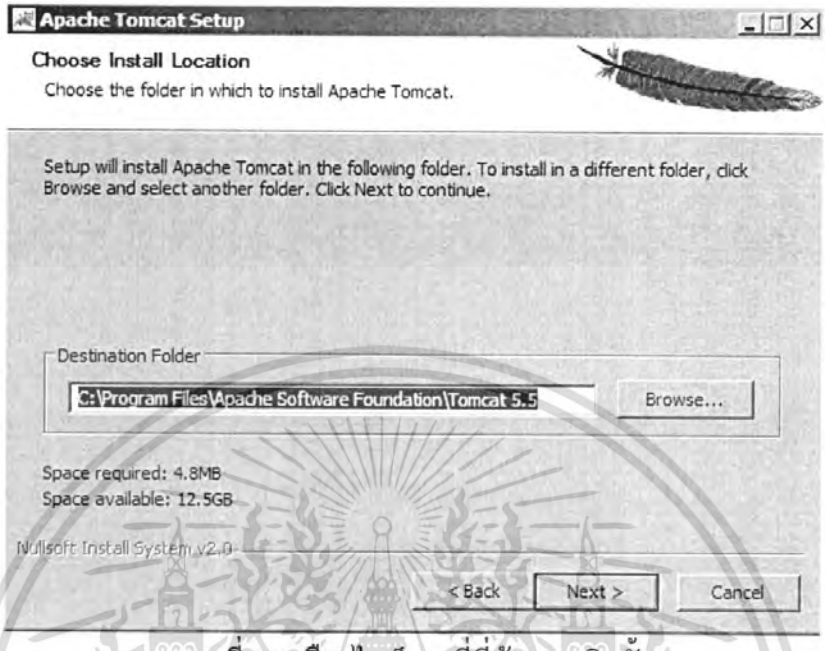
ภาพที่ ข.5 ข้อตกลงของซอฟต์แวร์

หลังจากที่คลิ๊กยอมรับข้อตกลงแล้วให้เลือกติดตั้ง Start Menu Item และ Webapps ดังรูป



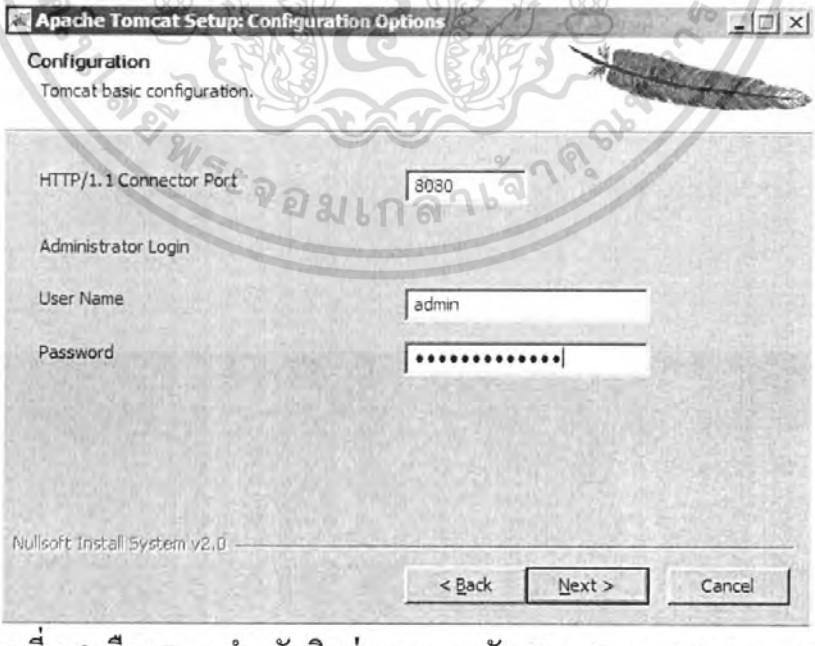
ภาพที่ ข.6 การเลือกติดตั้งส่วนต่างๆ ของโปรแกรม

