Lecture 12. Subwords

Ⅰ. Purely Character-Level Models

1. Purely Character-level model의 필요성

   1) 언어마다 단어 표현 방법이 다르기 때문에 character level로 접근하는 것이 더 효과적인 경우가 있다.

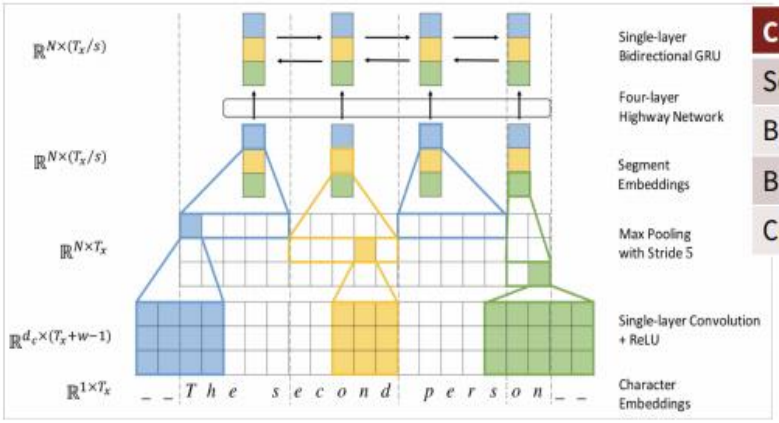
      ex) 중국어는 띄어쓰기가 없고 아랍어는 의미가 추가될 때 하나의 단어처럼 결합됨.

   2) 말그대로 large하고 open한 어휘를 다룰 수 있어야 하기 때문이다.

      ex) 체코어와 같이 풍부한 형태학(morphology)를 가진 언어와 외래어 표기와 같은 음역, sns등에서 자주 볼 수 있는 맞춤법에 맞지 않는 철자 표기(good vibeeeees-> good vibes, idc->I don't care) 등의 문제점 때문이다.

2. Purely character-level model의 성능

   1) 실제로 purely character level model은 특히 체코어 번역에서 우수한 성능을 보였다. Pure character level seq2seq system은 character level model로 영어-체코어 번역을 했지만 학습시간이 3주나 되었고 BLUE가 15.9에 불과하다는 단점이 있었다. 하지만 word level 모델에 비해 사람 이름을 잘 번역하는 경향이 나타났다.



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

4.2 Attention and Decoder

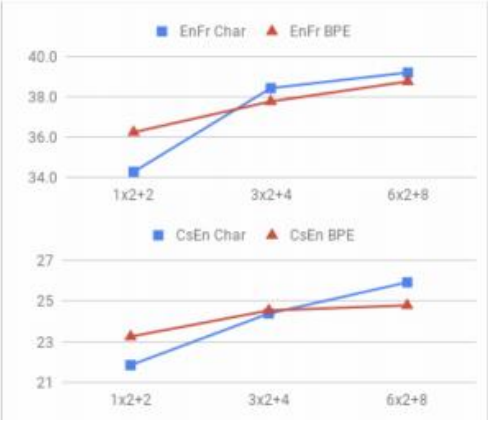
Similarly to the attention model in (Chung et al., 2016; Firat et al., 2016a), a single-layer feedforward network computes the attention score of next target character to be generated with every source segment representation. A standard two-layer character-level decoder then takes the source context vector from the attention mechanism and predicts each target character. This decoder was described as *base decoder* by Chung et al. (2016).

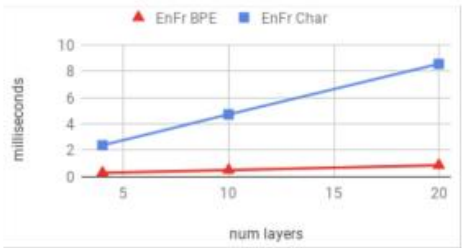
Bilingual	bpe2char	char2char
Vocab size	24,440	300
Source emb.	512	128
Target emb.	512	512
Conv. filters		200-200-250-250 -300-300-300-300
Pool stride		5
Highway		4 layers
Encoder	1-layer 512 GRUs	
Decoder	2-layer 1024 GRUs	

Table 1: Bilingual model architectures. The char2char model uses 200 filters of width 1, 200 filters of width 2, ··· and 300 filters of width 8.

