Subword Segmentation

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Fast campus

단어보다 더 작은 의미 단위: Subword

- 많은 언어들에서, 단어는 더 작은 의미 단위들이 모여 구성됨

언어	단어	조합
영어	Concentrate	con(=together) + centr(=center) + ate(=make)
한국어	집중(集中)	集(모을 집) + 中(가운데 중)

- 따라서 이러한 작은 의미 단위로 분절할 수 있다면 좋을 것

- 하지만 이를 위해선 언어별 subword 사전이 존재해야 할 것

Fast campus

Byte Pair Encoding (BPE) 알고리즘

- 압축 알고리즘을 활용하여 subword segmentation을 적용 [Sennrich at el., 2015]

- 학습 코퍼스를 활용하여 BPE 모델을 학습 후, 학습/테스트 코퍼스에 적용

- 장점:
  - 희소성을 통계에 기반하여 효과적으로 낮출 수 있다.
  - 언어별 특성에 대한 정보 없이, 더 작은 의미 단위로 분절 할 수 있다.
  - OoV를 없앨 수 있다. (seen character로만 구성될 경우)

- 단점:
  - 학습 데이터 별로 BPE 모델도 생성됨

BPE Training & Applying

- Training
  ① 단어 사전 생성 (빈도 포함)
  ② Character 단위로 분절 후, pair 별 빈도 카운트
  ③ 최빈도 pair를 골라, merge 수행
  ④ Pair 별 빈도 카운트 업데이트
  ⑤ 3번 과정 반복

- Applying
  ① 각 단어를 character 단위로 분절
  ② 단어 내에서 '학습 과정에서 merge에 활용된 pair의 순서대로' merge 수행

Fast campus

Training Example

vocab {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w e s t </w>': 6, 'w i d e s t </w>': 3}
pairs (l, o): 7, (o, w): 7, (w, </w>): 5, (w, e): 8, (e, r): 2, (r, </w>): 2, (n, e): 6, (e, w): 6, (e, s): 9, (s, t): 9, (t, </w>): 9, (w, i): 3, (i, d): 3, (d, e): 3
best pair ('e', 's') 9

vocab {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w es t </w>': 6, 'w i d es t </w>': 3}
pairs (l, o): 7, (o, w): 7, (w, </w>): 5, (w, e): 2, (e, r): 2, (r, </w>): 2, (n, e): 6, (e, w): 6, (w, es): 6, (es, t): 9, (t, </w>): 9, (w, i): 3, (i, d): 3, (d, es): 3
best pair ('es', 't') 9

vocab {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w est </w>': 6, 'w i d est </w>': 3}
pairs (l, o): 7, (o, w): 7, (w, </w>): 5, (w, e): 2, (e, r): 2, (r, </w>): 2, (n, e): 6, (e, w): 6, (w, est): 6, (est, </w>): 9, (w, i): 3, (i, d): 3, (d, est): 3
best pair ('est', '</w>') 9

vocab {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w est</w>': 6, 'w i d est</w>': 3}
pairs (l, o): 7, (o, w): 7, (w, </w>): 5, (w, e): 2, (e, r): 2, (r, </w>): 2, (n, e): 6, (e, w): 6, (w, est</w>): 6, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('l', 'o') 7

vocab {'lo w </w>': 5, 'lo w e r </w>': 2, 'n e w est</w>': 6, 'w i d est</w>': 3}
pairs (lo, w): 7, (w, </w>): 5, (w, e): 2, (e, r): 2, (r, </w>): 2, (n, e): 6, (e, w): 6, (w, est</w>): 6, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('lo', 'w') 7

Fast campus

Training Example

vocab {'low </w>': 5, 'low e r </w>': 2, 'n e w est</w>': 6, 'w i d est</w>': 3}
pairs (low, </w>): 5, (low, e): 2, (e, r): 2, (r, </w>): 2, (n, e): 6, (e, w): 6, (w, est</w>): 6, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('n', 'e') 6

vocab {'low </w>': 5, 'low e r </w>': 2, 'ne w est</w>': 6, 'w i d est</w>': 3}
pairs (low, </w>): 5, (low, e): 2, (e, r): 2, (r, </w>): 2, (ne, w): 6, (w, est</w>): 6, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('ne', 'w') 6

vocab {'low </w>': 5, 'low e r </w>': 2, 'new est</w>': 6, 'w i d est</w>': 3}
pairs (low, </w>): 5, (low, e): 2, (e, r): 2, (r, </w>): 2, (new, est</w>): 6, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('new', 'est</w>') 6

vocab {'low </w>': 5, 'low e r </w>': 2, 'newest</w>': 6, 'w i d est</w>': 3}
pairs (low, </w>): 5, (low, e): 2, (e, r): 2, (r, </w>): 2, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('low', '</w>') 5

vocab {'low</w>': 5, 'low e r </w>': 2, 'newest</w>': 6, 'w i d est</w>': 3}
pairs (low, e): 2, (e, r): 2, (r, </w>): 2, (w, i): 3, (i, d): 3, (d, est</w>): 3
best pair ('w', 'i') 3

vocab {'low</w>': 5, 'low e r </w>': 2, 'newest</w>': 6, 'wi d est</w>': 3}

Segmentation Example

lastest news

1) l a t e s t </w> n e w s </w>
2) l a t es t </w> n e w s </w>
3) l a t est </w> n e w s </w>
4) l a t est</w> n e w s </w>
5) l a t est</w> ne w s </w>
6) l a t est</w> new s </w>

applicable pairs in order

1) ('e', 's')
2) ('es', 't')
3) ('est', '</w>')
4) ('l', 'o')
5) ('lo', 'w')
6) ('n', 'e')
7) ('ne', 'w')
8) ('new', 'est</w>')
9) ('low', '</w>')
10) ('w', 'i')

Fast campus

Subword Segmentation Modules

- Subword-nmt
  - https://github.com/rsennrich/subword-nmt
- WordPiece
  - Upgrade BPE version. Currently unavailable...?
- SentencePiece
  - https://github.com/google/sentencepiece

OoV가 미치는 영향

- 입력 데이터에 OoV가 발생할 경우, <UNK> 토큰으로 치환하여 모델에 입력
  - e.g. 나는 학교에 가서 밥을 먹었다. → 나 는 <UNK> 에 가 서 <UNK> 을 먹 었 다 .

- 특히, 이전 단어들을 기반으로 다음 단어를 예측하는 task에서 치명적
  - e.g. Natural Language Generation

- 어쨌든 모르는 단어지만, 알고있는 subword들을 통해 의미를 유추해볼 수 있음
  - e.g. 버카충

Summary

- BPE 압축 알고리즘을 통해 통계적으로 더 작은 의미 단위(subword)로 분절 수행

- BPE를 통해 OoV를 없앨 수 있으며, 이는 성능상 매우 큰 이점으로 작용

- 한국어의 경우
  - 띄어쓰기가 제멋대로인 경우가 많으므로,
    normalization 없이 바로 subword segmentation을 적용하는 것은 위험
  - 따라서 형태소 분석기를 통한 tokenization을 진행한 이후,
    subword segmentaion을 적용하는 것을 권장