จากนั้นโปรแกรมจะให้เราเลือกไดเร็กทอรี่ที่ต้องการติดตั้งซึ่งสามารถเปลี่ยนแปลงได้ตามใจชอบ



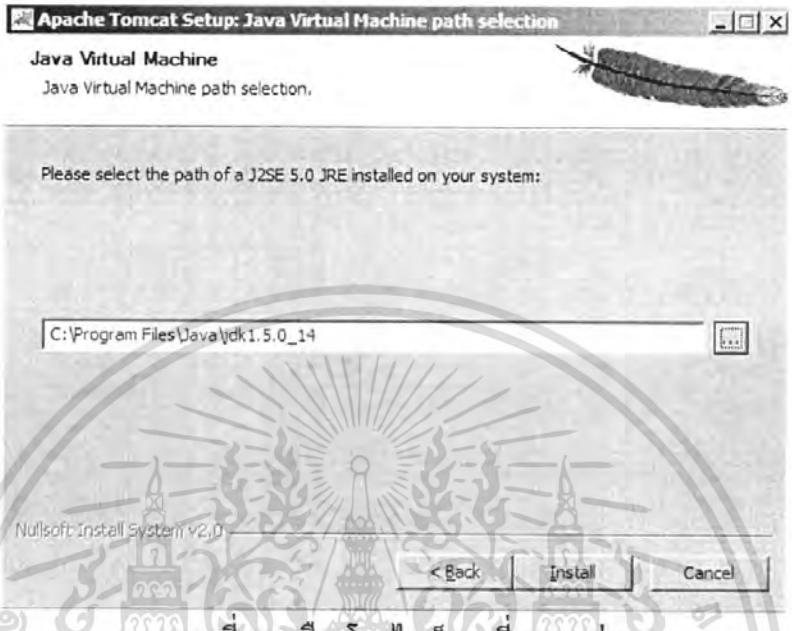
ภาพที่ ข.7 เลือกไดเร็กทอรี่ที่ต้องการติดตั้ง

จากนั้นโปรแกรมจะให้เราเลือก Port สำหรับติดต่อ Tomcat กับ User Name และ Password เพื่อใช้เข้าระบบ



ภาพที่ ข.8 เลือก Port สำหรับติดต่อ Tomcat กับ User Name และ Password

จากนั้นให้ทำการเลือกโฮมไดเร็กทอรี่ของจาว่า ซึ่งปกติแล้วระบบจะเลือกให้อัตโนมัติ แต่หากมีตัวแปลภาษาจาว่ามากกว่า 1 ตัวก็สามารถเลือกไปที่เวอร์ชั่นที่ต้องการได้ ในที่นี้ให้ใช้เวอร์ชั่น 1.5 เนื่องจาก Tomcat 5.5 สนับสนุนแค่ JRE 1.5 เท่านั้น

