รหัสโครงการ 11P34N003

การเพิ่มประสิทธิภาพการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ด้วยเครื่อง

Improvement of using machine learning techniques

โปรแกรมตัดคำภาษาไทย

รายงานฉบับสมบูรณ์

เสนอต่อ

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ
กระทรวงวิทยาศาสตร์และเทคโนโลยี
ได้รับทุนอุดหนุนโครงการวิจัย พัฒนาและวิศวกรรม
โครงการแข่งขันพัฒนาโปรแกรมคอมพิวเตอร์แห่งประเทศไทย ครั้งที่ 9
ประจำปีงบประมาณ 2549

ชื่อผู้พัฒนา

1.นายสิทธิโชค ทรัพย์ไพบูลย์กิจ
2.นายภานุวัฒน์ เมฆะ
3.น.ส. สุพัฒนวรี ทิพย์เจริญ

ชื่ออาจารย์ที่ปรึกษาโครงการ

ผศ.ดร.จีรยุทธ ไชยจารุวณิช

สถาบันการศึกษา

คณะวิทยาศาสตร์ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่

กิตติกรรมประกาศ (Acknowledgement)

งานวิจัยนี้ได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร.จีรยุทธ ไชยจารุวณิช อาจารย์ที่ปรึกษางานวิจัยนี้ ผู้ซึ่งกรุณาช่วยเหลือให้ความรู้ คำแนะนำและ คำปรึกษา รวมทั้งสละเวลาตรวจแก้ไขจนงานวิจัยนี้เสร็จสมบูรณ์ จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบคุณศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ สำนักงานพัฒนาวิทยาศาสตร์และ เทคโนโลยีแห่งชาติ และสำนักงานส่งเสริมอุตสาหกรรมซอฟต์แวร์แห่งชาติ ที่มอบทุนอุดหนุนโครงการ การแข่งขันพัฒนาโปรแกรมคอมพิวเตอร์แห่งประเทศไทย ครั้งที่ 11 โครงการตัดคำภาษาไทย

สุดท้ายนี้คณะผู้จัดทำโครงการหวังเป็นอย่างยิ่งว่า โครงงานวิจัย “การเพิ่มประสิทธิภาพการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ด้วยเครื่อง” (Improvement of using machine learning techniques) จะเป็นประโยชน์ต่อการนำไปพัฒนาและประยุกต์ใช้ กับงานด้านประมวลผลภาษาต่อไป

คณะผู้จัดทำโครงการ

บทคัดย่อ

โครงงานวิจัย “การเพิ่มประสิทธิภาพการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ด้วยเครื่อง” ได้นำเสนอ กระบวนการตัดคำจากเอกสารภาษาไทย โดยใช้เทคนิคการเรียนรู้ด้วยเครื่องเข้ามาช่วยในการรู้จำลักษณะของคำในภาษาไทย ซึ่งได้ทำการประยุกต์วิธีของกราฟโมเดล (Graphical Model) มาใช้ ได้แก่ คอนดิชั่นนอล แรนดอมฟิลด์ (Conditional Random Fields :CRFs) เนื่องจากมีความเหมาะสมและสอดคล้องกับลักษณะของปัญหา โดยการเพิ่มประสิทธิภาพการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ด้วยเครื่องนั้น แบ่งได้ออกเป็นสองส่วน ได้แก่ ส่วนของการสร้างโมเดล และส่วนของการตัดคำภาษาไทย ซึ่งประสิทธิภาพของระบบแสดง ค่าความถูกต้อง (Recall) เท่ากับ 87.14% และ ค่าความแม่นยำ (Precision) เท่ากับ 85.92% และค่าความเหวี่ยง (F-Measure) เท่ากับ 86.52% ซึ่งแนวทางในการพัฒนาต่อไป จะมีการปรับปรุง และ ประยุกต์ความรู้ด้านการคัดเลือกคุณบัติตัวแทน (Feature Selection) ตลอดจนความรู้ด้านการประมวลผลภาษาอื่นๆเข้ามาพิจารณาเพื่อเพิ่มประสิทธิภาพของระบบการตัดคำภาษาไทยต่อไป

This Thesis: “Improvement of using machine learning techniques” purpose Thai Word Segmentation Process with Machine Learning Techniques. It’s help to recognize the Thai word. We applied Graphical Model which is Conditional Random fields (CRFs) because it’s suitable for problem. The system composes of two parts, model and segmentation. The result shows that recall is 87.14%, precision 85.92% and F-measure 86.52% precision and F-measure. For future work, we plan to combine feature selection and natural language processing techniques to improve the performance of Thai Word Segmentation in the next step.

บทนำ

ปัจจุบันงานทางด้านการประมวลผลทางภาษาธรรมชาติ (Natural Language Processing) ได้เข้ามามีบทบาทในด้านการประมวลผลสารสนเทศเป็นอย่างมากไม่ว่าจะเป็นงานทางด้านการสืบค้นสารสนเทศ (Information Retrieval) หรืองานทางด้านการสกัดสารสนเทศ (Information Extraction) เพื่อรองรับข้อมูลที่มีปริมาณมากขึ้น ตลอดจนการเข้าถึงข้อมูลได้อย่างถูกต้องแม่นยำ โดยส่วนใหญ่ข้อมูลจะอยู่ในรูปแบบของข้อความในเอกสาร การประมวลผลข้อมูลต้องอาศัยความรู้ทางด้านภาษาศาตร์เข้ามาช่วย โดยเฉพาะในภาษาไทย ลักษณะของข้อความ รูปประโยค และ คำในภาษาไทย มีความเฉพาะเจาะจง จำเป็นต้องอาศัยการเตรียมข้อมูลที่มีประสิทธิภาพ ซึ่งหนึ่งในขั้นตอนที่สำคัญในการเตรียมข้อมูลสำหรับภาษาไทยก็คือ การตัดคำ

การตัดคำ [1,2,3] เป็นการหาขอบคำของคำที่ถูกต้อง เนื่องจากลักษณะของภาษาไทย ไม่มีกฎเกณฑ์ที่แน่นอนในการระบุขอบเขตของคำที่แน่นอนทำให้ การเว้นวรรค หรือการตัดคำไม่มีกฎเกณฑ์ตายตัวในการหาขอบเขตคำที่ถูกต้อง ดังนั้นการตัดคำจึงกลายเป็นปัญหาหนึ่งที่ได้รับความสนใจจากนักวิจัยในการพัฒนาเทคนิค ตลอดจนวิธีการที่เข้ามาช่วยในการตัดคำภาษาไทย เช่น SWATH, LEXTON, G2P/Romanize แต่เดิมการตัดคำหากเป็นการตัดคำข้อความจากเอกสารที่มีปริมาณไม่มาก ทำให้สามารถทำการตัดคำด้วยมือได้ แต่ในความเป็นจริงแล้ว เอกสารในปัจจุบันมีปริมาณเพิ่มมากขึ้น ไม่ว่าจะเป็นเอกสารทางด้านบทความ เอกสารบนเว็บไซต์ ฯลฯ ทำให้ การตัดคำด้วยมือไม่สามารถทำได้อันเนื่องมาจากข้อจำกัดทางด้านเวลาและแรงงาน ดังนั้นการใช้คอมพิวเตอร์เข้ามาช่วยในการประมวลผลจึงจำเป็นอย่างยิ่งที่จะต้องอาศัยความชาญฉลาดในการวิเคราะห์และ ประมวลผลตัดคำได้อย่างถูกต้อง ยิ่งไปกว่านั้น การตัดคำภาษาไทยยังเป็นเทคนิคที่ถูกนำไปใช้เพื่อประยุกต์กับงานเฉพาะด้านต่างๆมากมาย ไม่ว่าจะเป็น งานทางด้านการสืบค้นข้อมูลสรรสาร (Sansarn Look) , การแปลงระหว่างเสียงพูดและข้อความ (Vaja) และ การบริการถามตอบข้อมูลด้วยโปรแกรม เอ็มเอสเอ็น: อับดุล (ABDUL) [4] เป็นต้นแต่การตัดคำภาษาไทยในปัจจุบัน ยังคงมีการพัฒนาอย่างต่อเนื่อง เพราะ ปัญหาของการตัดคำภาษาไทยยังคงมีอยู่มากมาย ทั้งนี้ สาเหตุของปัญหาคือในบางครั้งมีการเกิดคำที่ไม่รู้จัก คอมพิวเตอร์ไม่เคยรู้จัก หรือเรียนรู้มาก่อน (Machine learning) รวมถึงคำที่เป็นชื่อเฉพาะทำให้การตัดคำเกิดการผิดพลาดขึ้น [2,3,4]

ด้วยเหตุนี้ทำให้ทีมพัฒนาเกิดแนวคิดในการการเพิ่มประสิทธิภาพการตัดคำภาษาไทยกด้วยเทคนิคการเรียนรู้ด้วยเครื่อง ซึ่งมีวิธีการมากมาย อันจะนำมาใช้เพื่อเพิ่มความสามารถในการวิเคราะห์ และ หาขอบเขตของคำได้อย่างถูกต้องแม่นยำ เช่น วิธีคอนดิชันนอลแรนดอมฟิลด์ (Conditional Random Fields Model) [5,6] ซึ่งเป็นวิธีการของกราฟโมเดลเข้ามาช่วยในการสร้างโมเดลที่รู้จำลักษณะของการตัดคำในภาษาไทย ตลอดจนวิธีการเตรียมข้อมูลต่างๆ (Preprocessing) ที่จะเข้ามาช่วยเพื่อทำให้การตัดคำภาษาไทยให้มีประสิทธิภาพมากยิ่งขึ้น

สารบัญ

วัตถุประสงค์ของโครงงาน

1. เพื่อทำการปรับปรุงประสิทธิภาพของวิธีการตัดคำภาษาไทย ให้มีประสิทธิภาพเพิ่มขึ้น ในด้านของความถูกต้องและแม่นยำ

2. เพื่อสร้างโปรแกรมจากโมเดลทางการคำนวณจากเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการตัดคำภาษาไทย ทั้งนี้เพื่อรองรับปริมาณของข้อความที่มีมากขึ้นในปัจจุบัน

