บทที่ 3

วิธีการดำเนินการวิจัย

งานวิจัยนี้นี้กล่าวถึงการปรับปรุงประสิทธิภาพของการตัดคำโดยใช้เทคนิคการจดจำนิพจน์ระบุนามที่ใช้วิธี Conditional Random Fields (CRFs) ในการเรียนรู้เพื่อสร้างโมเดล การจดจำนิพจน์ระบุนาม ซึ่งใช้คุณลักษณะ คำขึ้นต้นและคำลงท้ายของนิพจน์ระบุนาม ในการเรียนรู้ของโมเดล หลังจากนั้นจะนำผลที่ได้จากการตัดคำภาษาไทยของโปรแกรม (TLex) นำมาปรับปรุงประสิทธิภาพการตัดคำ โดยใช้โมเดลในการเรียนรู้ มาปรับปรุงประสิทธิภาพ โดยมีตารางการวิจัยดำเนินงานตามตารางที่ 3.1

ตาราง 3.1

ตารางการดำเนินการวิจัย

กิจกรรม	ระยะเวลา (เดือน)				
	พ.ย -ธ.ค 52	ม.ค-ก.พ 53	มี.ค-เม.ย 53	พ.ค.-มิ.ย 53	ก.ค.-ส.ค 53
1. ศึกษาขอบเขตของปัญหา	←→				
2.ทบทวนวรรณกรรม	←→				
3.วางขอบเขตและกำหนดเป้าหมาย	←→	←→			
6.ทำการทดลอง		←→	←→		
4.สอบเค้าโครงวิทยานิพนธ์			←→		
5. ออกแบบการทดลอง			←→		
7. สรุปผลการทดลอง			←→	←→	
8.จัดทำรายงานวิทยานิพนธ์				←→	
9. เขียนบทความ				←→	
10.เสนอบทความ				←→	
11.สอบวิทยานิพนธ์					←→

3.1 ขอบเขตการวิจัย

ขอบเขตการทดลองสำหรับงานวิจัยชิ้นนี้ประกอบด้วย ข้อมูลที่ใช้วิจัยและหมวดหมู่ที่ใช้ในการวิจัย

3.1.1 ข้อมูลที่ใช้ในการวิจัย

คลังข้อมูล (Corpus) ที่ใช้คือ จากคลังข้อมูลของ BEST2010 ขนาด 5 ล้านคำ ในการฝึกฝนและทดสอบ โดยจัดเก็บให้อยู่ในรูปแบบ Text File ดังแสดงตัวอย่างในภาพที่ 3.1

3.1.2 หมวดหมู่ที่ใช้ในการวิจัย

หมวดหมู่ที่ใช้ในการวิจัย นำมาจากหมวดหมู่ข้อมูลที่มาจากเว็บไซต์หรือหนังสือพิมพ์ที่กล่าวถึงในหัวข้อ ข้อมูลที่ใช้ทดลอง ที่ได้จัดแบ่งกลุ่มของข้อมูลไว้แล้ว หมวดหมู่ที่คัดเลือกมามี 7 หมวดหมู่ ดังนี้ หมวด บทความต่างๆ จำนวน 198 ไฟล์, สารานุกรม จำนวน 108 ไฟล์, ข่าวต่างๆ จำนวน 96 ไฟล์ และบทความวิชาการ จำนวน 36 ไฟล์ บทความพระพุทธศาสนา จำนวน 52 ไฟล์ และสุดท้ายบทความวิกิพีเดีย จำนวน 106 ไฟล์ รวมทั้งหมด 596 ไฟล์ ซึ่งข้อมูลเหล่านี้ถูกรวบรวมขึ้น โดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ โดยมีรายละเอียดตามภาพที่ 3.1