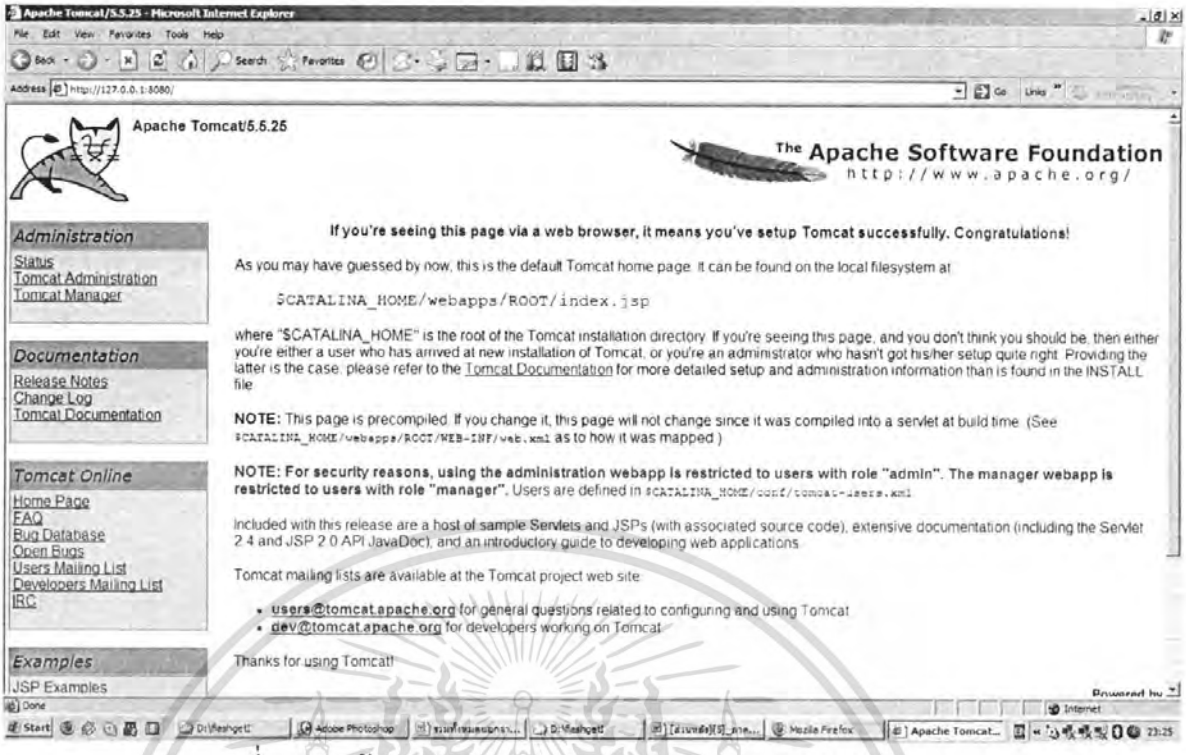


ภาพที่ ข.9 เลือกโฮมไดเร็กทอรี่ของจาว่า

หลังจากนั้นระบบจะทำการติดตั้งโปรแกรมให้อัตโนมัติ เมื่อเสร็จเรียบร้อยแล้วให้กด Finish จะทำการรันโปรแกรมโดยอัตโนมัติ ดังรูป

ภาพที่ ข.10 โปรแกรมทำการติดตั้งอัตโนมัติ

ให้ทดสอบว่า Tomcat ติดตั้งสมบูรณ์ดีรึเปล่าโดยเข้าไปที่เวบไซท์ http://127.0.0.1:8080 จากเวบบราวเซอร์ต่างๆ เช่น Internet Explorer จะปรากฏหน้าจอดังรูป