3. เพื่อพัฒนาความสัมพันธ์ระหว่างคนและคอมพิวเตอร์ ในการจัดการการประมวลผลทางภาษา

ปัญหาหรือประโยชน์ที่เป็นเหตุผลให้ควรพัฒนาโปรแกรม

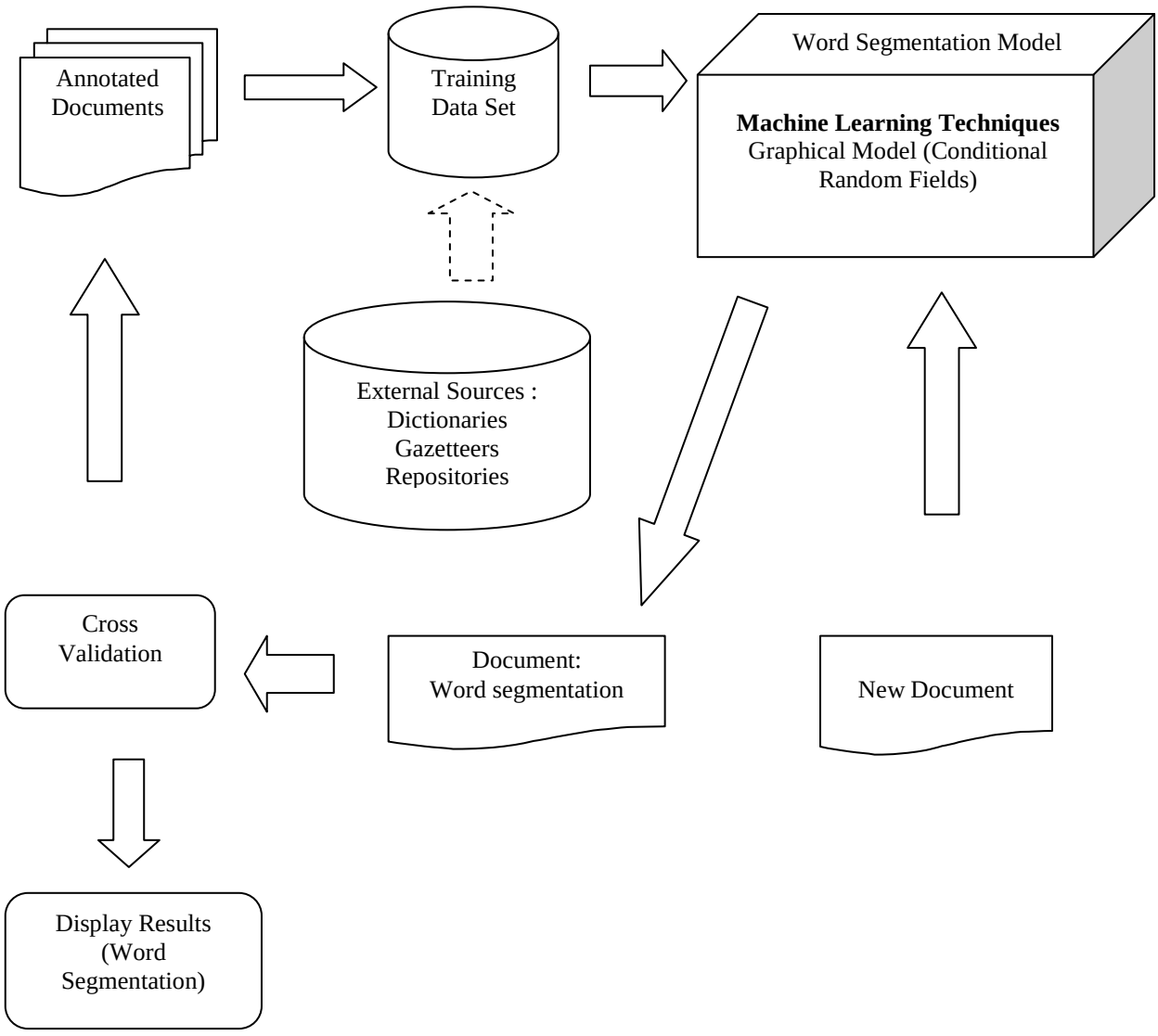
เนื่องจากการตัดคำเป็นเทคนิคพื้นฐานที่มีความจำเป็นต่อการประมวลผลข้อความในระดับงานที่สูง เช่น การกำหนดหน้าที่ของคำ (Part-Of-Speech Tagging) การแปลภาษาโดยเครื่อง (Machine Translation) การจดจำและสังเคราะห์เสียงพูด (Speech Recognition) การค้นคืนเอกสารและระบบสืบค้นข้อมูล (Information Retrieval)ตลอดจน การทำเหมืองข้อความ (Text Mining) รวมถึงการสร้างฐานความรู้และโครงสร้างความรู้เชิงความหมาย (Knowledge Base and Ontology) ทั้งนี้ประสิทธิภาพของงานต่างๆขึ้นอยู่กับความถูกต้องของการตัดคำ เพราะหากตัดคำไม่ถูกต้องแล้วความหมายของคำก็อาจจะเปลี่ยนไป และในปัจจุบัน ยังคงมีการพัฒนาเทคนิค โปรแกรมต่างๆเพื่อใช้ในการตัดคำภาษาไทยทั้งนี้มาจากปัญหาการตัดคำได้แก่ คำกำกวม สามารถให้ความหมายหรือตัดคำได้มากกว่า 1 อย่าง และ คำที่ไม่รู้จัก ซึ่งอาจจะหมายถึงคำที่เป็นชื่อเฉพาะ หรือคำที่เกิดขึ้นมาใหม่ ดังนั้น การพัฒนาโมเดลเพื่อช่วยสร้างโปรแกรมในการตัดคำในภาษาไทยด้วยการเรียนรู้ด้วยเครื่อง จึงเป็นกระบวนการที่สำคัญและเป็นประโยชน์ต่องานทางด้านการประมวลผลข้อมูลอื่นๆเป็นอย่างมาก

เป้าหมายและขอบเขตของโครงการ

โครงการตัดคำภาษาไทยนี้มีเป้าหมายในการสร้างโมเดลวิธีการตัดคำภาษาไทย ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยมีการนำเอาแนวคิดทางด้านกราฟโมเดลมาช่วยในการแก้ปัญหาตัดคำ โดยขอบเขตของโครงการสามารถแบ่งได้ตามหัวข้อดังนี้

เนื้อเรื่องย่อ (Story Board) ภาพประกอบ แบบจำลอง ทฤษฎีที่เกี่ยวข้อง

โครงการการเพิ่มประสิทธิภาพการตัดคำด้วยเทคนิคการเรียนรู้ด้วยเครื่องนั้น ต้องการปรับปรุงเทคนิคการตัดคำภาษาไทยให้มีประสิทธิภาพมากยิ่งขึ้น โดยการสร้างโมเดลที่ประยุกต์จากวิธีการทางกราฟโมเดล (CRFs) ตลอดจนเทคนิคการเรียนรู้ด้วยเครื่องต่างๆ โดยระบบงานสามารถแสดงได้ดังรูปที่ 1



รูปที่ 1 แบบจำลองการเพิ่มประสิทธิภาพการตัดคำด้วยเทคนิคการเรียนรู้ด้วยเครื่อง

การตัดคำเป็นวิธีการที่สำคัญที่นำมาใช้เป็นขั้นตอนพื้นฐานของการประมวลผลข้อมูลที่มีลักษณะเป็นข้อความ การตัดคำต้องใช้วิธีการและความรู้ทางด้านภาษาเข้ามาช่วย สำหรับปัญหาการตัดคำได้ถูกนำไปใช้งานในหลายๆภาษา ตามแต่กฎเกณฑ์และไวยากรณ์ทางภาษา ในภาษาอังกฤษ มีหลักการแบ่งคำด้วยช่องว่าง (Space) ซึ่งทำให้ง่ายต่อการตัดคำ ปัญหาจึงถูกมองไปในเรื่องของการระบุชื่อเฉพาะ (Named

Entity Recognition) มากกว่า ภาษาจีน ภาษาญี่ปุ่น ก็จะมีลักษณะที่แตกต่างกัน เช่นเดียวกันภาษาไทย ก็มีลักษณะที่ มีความซับซ้อนในลักษณะของประโยค การประสมคำ คำหนึ่งคำมีได้หลายความหมาย หรือ คำที่ใกล้เคียงกัน รวมถึงการเขียน การเว้นวรรค ก็ไม่ได้มีหลักเกณฑ์ที่แน่นอน จึงทำให้การตัดคำภาษาไทยมีความน่าสนใจ และ เป็นปัญหาที่มีนักวิจัยพยายามคิดค้นวิธีการตัดคำมากมาย จากงานวิจัยที่ทำการเปรียบเทียบวิธีการตัดคำที่มีอยู่ในปัจจุบัน [2] พบว่าวิธีการตัดคำแบ่งได้เป็น 2 ประเภทด้วยกัน คือ วิธีการตัดคำโดยใช้พจนานุกรม (Dictionary Based) และ การตัดคำโดยใช้วิธีการเรียนรู้ด้วยเครื่องซึ่งใช้วิธีการทางสถิติ (Machine Learning Based) ซึ่งวิธีการใช้พจนานุกรมก็มีเทคนิคมากมายที่ใช้วิธีการนี้ ได้แก่การตัดคำให้ได้ความหมาย (longest matching and maximal matching) ส่วนเทคนิควิธีการเรียนรู้ด้วยเครื่อง ได้แก่ วิธีการทางสถิติ ความน่าจะเป็น (Naïve Bayes)โครงสร้างการตัดสินใจแบบต้นไม้ (Decision Tree) และหลักการทางคณิตศาสตร์ (Support Vector Machine) และ การใช้กราฟโมเดล (Conditional Random Fields :CRFs) ผลการทดสอบพบว่า วิธีการตัดคำภาษาไทยโดยพจนานุกรมมีประสิทธิภาพมากกว่าวิธีการเรียนรู้ด้วยเครื่องแบบ Naïve Bayes Decision Tree และ Support Vector Machine แต่วิธีการที่มีประสิทธิภาพมากที่สุดในขณะนี้คือ CRFs คือให้ค่าความถูกต้องแม่นยำ (Recall and Precision) อยู่ที่ 95.75% และ 94.98% วิธีการตัดคำแต่ละแบบมีข้อดีและข้อจัดที่แตกต่างกันไป เช่น การตัดคำโดยใช้พจนานุกรมมีข้อดีคือมีความรวดเร็วและง่าย ความถูกต้องสูง สามารถตัดคำในลักษณะคำประสมได้ แต่ข้อจำกัดก็คือ ในปัจจุบันยังไม่มีพจนานุกรมที่ประกอบด้วยคำทุกคำครบถ้วน เนื่องจากมีคำใหม่เกิดขึ้นอยู่เสมอ โดยเฉพาะคำที่เป็นชื่อเฉพาะ ส่วนวิธีการเรียนรู้ด้วยเครื่องนั้นจะช่วยในการรู้จำคำ โดยลักษณะของโมเดล จะดีหรือไม่นั้น ขึ้นอยู่กับ ชุดข้อมูลที่ใช้รู้จำ (Training data set)