ภาพที่ 3.1

ตัวอย่างข้อความใน คลังข้อมูล

<NE>คมช.</NE>|เขย่า|ขวัญ| |<AB>รธน.</AB>|ไม่|ผ่าน|ใช้|ปี|40|+|35|เปิด|ทาง|นายก|ฯ| |คน|นอก|ขณะ|ที่|"|<NE>สมัคร</NE>|"|เข้า|ซุ้ม|<NE>พลังประชาชน</NE>|แล้ว|พร้อม|นั่ง|หัวหน้า|พรรค|หลาย|ฝ่าย|ยัง|ข้องใจ|ด้าน|กลุ่ม|สิทธิ|มนุษยชน|เมือง|ผู้ดี|ต้าน|"|<NE>ทักษิณ</NE>|"|ฮุบ|<NE>แมนฯซิตี้</NE>| |นิรโทษกรรม|สากล|ชี้|3|ปม|ต้อง|พิสูจน์|"|ฆ่า|ตัด|ตอน|-|<NE>ตากใบ</NE>|-|แทรกแซง|สื่อ|"|การ|รณรงค์|เพื่อ|ให้|ไ|<NE>จ.ระนอง</NE>|ใช้|อาสาสมัคร|กระจาย|กำลัง|ลง|ไป|ใน|พื้นที่| |<NE>จ.อุทัยธานี</NE>|ใช้|มวล|ชน|วิ่ง|ธง|ขับเคลื่อน|ประชาธิปไตย| |<NE>จ.ขอนแก่น</NE>|ตั้ง|เป้า|ผู้|ออก|ไป|ใช้|สิทธิ|ราว|60|%| |ส่วน|<NE>จ.ยะลา</NE>|ผู้ว่า|ราชการ|จังหวัด|ได้|มอบ|ร่าง|รัฐธรรมนูญ|เพื่อ|เผยแพร่|ให้|ประชาชน| |โดยที่|<NE>สมาชิกสภาร่างรัฐธรรมนูญ</NE>|นิด</NE>|โฆษก|<NE>คมช.</NE>| |กล่าว|ถึง|กรณี|ที่| |

3.2 เครื่องมือและซอฟต์แวร์ที่ใช้ในการพัฒนาระบบ

3.2.1 อุปกรณ์ที่ใช้ในงานวิจัยประกอบด้วย

1. ฮาร์ดแวร์

1. เครื่องคอมพิวเตอร์ 1 ชุด ประกอบด้วย หน่วยประมวลผลกลาง Centrino core 2 Duo 1.83 GHz หน่วยความจำหลัก 3 GB ฮาร์ดดิสก์ขนาด 120 GB

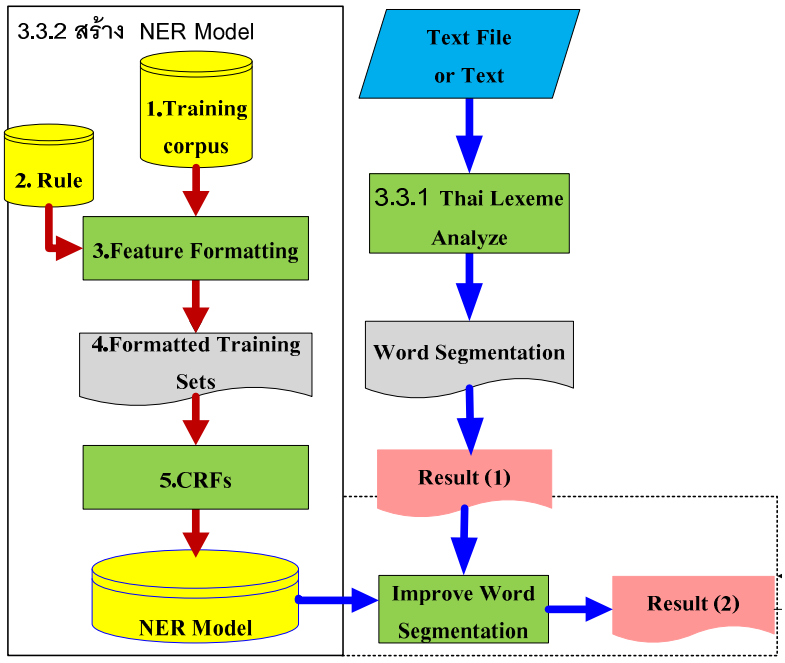
2. ซอฟต์แวร์

1. ระบบปฏิบัติการ Windows
2. เครื่องมือพัฒนาโปรแกรม jdk-6u10-windows-i586-p
3. โปรแกรม CRF++-0.53

3.3 ขั้นตอนการทดลอง

ขั้นตอนการวิจัยโดยภาพรวมทั้งระบบ แสดงได้ดังภาพที่ 3.2 โดยสามารถแบ่งออกได้เป็น 2 กระบวนการใหญ่ๆ ดังต่อไปนี้

