Week 13

태그	
열	

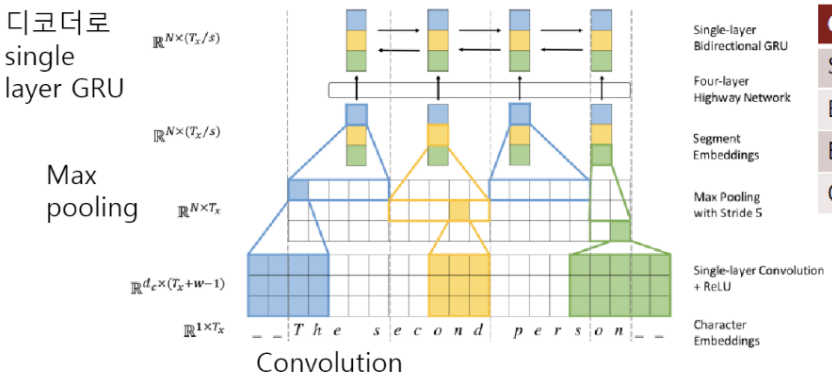
Subword Models

Linguistic knowledge

- Phonetics: 음성학
- Phonology: 음운론
- Phonemes: 음소 (뜻 구별의 최소 단위)
- Morphology: 형태학
- Morpheme: 형태소
- Semantic: 의미론적인

Purely character-level models

- Pure character-level seq2seq system (2015)
  - 영어-체코어 번역
  - Word-level 모델에 견줄 만한 character-level model
  - 학습 시간이 너무 느림 (3주)
  - BLEU 15.9의 성능에 불과함
- Fully Character-level neural machine translation without explicit segmentation (2017)



Cs-En	WMT 15	Test
Source	Target	BLEU
Bpe	Bpe	20.3
Bpe	Char	22.4
Char	Char	22.5

  - 체코어-영어 번역
  - 앞의 모델보다 더 나아진 성능
- Stronger character results with depth in LSTM seq2seq model
  - 영어 → 프랑스어 번역에서는 character based와 word based의 성능에 큰 차이가 없으나, 체코어 → 영어 번역에서는 character based가 더 우수
  - 언어의 특성에 따라 효과가 다르다

Subword models

- Word-level model과 같은 구조: 하지만 “word piece”라는 작은 unit 사용
- Hybrid 구조: 메인 모델은 단어 이용, 문자에는 다른 구조 이용

BPE (Byte Pair Encoding)

- 딥러닝과는 관련이 없으나 단어 조각들을 표현하는 데 유용함
- 자주 나오는 byte pair (n-gram)을 새로운 byte (a new gram)로 클러스터링

Hybrid models

- Character-based LSTM
  - Character level을 합친 output을 더 높은 레벨 모델의 input으로 넣음
- Character-aware neural language models
  - Char 단위로 구분한 상태에서 시작
  - Convolutional layer를 거쳐 feature representation
  - 최종 출력층은 word-level LSTM
- Hybrid Neural Machine Translation

- 대부분 word level 사용
- 필요할 때만 character level 사용
- Hybrid 모델의 성능이 우수하다