โครงการเพิ่มประสิทธิภาพการตัดคำด้วยวิธีเทคนิคการเรียนรู้ด้วยเครื่อง โดยอาศัยเทคนิคกราฟโมเดล เข้ามาช่วยในการหารูปแบบ (Pattern) ของคำเพื่อทำการตัดคำ โดยการปรับปรุงและเพิ่มประสิทธิภาพของกฎในการกำหนดลักษณะของการหารูปแบบของคำ โดยอาจจะมีการประยุกต์ใช้เทคนิคการตัดคำอื่นๆ ตลอดจนวิธีการเตรียมข้อมูลต่างๆเข้ามาช่วยในการประมวลผลเพื่อให้ได้ประสิทธิภาพในการตัดคำที่เพิ่มมากขึ้น ดังนี้

ทฤษฎีที่เกี่ยวข้อง หลักการและเทคนิคหรือเทคโนโลยีที่ใช้

กราฟโมเดลแบบไม่มีทิศทาง (Undirected Graphical Model)

ถูกนำมาใช้ในการนำเสนอทฤษฎีทางด้านความน่าจะเป็น และ การเรียนรู้ของเครื่อง แสดงถึงความสัมพันธ์ระหว่าง ตัวแปรสุ่ม (Random
Variables) ซึ่ง กราฟโมเดลสามารถแบ่งได้เป็น 2 ประเภท คือ กราฟแบบมีทิศทางและ ไม่มีทิศทาง ซึ่ง ในที่นี้ จะเน้นวิธีการใช้กราฟโมเดลแบบไม่มีทิศทาง (Undirected Graph) เนื่องจากเป็นเทคนิคที่มีลักษณะเป็น

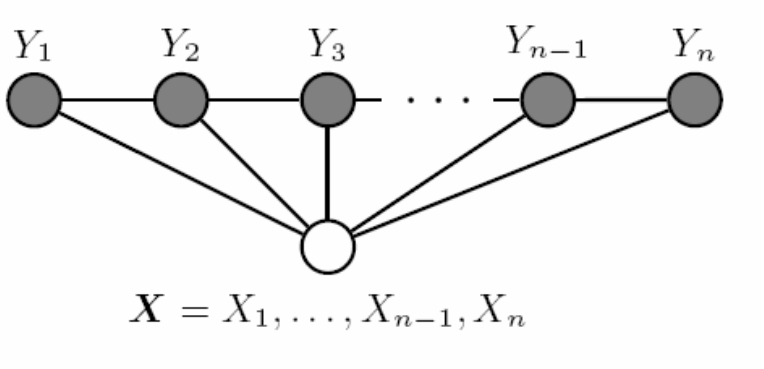
แบบจำลองในการแยกแยะโดยอาศัยความน่าจะเป็น (Discriminative Probabilistic Model) โดยสารมารถคำนวณหาค่าความน่าจะเป็นได้จากสูตร ดังนี้

$$p(\vec{v}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(\vec{v}_C) \tag{1}$$

ซึ่งเป็น Potential Function โดยหาได้จากผลคูณจากค่าที่ได้จากทุกๆ cliques และ Z คือ Normalize Term เป็นส่วนที่สำคัญ ต้องประมาณค่า ทั้งนี้เพื่อทำให้ค่าที่ได้ออกมาอยู่ระว่าง 0 ถึง 1 เนื่องจาเป็นค่าความน่าจะเป็น

Conditional Random Fields (CRFs)

CRFs เป็นวิธีการหนึ่งที่ใช้เทคนิคของกราฟไม่มีทิศทาง [5,6] ซึ่งเป็นวิธีที่ถูกใช้อย่างมากในงานที่เกี่ยวข้องกับการระบุหน้าที่ของคำ (Sequence Labeling) หรือการวิเคราะห์ลำดับข้อมูล (Parse) เนื่องจากมีความเหมาะสมกับปัญหา มีความยืดหยุ่นในเรื่องของการดูบริบทรอบๆข้าง [A. McCallum and C. Sutton] สามารถแก้ปัญหา “Long distance problem” ซึ่งแบบจำลองของ CRFs แสดงได้ดังรูปที่ 2



รูปที่ 2 แบบกราฟจำลองแนวคิดของ Conditional Random field

โดย Potential function ของ CRFs สามารถเขียนได้ดังสมการ

$$\exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right) \tag{2}$$

ซึ่ง $t_j(y_{i-1}, y_i, x, i)$ แทนTransition Function ของการเกิดข้อความ เช่น

$$t_j(y_{i-1}, y_i, x, i) = \begin{cases} b(x, i) & \text{if } y_{i-1} = IN \text{ and } y_i = NNP \\ 0 & \text{otherwise.} \end{cases}$$

โดยที่ b(x,i) เป็น Function ในการคืนค่าเป็น 1 และ 0 เท่านั้น โดยจะเป็น 1 เมื่อ พบอักขระ (x) ในข้อความ

ส่วน $s_k(y_i, x, i)$ แทน Feature function ที่กำหนดลักษณะของการเกิด Pattern ของอักรขะ ที่ต่อกันเป็นคำการคำนวณหาความน่าจะเป็นสามารถทำการหาได้จากสมการ

$$p(y|x,\lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j F_j(y,x)\right) \quad (3)$$

โดยที่เทอมของ

$$F_j(y,x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)$$

แทน Feature function ที่กำหนดลักษณะของการเกิด Pattern ของอักรขะ ที่ต่อกันเป็นคำ

และ λ_j หมายถึงพารามิเตอร์ (Parameter) เป็นตัวประมาณค่าที่ทำให้โมเดลมีประสิทธิภาพ ในขณะที่ Z(x) คือ Normalize Term เป็นส่วนที่สำคัญ ต้องประมาณค่า ทั้งนี้เพื่อทำให้ค่าที่ได้ออกมาอยู่ระว่าง 0 ถึง 1 เนื่องจากเป็นค่าความน่าจะเป็น จากสมการที่ (3) จะคำนวณหาpattern ของคำที่ให้ค่าความน่าจะเป็นที่สูงที่สุด โดยจะมีการคำนวณจากทุกๆเส้นทางทำให้ คำตอบที่ได้จะอยู่ในรูปแบบที่ดีที่สุด (Global optimization) เนื่องจากต้องหาทุกๆเส้นทางแล้วทำการเปรียบเทียบว่าค่าใดให้ความน่าจะเป็นมากที่สุด แล้วจึงได้คำตอบเป็น Pattern นั้นๆ

เครื่องมือที่ใช้ในการพัฒนา

1. ระบบปฏิบัติการ ไมโครซอฟท์วินโดวส์เอ็กซ์พี โปรเฟสชันแนล
2. ระบบปฏิบัติการยูนิกซ์ (UNIX)
3. Turbo C++ 4.5
4. Edit plus 3
5. CRF++

รายละเอียดโปรแกรมที่ได้พัฒนาในเชิงเทคนิค (Software Specification)

Input / Output Specification:

Input for Training: เอกสารข้อความที่ประกอบด้วยอักขระต่างๆ ทำการแปลงให้อยู่ในรูปของโครงสร้างที่มีการระบุชนิดของอักขระ ดังรูปที่ 3

เ	W	B
ก	C	I
ษ	C	I
ต	C	I
ร	C	I
ก	C	I
ร	C	I
ป	C	B
ล	C	I
ู	W	I
ก	C	I
ข	C	B
้	V	I
า	W	I
ว	C	I

รูปที่ 3 การระบุชนิดของอักขระของคำ

ใน Column ที่ 1 คือ อักขระแต่ละตัว

Column ที่ 2 คือ ชนิดของอักขระ

Column ที่ 3 คือ การบอกขอบเขตของคำ B คือ บอกจุดเริ่มต้นของคำ ส่วน I เป็นการบอกว่าเป็นตัวที่ตามมาไม่ใช่ตัวแรกของคำ

โดย ข้อมูลสำหรับใช้ในการสร้างโมเดลจะได้จาก Corpus ของ Best 2009 จากเว็บไซต์ (http://www.hlt.nectec.or.th/best/)

Input for Program: เอกสารข้อความที่ประกอบด้วยอักขระต่างๆ เช่น คำว่า เกษตรกรปลูกข้าว

Output Program: ได้คำที่ตัดแล้ว เช่น เกษตรกร|ปลูก|ข้าว โดย ถ้าหากคำใด เป็นชื่อเฉพาะ หรือ คำย่อ ก็อาจจะทำการ Label คำนั้น เช่น

ชาว|ยุโรป|เริ่ม|เข้า|มา|ติดต่อ|กับ|*ไทย*|อีก|ครั้ง|หนึ่ง|ใน|รัชสมัย|*พระบาทสมเด็จพระพุทธเลิศล้านภาลัย*
วัน|ที่|27| *พ.ค.*|

เช่นให้การ Label สีเขียว เป็นสถานที่ และ สีเหลืองคือ บุคคล สีแดงหมายถึงคำย่อ

Functional Specification

โปรแกรมประกอบไปด้วย ส่วนของการทำงาน ดังรูปที่ 1 คือประกอบด้วย Function ดังนี้

ฟังก์ชันในการรับ Input

ฟังก์ชันในการปรับปรุง โมเดล (Update Training data set)

ฟังก์ชันในการตัดคำ

ฟังก์ชันในการเลือก Label ชนิดของคำ

ฟังก์ชันในการตรวจสอบความถูกต้อง

ฟังก์ชันในการแสดงผลคำตอบ (คำที่ตัดแล้ว)

