เกษราภรณ์ ฤๅชัย 6140017022
ศดานันท์ อาศัยบุญ 6140216622

Contest 2

Experiment Setup

Langugage model นี้ไม่จำเป็นต้องใช้ข้อมูลในการ train (Supervised learning) จึงไม่มีการแบ่งสัดส่วนข้อมูล ใช้เพียงข้อมูล dev ในการปรับปรุงโมเดลให้ดีขึ้น แหล่งข้อมูลที่ใช้มาจากหนังสือพิมพ์ The New York Times Newswire Service (209,075,440 tokens) และ คลังหนังสือ Gutenberg (210,907,708 tokens)

Model

เทรน 5-gram based Language model และ 6-gram based Language model ด้วย KenLM toolkits เนื่องจาก

-ใช้เทคนิค Kneaser-Ney smoothing ในการทำนายความน่าจะเป็นของประโยค

-มีการ back-off น้ำหนักของแต่ละ n-gram ด้วย

Algorithm

1.ใช้ nltk ในการ tag part of speech แต่ละคำในประโยค

2. tokenize คำในประโยค

3.ให้โมเดลประเมินคะแนนแต่ละประโยคที่แก้ไขด้วย candidate (รูปแปร/คำที่เป็นไปได้อื่น ๆ) ที่เก็บไว้ใน list/dictionary of list

4.เลือก candidate ที่ได้คะแนนมากที่สุด และแทน candidate นั้น ในตำแหน่งตาม tokens

<u>การเลือก candidate มีรายละเอียดดังนี้</u>

-แก้ article ที่พบบ่อยและผิดบ่อย เช่น a,an,the,some,any

-แก้ คำกริยา โดย candidate มาจาก pyinflect module และ verb.csv จาก github.com/nozomiyamada

-แก้ คำนาม โดย candidate มาจาก pyinflect module และ noun.csv จาก github.com/nozomiyamada

-แก้คำบุพท โดย candidate มาจาก prep.csv จาก github.com/nozomiyamada

-แก้คำกริยาบางคำที่อยู่ในตำแหน่งของคำคุณศัพท์ได้ เช่น She got **punish.** โดยให้เช็คคำที่ tag เป็น Adjective ใน dictionary ของคำกริยาด้วย

เกษราภรณ์ ฤๅชัย 6140017022
ศดานันท์ อาศัยบุญ 6140216622

Result

ตารางเปรียบเทียบค่า precision recall และ F0.5 ของแต่ละโมเดล

ทดลองกับไฟล์ `bea19-sentences.txt(dev set)`

Model - Gram	Data	Model - Function	Precision	Recall	F 0.5
5 Gram	Gutenburg	ไม่มี candiadte ตัด article	0.0360	0.1242	0.0420
	Nyt_eng	ไม่มี candiadte ตัด article	0.0480	0.1537	0.0556
6 Gram	Nyt_eng	ไม่มี candiadte ตัด article	0.0606	0.1479	0.0687
	Nyt_eng	เพิ่ม candidate ตัด article	0.0677	0.1745	0.0771

Conclusion

โมเดลที่ได้ผลดีที่สุดคือ 6 Gram Version ใช้ Nyt_eng เทรนโมเดล และเพิ่ม candaidate ที่ตัด aiticle บางตำแหน่งออกด้วย (เพิ่ม “” ในลิส ของ article)

ปัจจัยที่ส่งผลต่อโมเดล

- **6 Gram ทำให้โมเดลมีประสิทธิภาพมากกว่า 5 Gram** เนื่องมาจากประโยคภาษาอังกฤษมี long dependency การดู n-gram ที่มากขึ้นอาจส่งผลมาก
- **Data Nyt_eng น่าจะมีข้อมูลใกล้เคียงกับไฟล์ที่ใช้ทดลองมากกว่า Gutenburg** ในด้านระดับของภาษา
- **การเพิ่ม Candidate ตัด article บางตำแหน่งออกส่งผลต่อประสิทธิภาพของโมเดล** อาจเพราะข้อผิดพลาดเรื่องการเติม article โดยไม่จำเป็นมีความเป็นไปได้สูงที่จะพบมาก