ภาพที่ 3.2
ภาพแสดงขั้นตอนการวิจัย

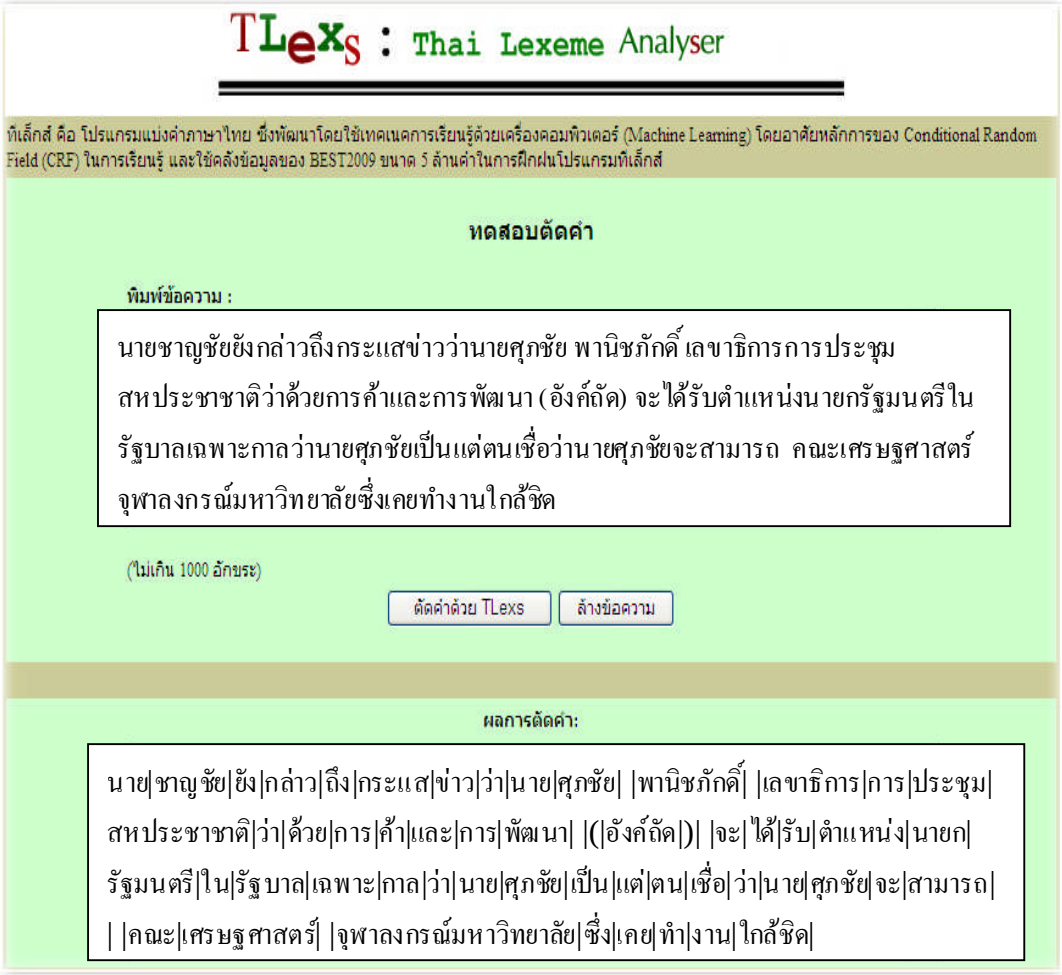


3.3.1 ทำการตัดคำด้วยโปรแกรม Thai Lexeme Analyser (TLex)

ทีเล็กส์ คือ โปรแกรมตัดคำภาษาไทย ซึ่งพัฒนาโดยใช้เทคนิคการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการของ Conditional Random Field (CRFs) ในการเรียนรู้ และใช้คลังข้อมูลของ BEST2009 ขนาด 5 ล้านคำในการฝึกฝน ผลลัพธ์ที่ได้จากการตัดคำ โดยโปรแกรม TLex คือผลลัพธ์ ที่ได้ก่อนนำไปปรับปรุงการตัดคำ หลังจากนั้นจะนำไปปรับปรุงโมเดลการจดจำนิพจน์ระบุนาม ในการแก้ไขปัญหาของนิพจน์ระบุนาม

ภาพที่ 3.3

ภาพแสดงตัวอย่างโปรแกรม TLex



3.3.2 **การสร้างโมเดลในการจดจำนิพจน์ระบุนาม** (Named Entity Recognition)

โดยการสร้างโมเดลการจดจำนิพจน์ระบุนาม เพื่อปรับปรุงการตัดคำให้มีประสิทธิภาพมากขึ้น ขั้นตอนการสร้าง โมเดลการจดจำนิพจน์ระบุนาม มีดังต่อไปนี้