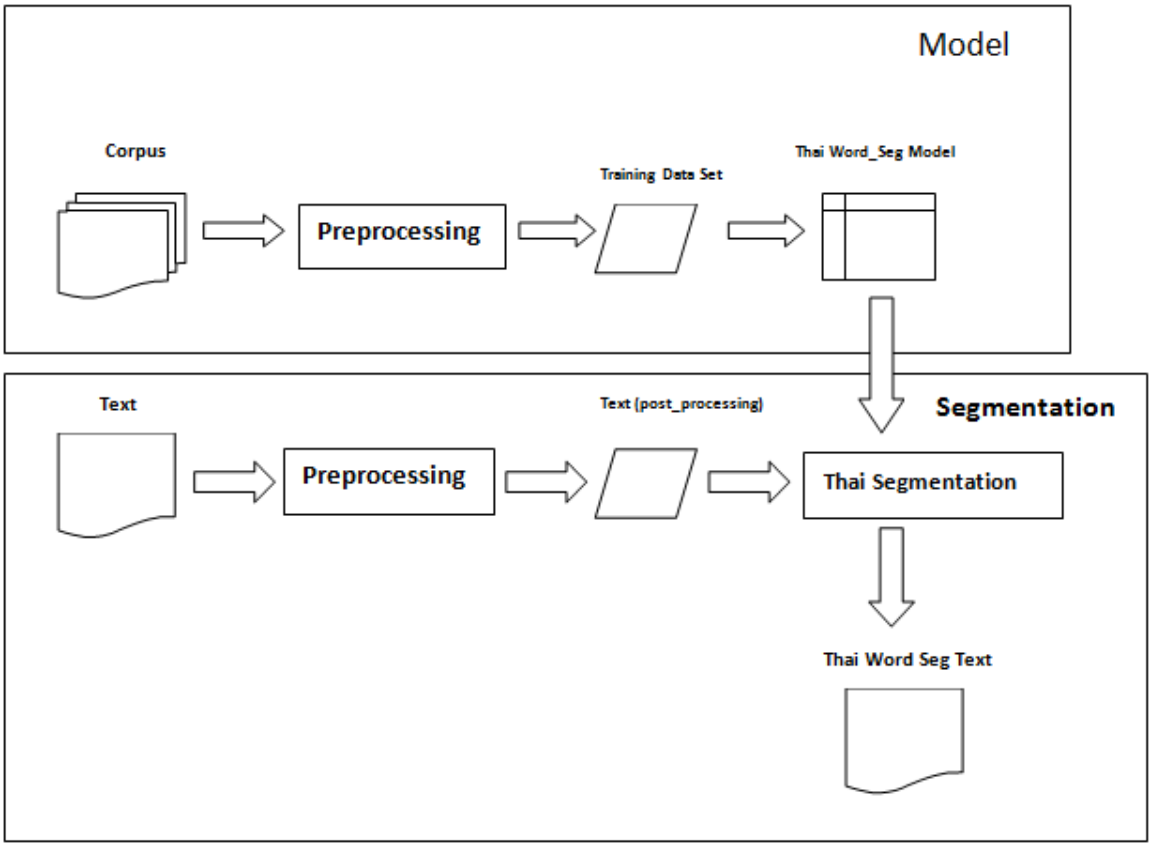
สำหรับการปรับปรุงโมเดลจำเป็นจะต้องสกัด อักขระให้อยู่ดังรูปที่ 3 ซึ่งอาศัยหลักการจัดประเภทของอักขระดังรูปที่ 4

Tag	Type	Example
c	Character that can be the final consonant in a word	กขฃคฅงจชซฌญฎฏฐฑฒฌณดตถทธนบปพฟภมยรลวศษสห ฬ อ
n	Character that cannot be the final consonant in a word	ฅ ฉ ผ ฝ ฌ ห ฮ
v	Vowel that cannot begin a word	ะ ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู ◌ั า ำ ๅ
w	Vowel that can begin a word	เ แ โ ใ ไ
t	Tonal character	◌่ ◌้ ◌๊ ◌๋
s	Symbol	ฯ ๆ ฯ.
d	Digit character	0-9
q	Quote character	'-' "-"
p	Space character inside a word	_
o	Other character	A-z

รูปที่ 4 ตารางประเภทของอักขระ

โครงสร้างของซอฟต์แวร์ (Design)

สำหรับโครงสร้างซอฟต์แวร์ (Design) ของระบบการตัดคำภาษาไทย จะทำการนำเสนอในรูปแบบของแผนภาพเพื่อให้ง่ายต่อความเข้าใจดังรูป



รูปที่ 5 โครงสร้างซอฟต์แวร์ (Design)

ขอบเขตและข้อจำกัดของโปรแกรมที่พัฒนา

ในการพัฒนาจะเป็นการเน้นเฉพาะการตัดคำภาษาไทย ที่มีข้อความเป็น Text (เฉพาะอักขระ) เท่านั้น โดยใช้เทคนิคการเรียนรู้ด้วยเครื่อง ซึ่งจะใช้ CRFs ในการสร้างโมเดล เพื่อการตัดคำในภาษาไทยจากชุดข้อมูลที่ได้มา

กลุ่มผู้ใช้โปรแกรม

กลุ่มผู้ใช้งานทั่วไปที่ต้องศึกษาหรือต้องการประยุกต์ใช้ การตัดคำภาษาไทยในการประมวลผล เช่น ผู้ที่ต้องการทำ Web search engine จะต้องการตัดคำภาษาไทยในส่วนของการเตรียมข้อมูลก่อนนำคำที่ตัดได้ไปประมวลผลและกลุ่มผู้ใช้งานที่ต้องการตัดคำภาษาไทยแต่ละหน้า Web เพื่อให้การขึ้นบรรทัดใหม่ของกลุ่มประโยค ไม่ให้มีการตัดคำที่ผิด และเว็บแปลคำศัพท์ภาษาไทยบนหน้า Web page ที่ต้องการแปลแบบอัตโนมัติ เป็นต้น

ผลของการทดสอบโปรแกรม

เพื่อทำการทดสอบประสิทธิภาพของระบบตัดคำภาษาไทย ผู้จัดทำโครงการได้ทำการสร้างชุดเอกสารส่วนที่รู้จำเพื่อใช้ในการสร้างโมเดล (Training Data Set) โดยเลือกมาจาก BEST Corpus training set 1 (Release 3) ประกอบด้วยไฟล์ 38 ไฟล์ ซึ่งมาจากหลายๆประเภทเอกสารได้แก่ : (24 news files, 5 encyclopedia files และ 9 novel files) รวมแล้วประกอบด้วยคำ 449,735 คำ, ในขณะที่ ชุดทดสอบ (Test set) ประกอบด้วยคำ 51,057 คำโดยเลือกจาก ไฟล์ที่มาจากหลายๆประเภท (news, encyclopedia และ novel) มาทำการทดสอบเพื่อทำการตรวจสอบประสิทธิภาพ (Cross Validation) ในการวัดประสิทธิภาพ ได้ใช้ มาตรฐานของค่าในการวัดความถูกต้องดังนี้

ค่าความถูกต้อง (Recall): หาได้จาก

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

ค่าความแม่นยำ (Precision): หาได้จาก

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

ค่าความเหวี่ยง (F-Measure): หาได้จาก

$$F = 2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall}).$$

โดย ค่า Relevant document: ในที่นี้หมายถึง จำนวนคำที่ตัดถูกต้อง

ค่า Retrieved document: หมายถึง จำนวนคำที่ตัดได้

โดยผลลัพธ์ที่ได้พบว่าระบบมีความถูกต้องในการตัดคำภาษาไทย ดังนี้

ค่าความถูกต้อง (Recall) เท่ากับ 87.14% และ ค่าความแม่นยำ (Precision) เท่ากับ 85.92% และค่าความเหวี่ยง (F-Measure) เท่ากับ 86.52% ดังตาราง

Corpus	ทั้งหมด (คำ)	ผลการทดสอบ		
		ค่าความถูกต้อง (Recall)	ค่าความแม่นยำ (Precision)	ค่าความเหวี่ยง (F-Measure)
BEST Corpus training set 1 (Release 3)	**Train:** (449,735) **Test:** **(51,057)**	**87.14**	**85.92**	**86.52**

ตาราง 1 แสดงผลการทดสอบประสิทธิภาพของระบบตัดคำภาษาไทย

ปัญหาและอุปสรรค

ปัญหาเบื้องต้นคือ ลักษณะการเขียนภาษาไทยจะเขียนติดต่อกันเป็นสายอักขระ โดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำดังเช่นภาษาอังกฤษ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่ต้องการการศึกษาวิจัยและพัฒนา เพื่อให้คอมพิวเตอร์สามารถคำนวณ เพื่อแบ่งสายอักขระไทยออกเป็นคำๆ ซึ่งจะส่งผลให้การทำงานของคอมพิวเตอร์ในการค้นหาคำใดๆ ทำได้อย่างถูกต้องและแม่นยำรวมถึงการจัดขอบขวาในโปรแกรมประมวลผลคำ(Word Processor) เป็นต้น

ส่วนปัญหาต่อมาคือปัญหาเรื่องเวลาในการแปลงไฟล์ ทำให้ต้องแบ่งออกเป็นไฟล์ย่อย ๆ หลังจากนั้น จึงนำมารวมเป็นไฟล์เดียวเพื่อนำไป train หรือ test อีกที จึงเสียเวลาในการทำขั้นตอนนี้มาก

แนวทางในการพัฒนาและประยุกต์ใช้ร่วมกับงานอื่นๆในขั้นต่อไป

เพื่อให้มีความถูกต้องที่มากขึ้น ควรจะเพิ่ม training set ที่ถูกต้องให้มากขึ้น ส่วนในการปรับ template นั้นควรมีการนำความรู้และประสบการณ์เกี่ยวกับความสัมพันธ์ของอักขระในคำภาษาไทยในการสร้าง template เพื่อให้มีการรู้จำของเครื่องมีประสิทธิภาพมากยิ่งขึ้น และการทำนายข้อมูลมีและสิทธิภาพมากยิ่งขึ้น

ข้อสรุปและข้อเสนอแนะ

โปรแกรมตัดคำภาษาไทยที่ได้ใช้เทคนิค Graphical model แบบ Conditional random field (CRF) นั้นได้ผลความถูกต้องได้ดีในระดับหนึ่ง เนื่องจากว่าเทคนิคนี้มีความยืดหยุ่นสูงและประยุกต์ใช้ได้อย่างดี เหมาะสมกับปัญหาที่เป็นสายอักขระ (Sequential Problem) กับการประมวลผลภาษาทางธรรมชาติ (National language processing) และวิธีการนี้ได้ใช้การเรียนรู้ของเครื่อง (Machine learning) โดยการ training ที่ได้รับการยอมรับว่าถูกต้องแล้วและการ training นี้เป็นไปตามแนวทางของ template ที่มีประสิทธิภาพ

เอกสารอ้างอิง (Reference)

[1] Wikipedia Foundation, Inc. “Word Segmentaion”. [Online. Available] Retrieved from http://en.wikipedia/wiki/Word Segmentation (20 July 2008).