2) 2017년에 나온 또 다른 character level 번역 모델은 앞선 모델보다 더 나은 성능을 보였다. 이 모델의 인코더 구조는 먼저 character 단위의 input을 주고 convolution layer를 거친 다음 max pooling과 single layer GRU를 거치는 구조이다. 앞선 모델보다 BLUE도 더 높다. 위 그림은 모델의 인코더에 해당하는 부분의 구조이다. 모델의 전체적인 구조는 table1과 같이 인코더 외에 2 layer GRU 디코더가 추가되어 있다.

3) 왼쪽 그래프는 character level의 seq2seq 모델과 word level의 BPE 모델의 성능을 비교한 그래프이다. 영어->프랑스어 번역에서는 두 모델의 성능이 크게 차이가 나지 않지만 체코어->영어 번역에서는 character based 모델인 BPE가 훨씬 우수한 성능을 보이고 있다. 오른쪽 그래프는 두 모델의 연산량 차이를 비교한 것인데 character level 모델의 연산량이 훨씬 더 많다.

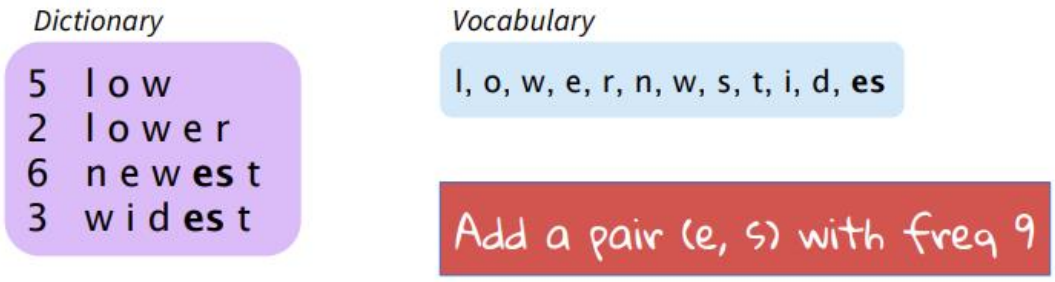




Ⅱ. Subword models

Subword model은 word level 모델과 동일하지만 더 작은 word인 word pieces를 이용한다.

1. BPE



Subword 모델의 대표적엔 예시. 딥러닝과 무관한 간단한 아이디어를 사용하고 있다. 자주 나오는 byte pair(n-gram)를 새로운 byte(a new gram)로 clustering하여 추가하는 방식이다. 왼쪽 그림과 같은 dictionary가 있을 때 es, est, lo가 자주 등장한다. 따라서 이 단어들을 새로 추가하여 각각 하나의 단어로 취급한다. 지정된 최대 길이를 넘으면 이러한 추가 과정을 중단한다. 이러한 점에서 시스템의 vocabulary를 자동적으로 결정한다고 힐 수 있다. BPE 모델은 효과적인 성능이 검증되어 널리 응용되고 있다.

2. Word piece

Word piece 모델은 BPE를 변형한 알고리즘이다. BPE는 빈도수에 기반해서 가장 많이 등장하는 쌍을 병합하는데 word piece 모델은 가장 많이 등장한 쌍이 아닌 병합되었을 때 코퍼스의 우도를 가장 높이는 쌍을 병합한다. BPE와 유사한 점은 자주 등장하는 piece는 unit으로 묶고 자주 등장하지 않는 것은 분리한다는 점이다.

3. Sentence piece

Word piece 토큰화를 수행하는 패키지. 구글에서 2018년에 공개한 비지도학습 기반 형태소 분석 패키지로, 사용하려면 'pip install sentencepiece'를 통해 설치해야 한다.