3.3.2.1 คลังข้อมูล (Corpus) ที่ใช้คือ จากคลังข้อมูลของ BEST2010 ขนาด 5 ล้านคำ ในการฝึกฝน และทดสอบ โดยจัดเก็บให้อยู่ในรูปแบบ Text File

3.3.2.2 กำหนดกฎในการเรียนรู้ของ CRFs

1.หลักฐานคำนำหน้านิพจน์ระบุนามที่ปรากฏอยู่หน้านิพจน์ระบุนาม ได้แก่คำนำหน้านิพจน์ระบุนามต่างๆ ซึ่งรายการคำนำหน้าชื่อที่เก็บได้มาจากคลังข้อมูลจะแสดงดังต่อไปนี้ "อ." , "นาย" , "นาง" , "ดร." , "พล.ต." , "พล.ท." , "พ.ญ." , "พ.อ." , "ส.ก." , "ร.ต." , "ร.อ." , "ภ.ญ." , "ร.ท." , "จอมพล" , "น.ส." , "นางสาว" , "น.ต." , "อาจารย์" , "หัวหน้า" ,"พล.อ." , "น.พ." , "ม.ล." , "คุณหญิง" ,"หลวงพ่อ" , "ร.อ.หญิง" , "พ.ต.ท." , "พ.ต.หญิง" ,"ส.ต.อ." , "พล.ร.ท." , "ร.ต.อ." , "พล.ต.ท." , "ม.ร.ว." , "พล.อ.อ." , "พล.ต.ต." , "พล.ต.อ." , "ส.ต.ต." , "พ.ต.ต." , "พ.ต.อ." , "พล.ต.ท." , "ร.ต.ต" , "จ.ส.ต." , "ร.ต.ท." , "พ.ญ.คุณหญิง" ,"ว่าที่ร.ต." , "พ.ต.อ.น.พ." , "กรม" , "พรรค" , "องค์การ" , "กระทรวง" , "สำนักงาน" ,"สำนัก" ,"กองทัพ" , "บริษัท" , "ธนาคาร" , "องค์กร" , "กอง" , "โรงเรียน" , "โรงพยาบาล" ,"วิทยาลัย" , "บรรษัท" , "สหภาพ" , "สภา" , "มูลนิธิ" , "ศาล" , "สถาบัน" , "ศูนย์" , "ทวีป" ,"ร้านกาแฟ" , "ร้านอาหาร" , "สวนสัตว์" , "ค่าย" , "สถานทูต" ,"สนามกอล์ฟ" , "ท่าอากาศยาน" , "สุเหร่า" , "บึง" , "คุก" , "มัสยิด" , "สวนสาธารณะ" , "ร.ร." , "พระธาตุ" , "เกาะ" ," สโมสร" , "สมาคม" , "สาธารณรัฐ" , "ราชอาณาจักร" , "จังหวัด" , "แม่น้ำ" , "ลุ่มน้ำ" ,"เขา" , "อ่างเก็บน้ำ" , "ท่าน้ำ" , "ท่าเรือ" , "ท่าเรือน้ำลึก" , "ปราสาท" , "ปราสาทหิน" ,"คลอง" , "วัด" , "วัง" , "รพ." , "อนุสาวรีย์" , "พระบรมราชานุสาวรีย์" , "เขื่อน" , "อนุสรณ์" ,"มณฑล" , "นคร" , "ต." , "อาคาร" , "ตึก" , "โรงแรม" , "สนามบิน" , "ประเทศ" , "แขวง" ,"หมู่บ้าน" , "สนามม้า" , "แฟลต" , "ซอย" , "ร้าน" , "เคหะชุมชน" , "โรงงาน" , "สวน" ,"บ้าน" , "ธนาคาร" , "ศาลเจ้า" , "สน." , "อุทยาน" , "ถ." , "เขต เทศบาลตำบล" , "ศูนย์ประชุม" , "หอประชุมอำเภอ" , "ทางด่วนสาย" , "ถนนสาย" , "สำนักงานสาขา" ,"หอประชุม" , "ที่ว่าการ" , "มูลนิธิ" , "ภูมิภาค" , "สถานี" , "ถนน" , "ที่ทำการสาขา" , "ห้อง" ," สถาบัน" , "อำเภอ" , "หอ" , "ที่ทำการ" , "ภาค" , "ตลาดสดเทศบาล" , "ตลาดสด" ,"ตลาด" , "สนามกีฬาเทศบาล" , "สนามกีฬากลาง" , "สนามกีฬา" , "สนาม" , "ศาลา" ,"จ." , "เมือง" , "กรุง" , "เขต" ฯลฯ หรือดูตัวอย่างได้ที่ภาคผนวก ก