[2] Choochart Haruechaiyasak, Sarawoot Kongyoung and MatthewN. Dailey, Comparative Study on Thai Word Segmentation Approaches, Human Language Technology Laboratory, National Electronics and Computer Technology Center, Pathumthani.

[3] Krisda Khankasikam and Nuttanart Muansuwan, 2005, Thai Word Segmentation a Lexical Semantic Approach, 2005, Department of Computer Engineering King Mongkut’s University of Technology Thonburi Bangkok,

[4] Human language Technology, NECTEC ,2007A Driving Force for National Science and Technology Capability, 2007, Chiang Mai University, Chiang Mai.

[5] Hanna M. Wallach. Conditional Random Fields: An Introduction. Technical Report MS-CIS-04-21. Department of Computer and Information Science, University of Pennsylvania, 2004.

[6] Andrew McCallum and Charles Sutton, 2007, An Introduction to Conditional Random Fields for Relational Learning, Department of Computer Science University of Massachusetts, USA.

[7] Hanna M.Wallach, 2004, Conditional Random Fields: An Introduction, University of Pennsylvania CIS Technical Report MS-CIS-04-21, February 24..

ภาคผนวก (Appendix)

คู่มือการติดตั้งโปรแกรม

คู่มือการติดตั้งโปรแกรม AppServ (PHP แอพพลิเคชั่นเซิร์ฟเวอร์)

โปรแกรม AppServ v2.4.4a Setup เป็นชุดติดตั้งโปรแกรม PHP แอพพลิเคชั่นเซิร์ฟเวอร์ สำหรับติดตั้งบนระบบปฏิบัติการ Windows โดยในชุดติดตั้ง AppServ นี้ ประกอบด้วยโปรแกรมต่าง ๆ ดังต่อไปนี้

1. Apache สำหรับทำหน้าที่เป็นเว็บเซิร์ฟเวอร์
2. PHP สำหรับทำหน้าที่เป็นตัวแปรภาษา PHP
3. MySQL สำหรับทำหน้าที่เป็นดาต้าเบสเซิร์ฟเวอร์

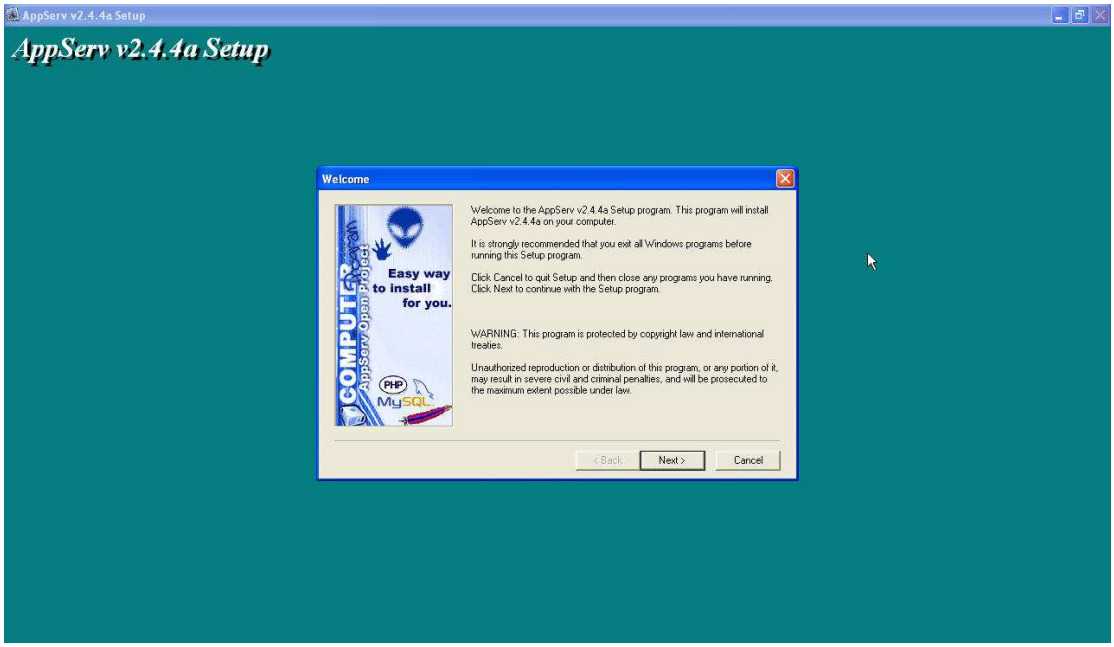
เนื้อหาในส่วนนี้จะกล่าวถึงวิธีการติดตั้งโปรแกรม AppServ (PHP แอพพลิเคชั่นเซิร์ฟเวอร์) ซึ่งสามารถดาวน์โหลดได้ที่เว็บไซต์ http://www.appservnetwork.com/

โดยก่อนทำการติดตั้ง ให้ตรวจสอบภายในเครื่องก่อนว่าได้มีการติดตั้งโปรแกรม AppServ เอาไว้ในเครื่องหรือไม่ โดยสามารถดูได้จาก start->Programs แล้วดูว่ามีโปรแกรม AppServ หรือไม่ ถ้ามีให้ทำการ Uninstall ออกก่อน เพราะอาจเกิดข้อผิดพลาดในการทำงานของโปรแกรมได้

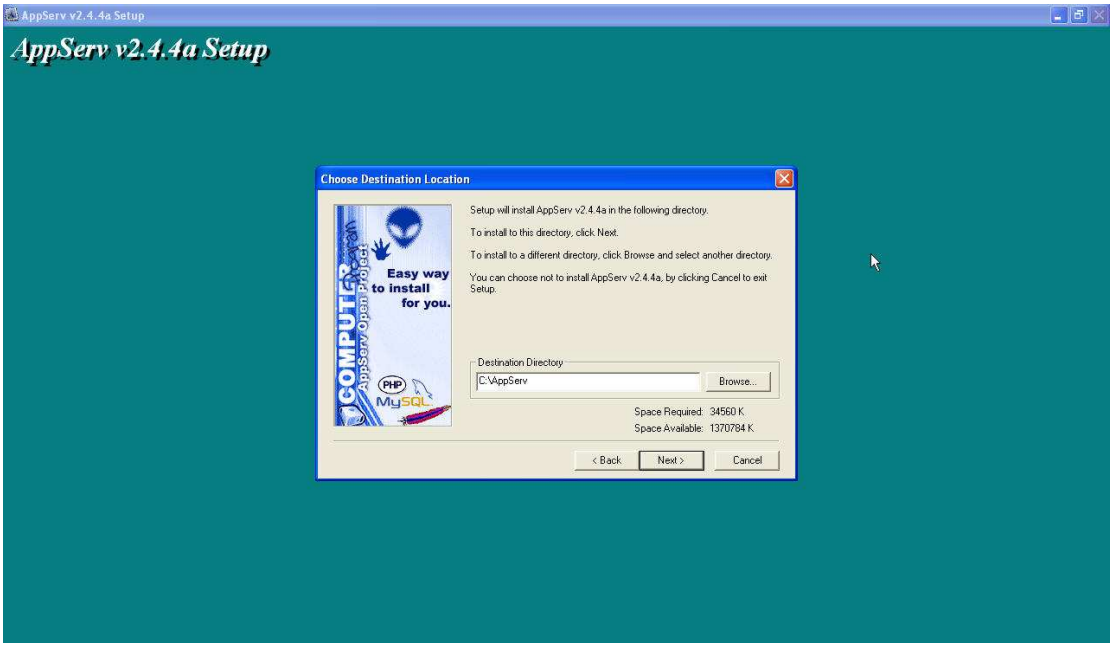
<u>การติดตั้งโปรแกรม AppServ (PHP แอพพลิเคชั่นเซิร์ฟเวอร์)</u>

ก่อนการติดตั้ง จะตั้งมั่นใจก่อนว่าขณะนั้น ผู้ติดตั้งมีสิทธิในการติดตั้งโปรแกรมในเครื่องคอมพิวเตอร์ (มีสิทธิ์เทียบเท่ากับ Administrator)

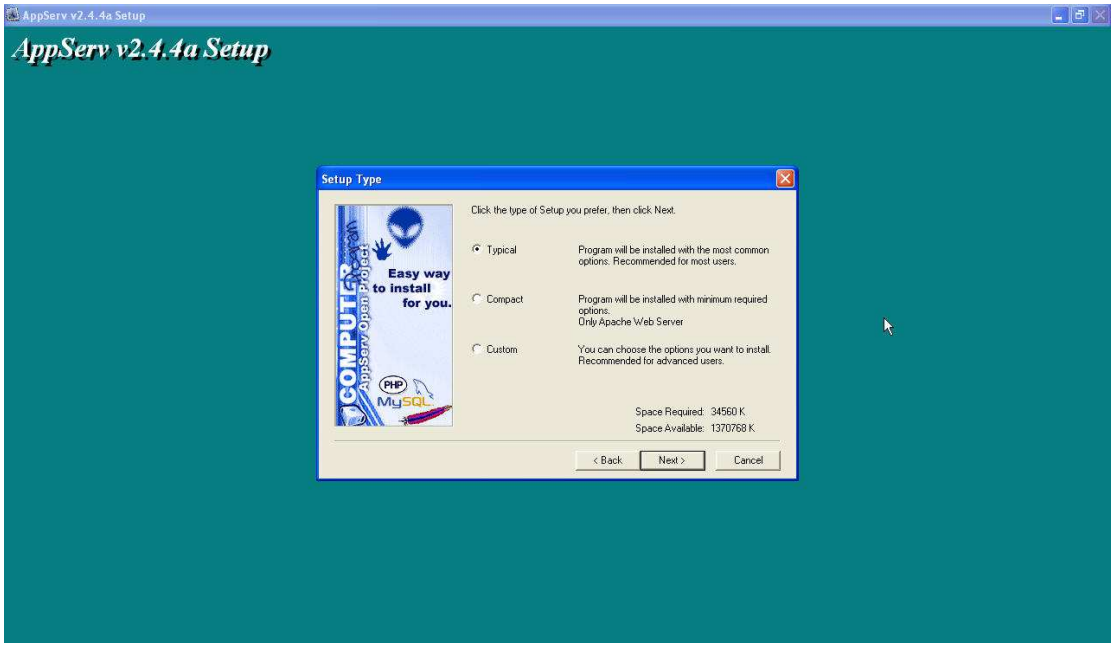
1. ให้เปิด Folder ที่ชื่อ Step_1 ที่อยู่บนแผ่นซีดี จะเห็นไฟล์ที่ชื่อ appserv-win32-2.4.4a.exe สั่ง run โดยกดดับเบิลคลิกที่ชื่อไฟล์ จากนั้นจะปรากฏหน้าจอ ดังรูป