4. 다른 word piece/sentence piece 모델

-BERT : vocab size가 크지만, 엄청 크지는 않기 때문에 word piece를 사용할 필요가 있다.

따라서 상대적으로 등장 빈도가 높은 단어들과 더불어 word piece를 사용한다.

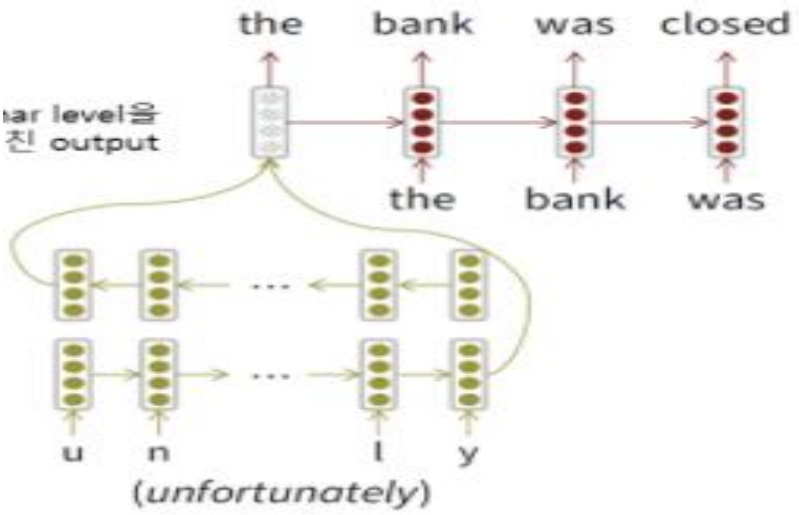
ex) 사전에 없는(등장 빈도가 높지 않은) Hypatia 경우, H yp ati a 처럼 4개의 word vector piece로 쪼개진다.

Ⅲ. Hybrid models

Hybrid model은 기본적으로 단어를 word 단위로 취급하고, 몇몇만 character 단위로 취급하는 모델이다. 주로 고유명사나 사전에 없는 단어를 character 단위로 취급한다.

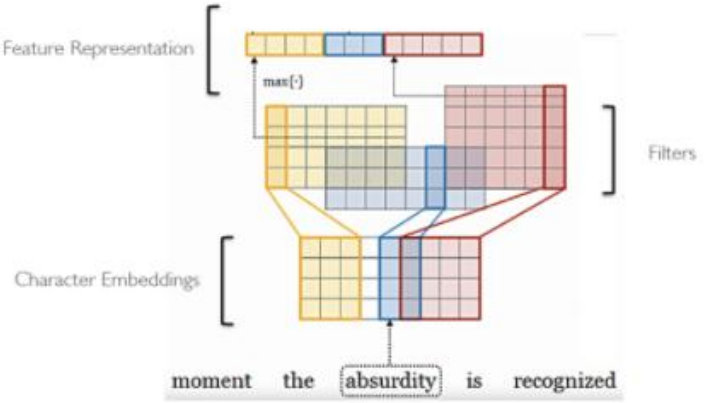
1. Hybrid model

   Hybrid model의 일종인 character-based LSTM의 구조는 아래와 같다. 처음 input 단어를 character level로 다루고 이것을 합친 output이 더 높은 레벨의 모델의 input이 되는 구조이다.