2. หลักฐานหลังนิพจน์ระบุนาม ที่ปรากฏอยู่ด้านหน้าและหลังนิพจน์ระบุนาม ได้แก่ บริษัท...จำกัดแห่งชาติ, แห่งประเทศไทย,จำกัด เป็นต้น

3.3.2.3 กำหนดกฎในการฝึกฝน (Feature Formatting)

กฎในการฝึกฝนโดยทำการดูนิพจน์ระบุนามที่มีคำขึ้นและคำลงท้าย เพื่อกำหนดประเภทและคุณลักษณะในการเรียนของโมเดล โดยมีรายละเอียดดังตารางที่ 3.2

ตาราง 3.2

ตาราง กำหนดกฎในการฝึกฝน

VALUE	TYPE	TAG	LABLE
นาย , นาง , ดร. , คณะ มหาวิทยาลัย ฯลฯ	เป็นคำอยู่หน้าชื่อเฉพาะ	P	B-NE
แห่งชาติ,โพลล์ ,จำกัด (มหาชน),ออนไลน์ ,ฯลฯ	เป็นคำอยู่หลังชื่อเฉพาะ	S	I-NE
กิน,นอน ,เที่ยว, ไป ,ฯลฯ	เป็นคำทั่วไป	O	I

3.3.2.4 การกำหนดข้อมูลในการฝึกฝน

ตัวอย่างข้อความในคลังข้อความ “<NE> รศ.ดร.เกษียร</NE><NE> มหาวิทยาลัยธรรม </NE> หมายเหตุ ”

ภาพที่ 3.4

ภาพแสดงตัวอย่างในการฝึกฝน

รศ.	P	B-NE
ดร.	P	I-NE
เกษียร	O	I-NE
มหาวิทยาลัย	P	B-NE
ธรรมศาสตร์	O	I-NE
หมาย	O	I
เหตุ	O	I

การฝึกฝนในการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการของ Conditional Random Fields จะประกอบด้วย 3 คอลัมน์ โดยคอลัมน์ที่ 1 คือ คำ ที่ต้องการให้

คอมพิวเตอร์เรียนรู้ คอลัมน์ที่ 2 คือ ตัวแยกประเภทของคำว่าเป็นคำขึ้นต้นหรือคำลงท้าย โดยดูได้จาก ตารางที่ 3.2

P คือคำขึ้นต้น เช่น นาย , นางสาว , รศ , ดร ฯลฯ

O คือ คำปกติทั่วไป

S คือ คำที่ลงท้าย เช่น คำว่า แห่งชาติ, จำกัด, จำกัด (มหาชน) ฯลฯ

ในส่วน คอลัมน์ที่ 3 คือ การกำหนด Label ให้กับคำที่บ่งบอกถึงคำขึ้นต้นหรือคำต่อท้ายโดยสอดคล้องกับคอลัมน์ที่ 2 และเพื่อให้เป็นผลลัพธ์ของค่าในการฝึกฝน

B-NE จะบ่งบอกได้ว่าเป็นขึ้นต้นและเป็นนิพจน์ระบุนาม

I-NE จะบ่งบอกได้ว่าเป็นคำที่เกี่ยวข้องกับนิพจน์ระบุนาม

I เป็นคำปกติทั่วไป

3.3.2.5 ทำการสร้างโมเดลการจดจำนิพจน์ระบุนาม โดย ใช้วิธีการการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการของ Conditional Random Fields

3.4 การทดสอบ

การทดสอบ จะทำการทดสอบข้อมูลโดยลักษณะเดียวกับ ข้อมูลที่ใช้ในการฝึกฝนจะประกอบด้วย 4 คอลัมน์ โดยคอลัมน์ที่ 1 - 3 เป็น ลักษณะเดียวกันกับการฝึกฝนมีคอลัมน์ที่ 4 เป็นผลลัพธ์ได้มาจากการทดสอบจากโมเดล โดยดูได้จากภาพที่ 3.5