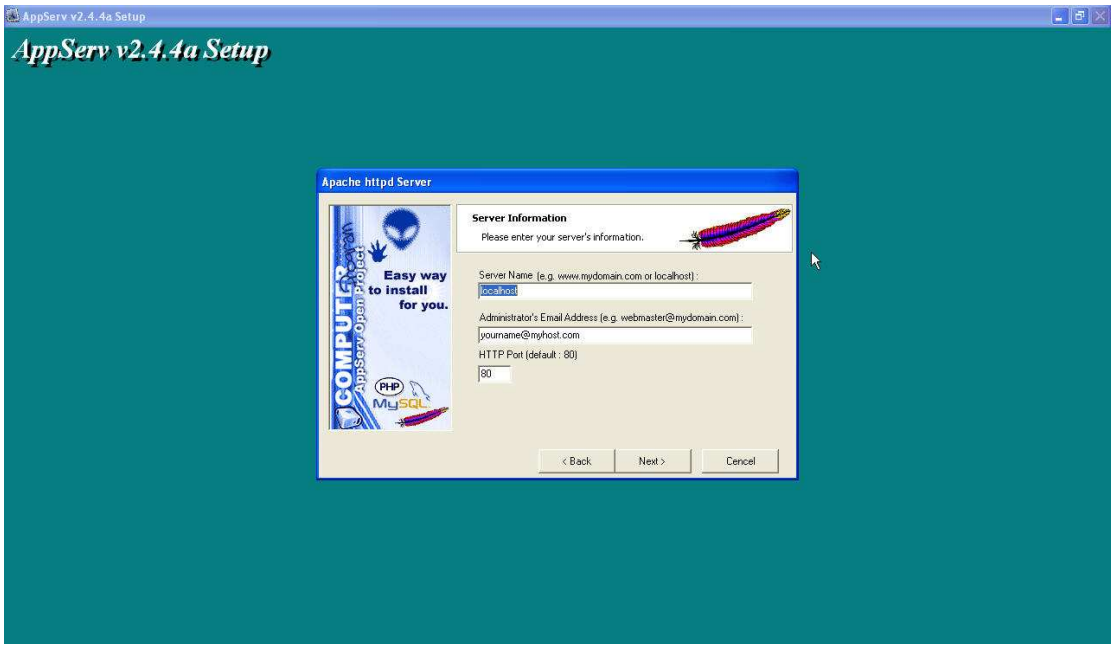
2. จากนั้นที่หน้าต่าง Welcome ให้กดปุ่ม Next จะปรากฏหน้าจอ ดังรูป



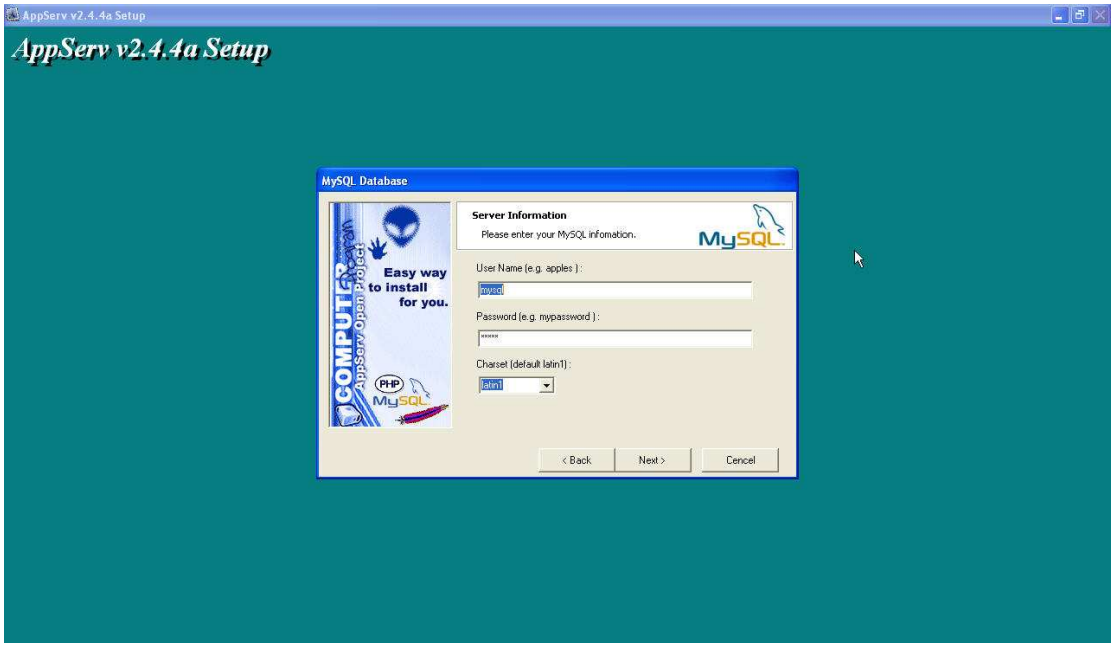
3. จากนั้นที่หน้าต่าง Choose Destination Location เป็นการกำหนดไดเร็คทอรี่ที่จะติดตั้งโปรแกรม ซึ่งกำหนดค่าเริ่มต้นเป็น **C:\AppServ** ให้กดปุ่ม Next จะปรากฏหน้าจอ ดังรูป (ไม่แนะนำให้เปลี่ยนค่า path ตรงนี้ครับ เนื่องจากจะกระทบกับการตั้งค่าของ Web Server ที่กำหนดไว้ได้)



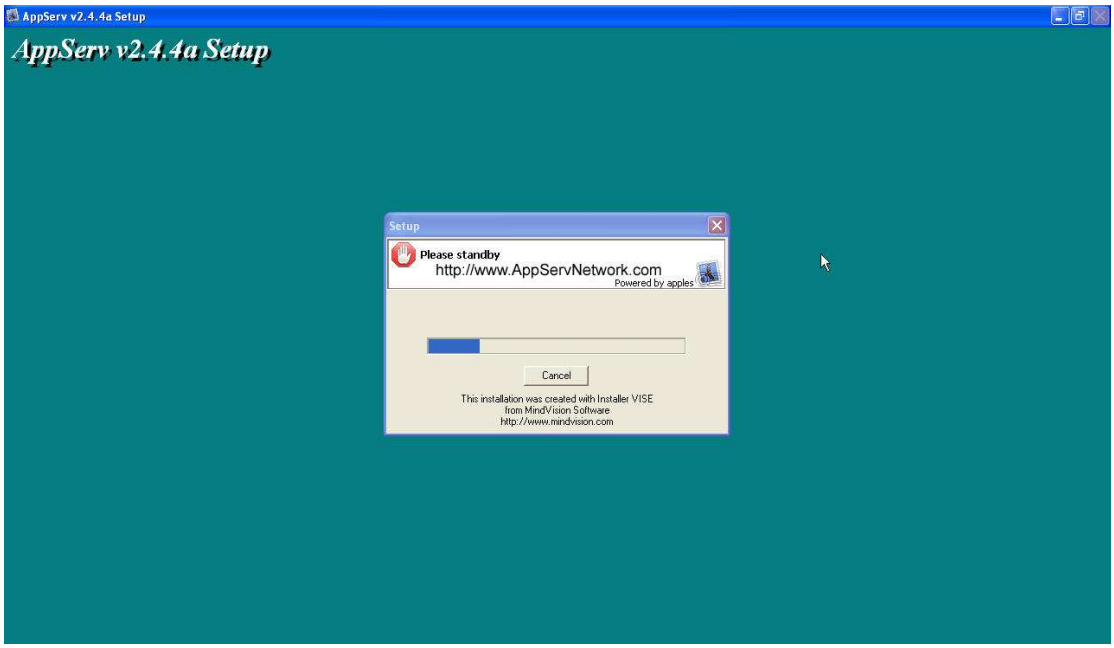
4. จากนั้นที่หน้าต่าง Setup Type จะเป็นการเลือกประเภทของการติดตั้ง ให้เลือกที่ Typical แล้วกดปุ่ม Next จะปรากฏหน้าจอ ดังรูป