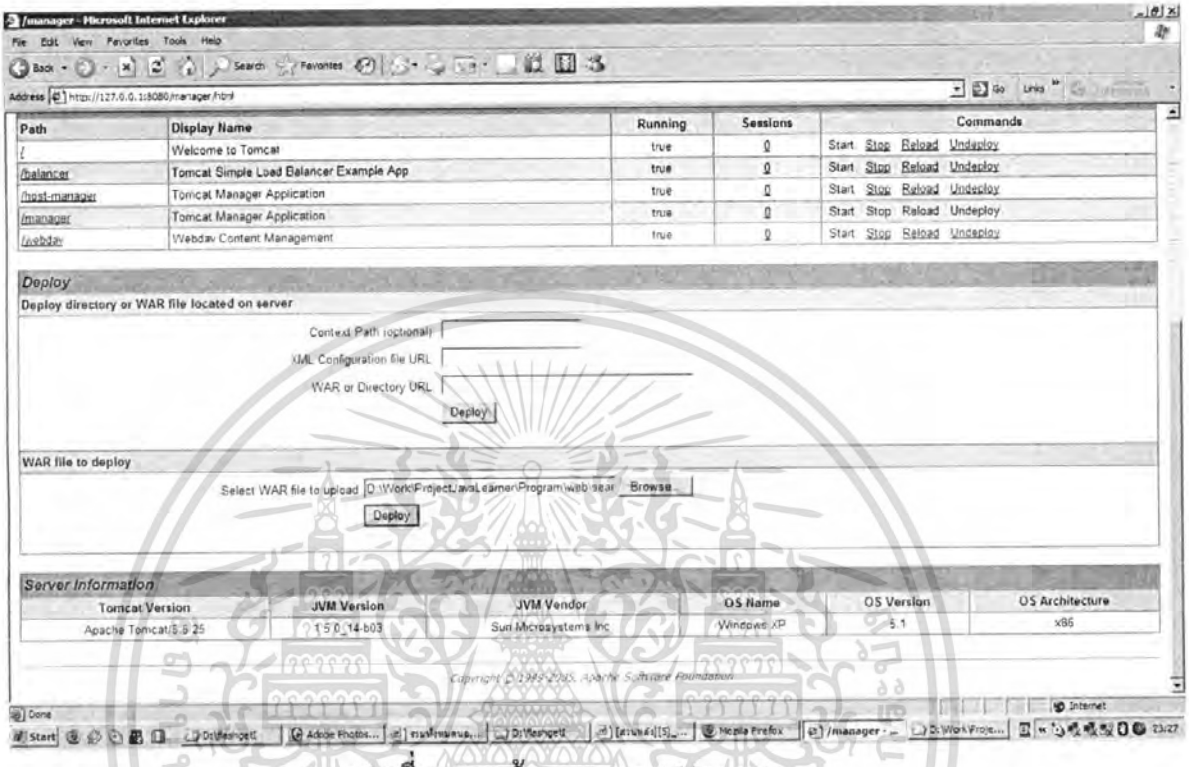


ภาพที่ ข.11 หน้าจอหลักของ Apache Tomcat Java Web Server

ให้คลิ๊กที่ Tomcat Manager เพื่อทำการลงทะเบียนโปรแกรมการสืบค้นด้วยประโยคคำถาม จะปรากฏกล่องข้อความให้กรอก User Name และ Password ตอนติดตั้ง ดังรูป

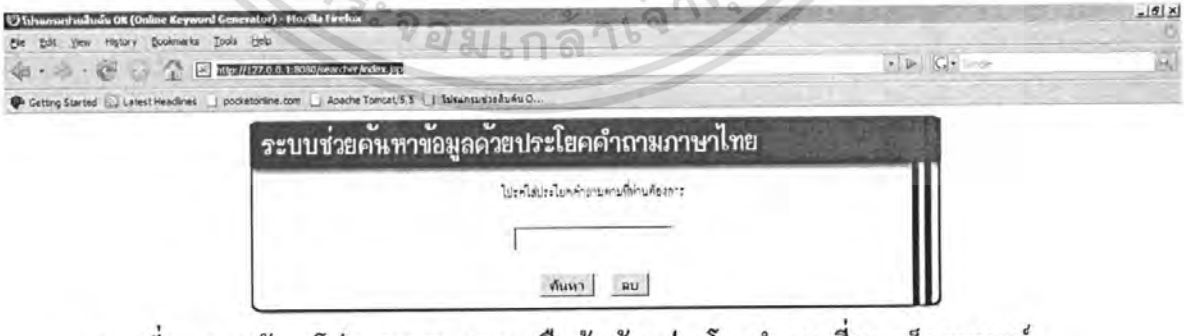
ภาพที่ ข.12 กล่องข้อความ User Name และ Password

หลังจากนั้นให้เลื่อนลงมาที่ช่อง War file to deploy แล้วกดปุ่ม Browse เพื่อทำการเลือกไฟล์ searcher.war ซึ่งเป็นไฟล์ที่เก็บระบบการสืบค้นด้วยประโยคคำถามไว้จากนั้นกดปุ่ม Deploy เพื่อทำการติดตั้ง



ภาพที่ ข.13 หน้าจอ Tomcat Manager

หากการติดตั้งทุกอย่างสมบูรณ์เรียบร้อย ให้ลองเรียก http://127.0.0.1:8080/searcher/index.jsp เพื่อทดสอบว่าระบบทำงานถูกต้องหรือไม่ หากถูกต้องจะปรากฏหน้าจอโปรแกรมดังรูป



ภาพที่ ข.14 หน้าจอโปรแกรมระบบการสืบค้นด้วยประโยคคำถามที่ลงเสร็จสมบูรณ์