ภาพที่ 3.5

ภาพแสดงตัวอย่าง การทดสอบโมเดลการจดจำนิพจน์ระบุนาม

รศ.	P	B-NE	B-NE
ดร.	P	I-NE	I-NE
เกษียร	O	I-NE	I-NE
มหาวิทยาลัย	P	B-NE	B-NE
ธรรมศาสตร์	O	I-NE	I-NE
หมาย	O	I	I
เหตุ	O	I	I

3.5 การประเมินผล/วัดประสิทธิภาพ

การประเมินผลโมเดลการจดจำนิพจน์ระบุนามและการประเมินผลการปรับปรุงประสิทธิภาพการตัดคำจะใช้วิธีการตาม 3.5.1 และ 3.5.2 ในส่วนการประเมินผลจะทำการวัดประสิทธิภาพโมเดลการจดจำนิพจน์ระบุนามจะใช้วิธีการตาม 3.5.1 และในส่วนการปรับปรุงประสิทธิภาพการตัดคำจะใช้วิธีการตาม 3.5.2 เป็นการวัดประสิทธิภาพแต่ละวิธี โดยมีรายละเอียดดังนี้

3.5.1 การประเมินประสิทธิภาพโมเดลการจดจำนิพจน์ระบุนาม จะดูจากคลังข้อความชุดทดสอบจำนวน 1 ล้านคำซึ่งประกอบด้วย 5 บทความ บทความข่าวสาร บทความกฎหมาย บทความวิชาการ บทความพุทธศาสนาและสุดท้ายบทความวิกิพีเดีย ซึ่งได้มาจากหน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา (HLT) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) โดยใช้วิธีการประเมินผล ดังนี้

ใช้วิธีการวัดค่า F-measure ซึ่งรายละเอียดตามที่ได้ ค่า F1 ซึ่งค่า F1 สามารถหาได้จาก ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall)

$$\text{ค่าความแม่นยำ(Precision)} = \frac{\text{จำนวนนิพจน์ระบุนามที่เลือกมาแล้วถูกต้อง *100}}{\text{จำนวนนิพจน์ระบุนามที่เลือกออกมาทั้งหมด}} \quad (1)$$

$$\text{ค่าความครบถ้วน (Recall)} = \frac{\text{จำนวนนิพจน์ระบุนามที่เลือกมาแล้วถูกต้อง *100}}{\text{จำนวนนิพจน์ระบุนามจริงทั้งหมดในข้อมูล}} \quad (2)$$

การประเมินประสิทธิภาพ จะได้สูตรดังนี้

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} * \text{Recall}} \quad (3)$$

3.5.2 การประเมินประสิทธิภาพการปรับปรุงการตัดคำโดยใช้การโมเดลการจดจำนิพจน์ระบุนามทำการทดสอบกับจำนวน 37 บทความ ขนาด 72,000 คำ โดยตรวจสอบจากคำตอบที่ได้จากการ

ตัดคำที่มาจากหน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา (HLT) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ โดยมีสูตรในการหาค่าความถูกต้องและความครบถ้วนดังต่อไปนี้

ใช้วิธีการวัดค่า F-measure ซึ่งรายละเอียดตามที่ได้ ค่า F1 ซึ่งค่า F1 สามารถหาได้จาก ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall)

ซึ่งค่าเหล่านี้จะหาได้จากสูตร ดังนี้

$$\text{ค่าความแม่นยำ(P)} = \frac{\text{จำนวนคำที่ตัดออกมาแล้วถูกต้อง}}{\text{จำนวนที่ตัดคำออกมาทั้งหมด}} \quad (4)$$

$$\text{ค่าความครบถ้วน (R)} = \frac{\text{จำนวนคำที่ตัดออกมาแล้วถูกต้อง}}{\text{จำนวนตัดคำที่ถูกต้องของบทความ}} \quad (5)$$

การประเมินประสิทธิภาพ จะได้สูตรดังนี้

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} * \text{Recall}} \quad (6)$$

3.6 สมมุติฐานของการวิจัย

1. โมเดลนิพจน์ระบุนาม ที่ได้จากการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการของ Conditional Random Fields สามารถจดจำนิพจน์ระบุนาม ได้อย่างมีประสิทธิภาพ

2. โมเดลนิพจน์ระบุนาม ที่ได้จากการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการของ Conditional Random Fields สามารถ สามารถเพิ่มประสิทธิภาพการตัดคำ ของโปรแกรม TLex ได้อย่างมีประสิทธิภาพ