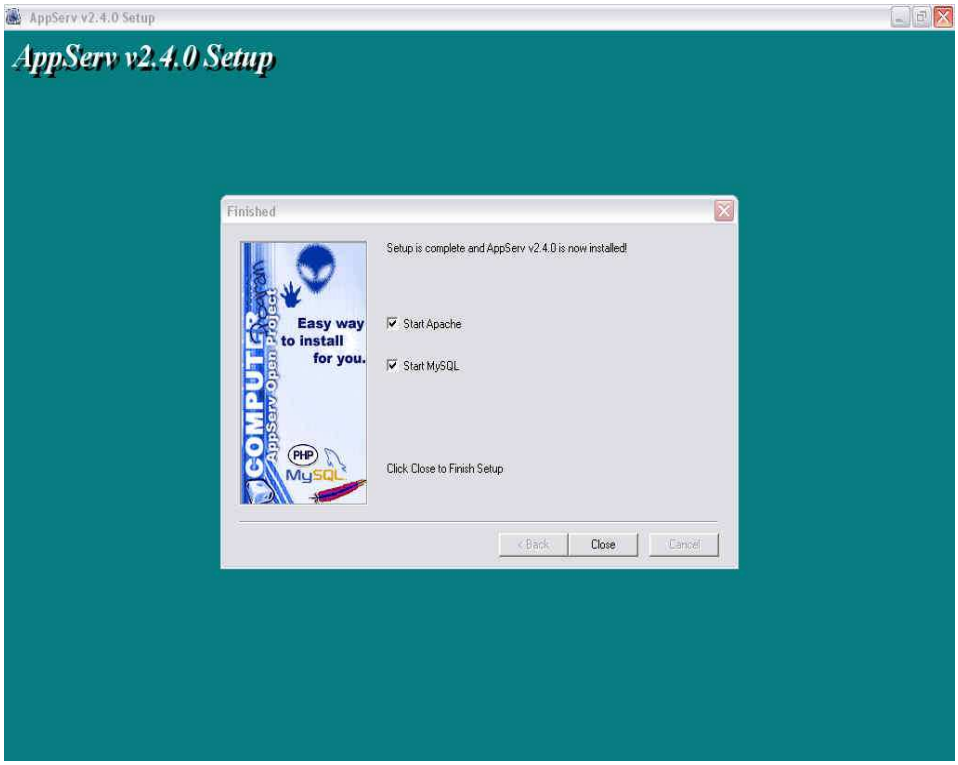
5. จากนั้นที่หน้าต่าง Apache Httpd Server ดำเนินการตั้งค่า Web Server ซึ่งกำหนดค่าเริ่มต้นของ Server Name เป็น localhost และ HTTP Port เป็น ไม่แนะนำให้เปลี่ยนค่าเช่นกันครับ จากนั้นให้กดปุ่ม Next จะปรากฏหน้าจอ ดังรูป



6. จากนั้นที่หน้าต่าง MySQL Database จะเป็นการตั้งค่าของ MySQL Database ซึ่งกำหนดค่าเริ่มต้นของ User Name เป็น mysql, Password เป็น mysql และ Charset เป็น latin1 ไม่แนะนำให้เปลี่ยนค่าเช่นกันครับ กดปุ่ม Next จะปรากฏหน้าจอ ดังรูป



7. ชุดติดตั้งจะทำการติดตั้งโปรแกรม Apache + PHP + MySQL ลงในระบบ หลังจากการติดตั้งเสร็จเรียบร้อย จะปรากฏหน้าจอดังรูป



8. คลิกเพื่อเช็คเลือกให้สตาร์ท Apache และ MySQL เมื่อเปิดเครื่องใหม่ทุกครั้ง หลังจากนั้น กดปุ่ม Close เพื่อสิ้นสุดการติดตั้ง

การเริ่มต้นใช้งาน (Start) PHP แอพพลิเคชั่นเซิร์ฟเวอร์

โดยปกติเมื่อเปิดเครื่องคอมพิวเตอร์ โปรแกรม Apache เว็บเซิร์ฟเวอร์ และ MySQL ดาต้าเบสเซิร์ฟเวอร์จะทำงานโดยอัตโนมัติ แต่หากโปรแกรมไม่ทำงาน เราสามารถสั่งให้เริ่มต้นทำงานได้ โดยวิธีการดังนี้

1. รัน Apache เว็บเซิร์ฟเวอร์ โดยคลิกที่เมนู

Start -> Programs -> AppServ -> Manual Control Server -> Apache Start

2. รัน MySQL ดาต้าเบสเซิร์ฟเวอร์ โดยคลิกที่เมนู

Start -> Programs -> AppServ -> Manual Control Server -> MySQL Start

การหยุดใช้งาน (Stop) PHP แอพพลิเคชั่นเซิร์ฟเวอร์

หากต้องการหยุดการทำงานของโปรแกรม Apache เว็บเซิร์ฟเวอร์ และ MySQL ดาต้าเบสเซิร์ฟเวอร์ สามารถสั่งให้หยุดทำงานได้ โดยวิธีการดังนี้

1. หยุดการทำงาน Apache เว็บเซิร์ฟเวอร์ โดยคลิกที่เมนู

Start -> Programs -> AppServ -> Manual Control Server -> Apache Stop

2. หยุดการทำงาน MySQL ดาต้าเบสเซิร์ฟเวอร์ โดยคลิกที่เมนู

Start -> Programs -> AppServ -> Manual Control Server -> MySQL Stop

การลบ (Uninstall) โปรแกรม AppServ (PHP แอพพลิเคชั่นเซิร์ฟเวอร์)

หากต้องการลบโปรแกรม AppServ ออกจากเครื่องคอมพิวเตอร์ที่ติดตั้ง สามารถทำได้โดยวิธีการดังต่อไปนี้ ตามลำดับ

1. หยุดการทำงาน Apache เว็บเซิร์ฟเวอร์ โดยคลิกที่เมนู

Start -> Programs -> AppServ -> Manual Control Server -> Apache Stop

2. หยุดการทำงาน MySQL ดาต้าเบสเซิร์ฟเวอร์ โดยคลิกที่เมนู

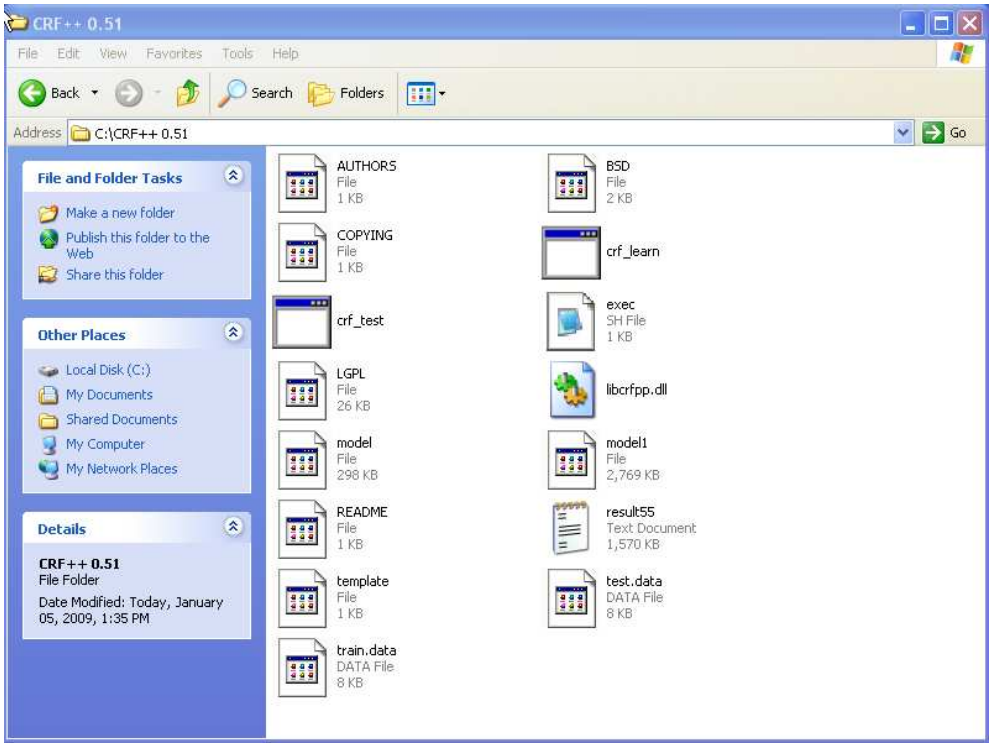
Start -> Programs -> AppServ -> Manual Control Server -> MySQL Stop

3. ลบโปรแกรมโดยคลิกที่เมนู

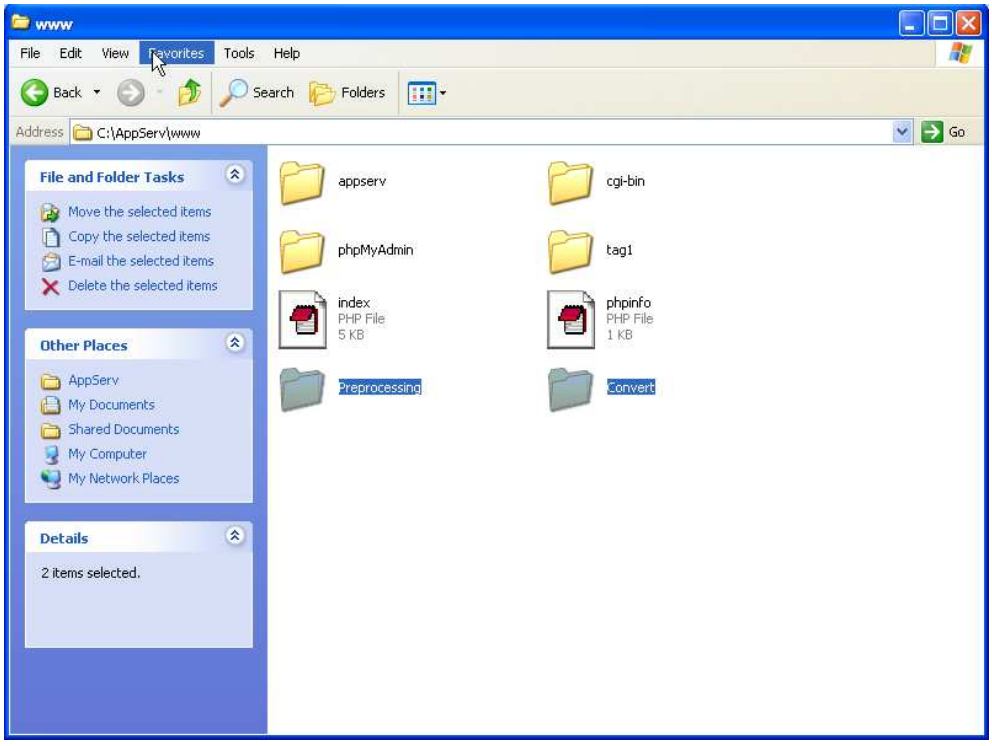
Start -> Programs -> AppServ -> Uninstall AppServ v2.4.4a

การติดตั้งโปรแกรม CRF++

1. คัดลอกไฟล์ลงใน Local Disk(C:) เวลาใช้งานจะเรียกใช้ทาง dos ผ่านทาง CRF++



2. คัดลอกไฟล์ convert.php และ tag.php เป็นไฟล์ในการทำ Preprocessing ลงใน C:\AppServ\www เวลาใช้งานจะเรียกใช้ทาง dos หรือ browser ก็ได้