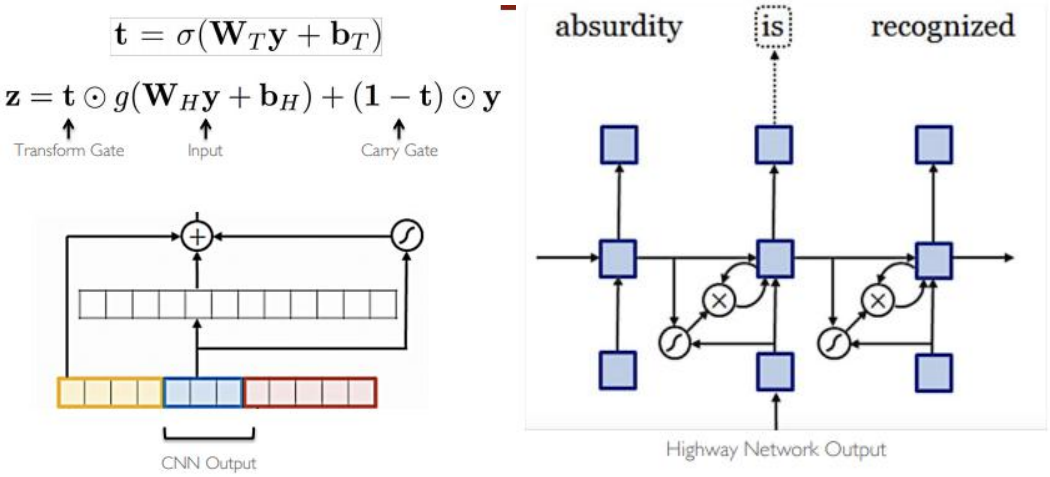
2. Character Aware Neural Language Model

   Subword의 관계성을 인코딩하는 모델. 공통된 subword에 의해 의미적인 관계가 존재하는 경우에 유용하다. 이 모델은 다른 모델이 가진 rare-word problem을 해결하고 더 적은 파라미터 수로 비슷한 성능을 낸다는 장점이 있다



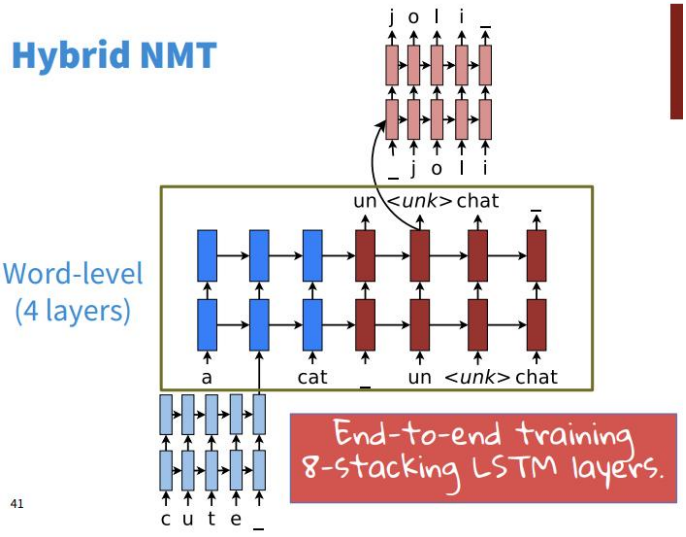
- Convolutions over character-level inputs.
- Max-over-time pooling (effectively n-gram selection).

먼저 input 값에 해당하는 단어를 character 단위로 구분한 다음, 구분된 단어가 convolutional layer를 거쳐 feature representation 단계로 넘어간다. 생성된 feature representation은 highway network를 거치고(LSTM과 유사) highway network의 output은 word level LSTM을 거쳐 최종 output을 출력한다.



다른 모델(size : 52m)보다 훨씬 더 적은 수의 파라미터 수(size : 19m)로 비슷한 성능을 낸다. Highway network 전부분에서는 단어의 철자의 유사도만 파악하고 단어의 의미를 파악하지 못했다. 하지만, highway block 부분을 거친 이후에는 이 단어의 의미를 파악하여 유사한 단어를 출력한다.

3. Hybrid NMT



대부분 word level에서 접근하고 필요할 때만 character level로 접근한다. Beam search를 이용한다.

4. FastText Embedding

Word2vec과 같은 차세대 word vector learning library로 주목받고 있는 임베딩 방법이다. 한 단어의 n-gram과 원래의 단어를 모두 학습에 사용한다. 예시로 where이라는 단어가 있을 때 원래의 단어인 where와 이 단어의 n-gram인 wh, whe, her, ere, re를 모두 학습에 사용한다. N-gram인 her는 기존에 존재하는 단어 her과 다르듯, 이러한 점을 반영한다는 점에서 효과적이다.