เกษราภรณ์ ฤๅชัย 6140017022
ศดานันท์ อาศัยบุญ 6140216622

Contest 3

Experiment Setup

ใช้ BIO tag ไปบน syllable และ คำของ training data และ dev data

Model

ใช้ conditional random field (CRF) และ algorithm ชื่อว่า lbfgs ในการเทรนโมเดล เนื่องจาก
- ใช้งานง่าย และให้ผลดี หากใช้ feature ที่เหมาะสม
- จำนวนข้อมูลที่ใช้เทรนไม่เยอะมากจึงเลือกใช้ lbfgs เพราะให้ผลดีกว่า averaged perceptron

Feature

1) รวบรวม gazetteer จากเว็บต่างๆ ได้แก่ wikipedia.org , http://www.codetukyang.com , http://www.mua.go.th และ http://www.phpthai.com ซึ่งประกอบไปด้วย entity name ของ
-place name ได้แก่ ชื่อทวีป ชื่อประเทศ ชื่อจังหวัด ชื่อตำบล ชื่ออำเภอ ชื่อห้างสรพพสินค้า
-organization name ได้แก่ ชื่อมหาวิทยาลัย ชื่อย่อมหาวิทยาลัย ชื่อกระทรวง
-person name ได้แก่ ชื่อจริงของดารานักแสดง
-place prefix ได้แก่ คำที่มักนำหน้าชื่อสถานที่ เช่น ห้าง บริเวณ
-organization prefix ได้แก่ คำที่มักนำหน้าชื่อองค์กร เช่น มหาวิทยาลัย กระทรวง
-person prefix ได้แก่ คำที่มักนำหน้าชื่อคน ตัวย่อยศ ตำแหน่ง เช่น นาย นางสาว ดร. รศ.
เมื่อรวบรวมข้อมูลข้างต้นก็นำมาทำเป็น function แยกกันทั้งหมดโดยเก็บข้อมูลอยู่ในรูปของ list ให้ตรวจว่าคำนั้นอยู่ใน list ที่ทำไว้หรือไม่แล้ว return ค่า True หรือ False

2) ฟังก์ชัน is_space เพื่อตรวจสอบว่าคำนั้นเป็นช่องว่างหรือไม่โดย return ค่า True หรือ False

3) ฟังก์ชัน word_shape เพื่อตรวจสอบว่าคำนั้นมีลักษณะเป็นอย่างไรโดย return string “ก” ถ้าเป็นพยัญชนะ “1” ถ้าเป็นตัวเลข punctuation ถ้าเป็น punctuation เช่น “ทบ.3” return “กก.1”

4) ฟังก์ชัน word_feature เพื่อตรวจสอบว่าคำตำแหน่งปัจจุบัน คำก่อนหน้าหนึ่งคำ (w-1) ก่อนหน้าสองคำ (w-2) และ คำถัดไปจากคำนั้นหนึ่งคำ (w+1) ถัดไปอีกสองคำ (w+2) คือคำว่าอะไรบ้าง

5) ฟังก์ชัน pos_feature เพื่อตรวจสอบว่าคำตำแหน่งปัจจุบัน คำก่อนหน้าหนึ่งคำ (w-1) ก่อนหน้าสองคำ (w-2) และ คำถัดไปจากคำนั้นหนึ่งคำ (w+1) ถัดไปอีกสองคำ (w+2) คือมี part of speech ชนิดใดบ้าง

เกษราภรณ์ ฤๅชัย 6140017022
ศดานันท์ อาศัยบุญ 6140216622

6) ฟังก์ชัน pos_conjunctive_feature นำผลลัพท์จาก pos_features มาเชื่อมกันด้วยเครื่องหมาย “_”

นำ list ของ word เข้าฟังก์ชันข้างต้นทั้งหมดและนำใส่รวมกันใน list ของ feature

Result

ตารางเปรียบเทียบ performance metrics

Model and Algorithm	Feature	Precision	Recall	F1
CRF-lbfgs	syllable-level tokenize word_feature word_shape pos_feature pos_conjunctive_feature is_space gazetteer	0.802	0.802	0.802
CRF-lbfgs	syllable-level tokenize word_feature word_shape pos_feature pos_conjunctive_feature is_space	0.915	0.851	0.882
CRF-average perceptron	syllable-level tokenize word_feature word_shape pos_feature pos_conjunctive_feature is_space	0.770	0.762	0.766

-วิเคราะห์ feature ที่มีส่วนช่วยมากที่สุด

การใช้ feature word_feature word_shape, pos_feature, pos_conjunctive_feature, is_space โดยไม่ใช้ gazeteer อาจเพราะใช้การตัดคำแบบ syllable ทำให้ แต่ละ token ไม่พบใน list ของ gazetteer ใดเลย

-โมเดลใช้ algorithm แบบ lbfgs แล้วได้ผลดีมากที่สุดเมื่อเทียบกับ average perceptron