คู่มือการใช้งาน

ระบบสามารถแบ่งได้เป็น 2 ส่วนด้วยกัน คือ

ส่วนของการสร้างโมเดล สามารถทำการสร้างได้ดังขึ้นตอนต่อไปนี้

1. ทำการ download corpus ทั้ง 6 ชุด ที่ได้จาก
   http://www.hlt.nectec.or.th/best/index.php?option=com_docman&task=cat_view&gid=13&Itemid=33

โดยลักษณะของข้อมูลจาก corpus มีลักษณะที่มีการแบ่งคำไว้ ดังรูป

```
http://www.bangkokhealth.com/healthnews _ htdoc/healthnews _ detail.asp?Number=10044
คู่มือ|ปลอดภัย| |ห่างไกล|
<NE>กรมควบคุมโรค</NE> | ออก|คำ|แนะ|สำหรับ|ประชาชน|เพื่อ|ป้องกัน|โรค|ติดต่อ|จาก|สัตว์|ปีก| |ทั้ง|ผู้|บริโภค| |ผู้|ชำแหละ|และ|เกษตรกร|ผู้|เลี้ยง|ไก่|
สืบเนื่อง|จาก|สถานการณ์|การ|ระบาด|ของ|โรค|ใน|ไก่|ใน|ขณะ|นี้| |ส่งผล|ให้|ประชาชน|ตื่น|ตระหนก|เกรง|จะ|ติด|เชื้อ| |ส่วน|ใหญ่|จึง|งด|บริโภค|เนื้อ|และ|ไข่|ไก่|ทำ|
ให้|ผู้|ผลิต|และ|จำหน่าย|ประสบ|ภาวะ|ขาดทุน|อย่าง|หนัก| | <NE>กรมควบคุมโรค</NE> | <NE>กระทรวงสาธารณสุข</NE> | จึง|ได้|ออก|คำ|แนะ|สำหรับ|
ประชาชน|เพื่อ|ป้องกัน|โรค|ติดต่อ|จาก|สัตว์|ปีก| |โดย|ยืนยัน|ว่า|ประชาชน|ทั่วไป|มี|ความ|เสี่ยง|น้อย|ต่อ|การ|ติด|โรค| |และ|เพื่อ|ให้|เกิด|ความ|ปลอดภัย|สูง|สุด|
  | <NE>กรมควบคุมโรค</NE> | จึง|ขอ|แนะนำ| |ดัง|นี้|
ผู้|บริโภค|
1|.|ผู้|บริโภค|ไก่|และ|ผลิตภัณฑ์|จาก|ไก่| |ควร|รับประทาน|เนื้อ|ที่|ปรุง|สุก|เท่านั้น| |เนื่องจาก|เชื้อ|โรค|ต่างๆ| |ที่|อาจ|ปนเปื้อน|มา| |ไม่ว่า|จะ|เป็น|ไวรัส| |แบคทีเรีย| |หรือ|พยาธิ|
2|.|สำหรับ|เนื้อ|ไก่|ที่|มี|ขาย|อยู่|ตาม|ท้องตลาด|ใน|ขณะ|นี้| |ถือ|ว่า|มี|ความ|ปลอดภัย|สามารถ|บริโภค|ได้|ตาม|ปกติ| |แต่|ต้อง|รับประทาน|เนื้อ|ไก่|สุก|เท่านั้น| |งด|การ|รับประท
3|.|ส่วน|ไข่|ไก่|ก็|ควร|เลือก|ฟอง|ที่|สด|ใหม่|และ|ไม่|มี|มูล|ไก่|ติด|เปื้อน|ที่|เปลือก|ไข่| |ก่อน|ปรุง|ควร|นำ|มา|ล้าง|ให้|สะอาด| |และ|ปรุง|ให้|สุก|ก่อน|รับประทาน|
ผู้|ชำแหละ|
```

จากไฟล์ที่ได้ ทำการเปลี่ยนรูปแบบเพื่อสร้าง data set โดยใช้คำสั่ง

```
C:\>php tag.php
```

ผลลัพธ์ที่ได้จะได้ data set ที่มีโครงสร้างดังรูป

```
<       o       B
N       o       I
E       o       I
>       o       I
ไ       w       I
น       c       I
ห       n       I
ล       c       I
ว       c       I
ง       c       I
<       o       I
/       o       I
N       o       I
E       o       I
>       o       I
ค       c       B
ี       v       I
ข       c       B
ี       v       I
้       t       I
น       c       I
ท       c       B
ร       c       I
ง       c       I
```

2. จากนั้นใช้ CRF ++ 0.51 package ในโฟลเดอร์ example และ โฟลเดอร์seg เพื่อใช้ในการ Train หรือ เพื่อฝึกฝนให้กับเครื่อง เพื่อให้เครื่องได้ทราบถึงรูปแบบของการตัดคำของข้อมูลที่มีการตัดคำที่ถูกต้อง ซึ่งข้อมูลนั้นได้มากจากการ corpus ทั้ง 6 ชุด

การสร้างโมเดล ได้ใช้ package จาก CRF++ เพื่อทำการสร้างโมเดล โดยใช้คำสั่ง

```
crf_learn -f 3 -c 4.0 template train.txt model
```

ส่วนของการสร้างโมเดล สามารถทำการสร้างได้ดังขึ้นตอนต่อไปนี้

สำหรับการ test จะเลือกตัวอย่างมาต่างหากจากชุดข้อมูลที่ได้จาก corpus โดยที่ชุดข้อมูล test กะ train ไม่ซ้ำกัน เพื่อไม่เกิด bias (หรือ download จาก http://crfpp.sourceforge.net/#download)

1. ทำการแปลงไฟล์ รูปแบบไฟล์จาก txt file ให้อยู่ในรูปแบบ ดังรูปเพื่อทำการ ตัดคำจากโมเดล

```
<       o
/       o
N       o
E       o
>       o
ค       c
ื       v
ข       c
ื       v
้       t
น       c
ท       c
ร       c
่       c
ป       c
ล       c
ื       v
้       t
ม       c
พ       c
ร       c
ะ       v
ท       c
ั       v
ย       c
พ       c
ส       c
ก       c
น       c
ื       v
ก       c
ร       c
<       o
N       o
E       o
>       o
```

จากนั้นทำการตัดคำภาษาไทยโดยใช้คำสั่ง

```
crf_test -m model test.txt > seg_text.txt
```

โดยผลลัพธ์ที่ได้จากการทำนาย จะอยู่ในรูปแบบดังรูป

```
ไ       w       I
น       c       I
ห       n       I
ล       c       I
ว       c       I
ง       c       I
<       o       I
/       o       I
N       o       I
E       o       I
>       o       I
ด       c       B
ี       v       I
ข       c       B
ี       v       I
้       t       I
น       c       I
ท       c       B
ร       c       I
ง       c       I
ป       c       B
ล       c       I
ี       v       I
้       t       I
ม       c       I
พ       c       I
ร       c       I
ะ       v       I
ท       c       I
ี       v       I
ย       c       I
พ       c       B
ส       c       I
ก       c       I
น       c       I
ี       v       I
ก       c       I
ร       c       I
```

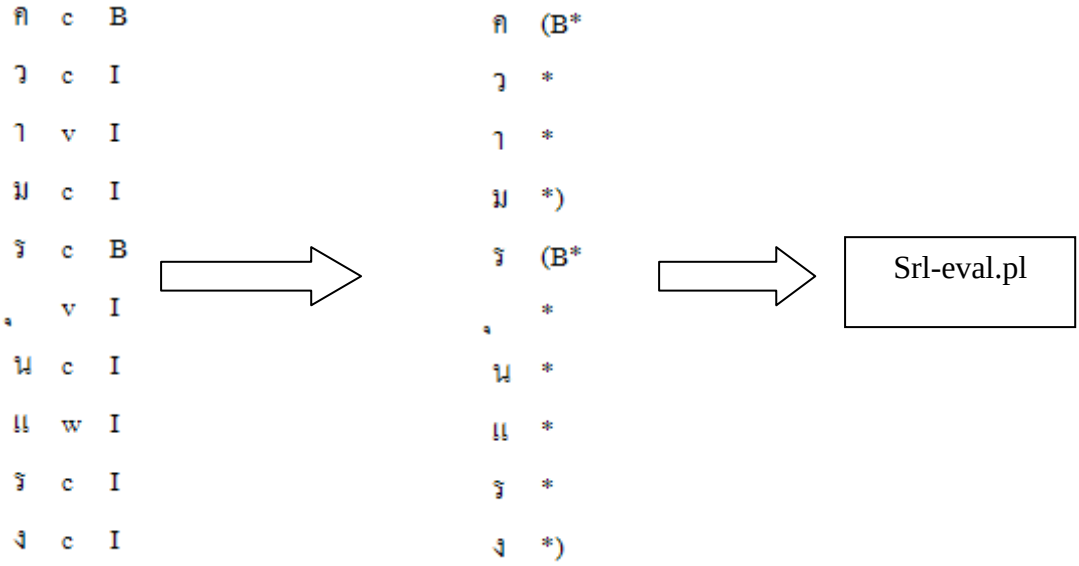
3. นำส่วนของไฟล์ ที่ทำการ test เสร็จเรียบร้อยแล้ว ไฟล์ทำการแปลงรูปแบบต่อให้กลับมาอยู่ในรูปแบบที่กฎการแข่งขันได้ระบุไว้ เช่น ฉัน|ไป|โรงเรียน โดยใช้คำสั่ง

```
php convert.php
```

เพื่อแปลงกลับให้อยู่ในรูปของเอกสารที่ได้รับการแบ่งคำเรียบร้อยแล้ว

ส่วนของการวัดประสิทธิภาพของระบบตัดคำภาษาไทย

ในส่วนที่ 5 จะไม่มีในโปรแกรมเพื่อทำการแข่งขัน แต่จะอยู่ในส่วนของการทดสอบความถูกต้อง ซึ่งส่วนนี้จะทำการประมวลผลเทียบความถูกต้องกับเฉลยที่ได้จาก BEST ซึ่งการทดสอบนั้นจะประมวลผลตามมาตรฐานของ CoNLL-2005 Shared Task โดยประมวลผลผ่านโปรแกรม srl-eval.pl **(http://www.lsi.upc.es/~srlconll/soft.html)** โดยมีการแปลงรูปแบบของคำตอบดังรูป



เมื่อทำการเรียกใช้โปรแกรม srl-eval.pl แล้ว จะแสดงผลลัพธ์ความถูกต้อง ดังรูป