Sub-word tokenization methods

Sang Yup Lee

주요 토큰화 방법들

- Byte Pair Encoding (BPE)
  - Byte-level Byte Pair Encoding (BBPE)
- WordPiece Encoding
- SentecePiece Encoding

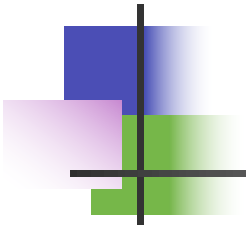
Byte Pair Encoding (BPE)

BPE

- 토큰화 단계
  - 1) 사전 구축
  - 2) 사전을 이용한 토큰화
- BPE 에서의 사전 구축
  - 사전 구축을 위한 데이터셋
    - 예) 사용된 단어들:
      'first', 'best', 'song', 'soft', 'son'

단어	빈도
f, i, r, s, t	2
b, e, s, t	2
s, o, n, g	1
s, o, f, t	1
s, o, n	1

<말뭉치 데이터에 존재하는 단어들과 단어들의 출현 빈도>



BPE

- 사전 구축 (cont'd)
  - 생성하고자 하는 사전의 크기 (사전에 포함하고자 하는 토큰의 수) 결정
    - 가정: 사전 크기 = 13
    - 순서
      - 말뭉치 데이터에 존재하는 모든 고유한 문자들([b, e, f, g, i, n, o, r, s, t])을 사전에 추가
        - 사전 = { b, e, f, g, i, n, o, r, s, t}
        - 세 개의 토큰 추가 필요
      - 가장 자주 출현하는 두 개 이상의 문자들로 구성된 문자열을 찾고 해당 문자열을 새로운 토큰으로 추가 => 'st'
        - 사전 = { b, e, f, g, i, n, o, r, s, t, st}
      - 최정 사전: { b, e, f, g, i, n, o, r, s, t, st, so, son}

BPE

- 구축된 사전을 이용한 토큰화
  - 가정: 토큰화를 하고자 하는 텍스트 데이터 = ‘soft’
  - 순서
    - s, o, f, t 로 분할
    - so, of, ft 가 사전에 있는지 파악 => so 존재
    - ft가 있는지 파악 => 없음 => f,t로 구분
    - 최종 결과 => [so, f, t]
  - 그렇다면 데이터 = ‘sole’ 의 경우는?

Other methods

- WordPiece
  - BPE와 WordPiece가 갖는 주요한 차이는 입력된 시퀀스 데이터를 토큰 단위로 분할하는 과정에 있는 것이 아니라, 토크나이저가 갖는 어휘 사전을 구축하는 과정 차이
  - 어휘 사전을 구성하는 토큰을 정할 때 WordPiece는 토큰의 우도(likelihood, 즉 해당 토큰이 데이터에 존재할 확률을 의미합니다)를 이용해 우도가 큰 순서대로 어휘 사전에 추가
  - 예) st의 likelihood

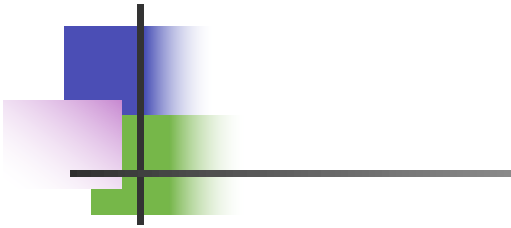
$$\frac{p(st)}{p(s)p(t)} = \frac{\frac{\#st}{n}}{\frac{\#s}{n} \times \frac{\#t}{n}}$$

Other methods

- BBPE
  - BPE와 유사, 하지만, 바이트 단위
  - BPE가 문자(character) 단위로 작동하는 반면, BBPE는 바이트 단위(byte level)로 작동
  - 다국어를 처리할 때 유리
    - 영어: a character = one byte (8 bits)
    - 한글: a character >= two bytes (16 bits)
      - UTF-8의 경우, a character = three bytes
      - '위' => 11101100 10011100 10000100

Other methods

- SentencePiece
  - 앞의 방법들은 공통적으로 공백문자를 기준으로 구분
  - 공백문자를 기준으로 단어로 구분되지 않는 언어 (중국어, 일본어 등)
  - SentencePiece는 공백문자도 사전에 포함
  - 사전을 구축하기 위해 BPE (또는 Unigram) 방법 사용
  - Unigram
    - 토큰을 추가하는 식으로 사전을 구축하지 않고, 많은 수의 토큰에서 토큰을 제거하는 방식으로 사전을 구축
    - 단독으로 사용되지 않고, SentencePiece 에서 사용



Q & A