Machine Translation WMT’14
With sequence to sequence

정규헌

WMT’14 Dataset

- English - French corpus
- europal-v7.fr-en.en : English
- europal-v7.fr-en.fr : french
- 2007724개 문장

Process

1. Dataset을 split, csv 파일로 생성(Tabular Dataset으로 load하기 위해)
2. Spacy를 이용해 tokenizing
3. Modeling
4. Training
5. Evaluation (PPL,BLEU)

Machine Translation

- Tokenizer : spacy
- Dataset load : TabularDataset을 이용
- Batch size : 8
- Optimizer : Adam
- Model
  - Seq2seq
  - Seq2seq with attention
- Evaluation : PPL(Perplexity), BLEU

Sequence to sequence

- With no attention
  - 시간이 attention을 추가할 때보다 더 적게 걸림
  - 학습 수렴이 아주 느리다
- With attention
  - 연산 시간이 약 1.5배 더 걸림
  - 학습 수렴이 attention이 no attention에 비해 아주 빠름

- 문장 길이가 길때
  - 학습시간이 매우 오래 걸림
  - PPL이 아주 높음
- 문장 길이가 짧을 경우
  - 학습시간이 짧게 걸림
  - PPL이 더 낮음

Source sentence : Mais l' alliance occidentale ne combat pas dans les Balkans aux seuls motifs de mettre un terme à la souffrance et de défendre le droit et la justice .

Target sentence : The western alliance is not just fighting in the Balkans to prevent suffering and to secure justice and the rule of law in the Balkans

seq2seq 20 : But the Western alliance does not stop the the only grounds for the and to defend and law .

Seq2seq 30 : The Western alliance is not fighting in the <unk> to the the reasons of putting the end to defend and and justice and justice

But the Western alliance does not fight in the Balkans just for the sake of ending suffering and defending law and justice.

Evaluation

Model	Sentence length	Batch size	perplexity	BLEU
seq2seq	All sentence	8	Over 200	
seq2seq	50	8	120	
seq2seq	30	8	82	
seq2seq_attention	All sentence	8	200	3.84
seq2seq_attention	30	8	81	23.5
seq2seq_attention	20	8	56	24.39

Conclusion

문제점 및 느낀점

1. 문장 길이를 제한해 데이터 셋을 줄인 상태로 실험을 진행했기 때문에 데이터의 양이 줄어들어서 성능이 더 잘 나왔을 가능성이 존재

➡ **실험에 필요한 요인들을 미리 정해놓고 실험을 했어야 함**

2. 문장 길이를 제한해서 학습을 진행할 때 모든 모델들의 성능을 측정할 때 동일한 test셋을 사용해야 했는데 test셋도 동일하게 문장길이를 제한해서 성능을 측정해서 성능이 더 크게 차이났을 가능성이 존재

➡ **평가 척도를 설정할 때 객관적인 방법인지 확인했어야 함**

3. 실험 진행당시 이런 설계를 해두지 않은 상태에서 진행해 너무 많은 시간을 소모, 다시 측정하는데 무리가 있었음

➡ **처음부터 데이터를 완전히 줄여서 설계를 마친 뒤, 전체 데이터를 사용하는 방법을 진행하는게 더 좋았을 것 같다.**