BPE

1. BPE (Byte Pair Encoding)

· 연속되는 글자 및 데이터의 쌍을 찾아 해당 데이터 및 글자로 치환하는 것.

ex) aaabd aaabac

여기서 Z = aa 라고 하면,

ZabdZabac

Y = ab,

ZYdZYaC

X = ZY

XdXac 데이터를 획기적으로 줄여는 것이 가능함.

2. 자연어 에서의 BPE.

· 단어를 여러개의 Segmentation으로 분리하는 알고리즘.

↳ 단어의 OOV 문제를 획기적으로 줄여는 것이 가능!

* 순서.

1) 단어를 최소 단위인 자모 단위로 분할.

2) 자모의 쌍 (bigram) 단위로 최대 빈도의 단어를 찾아 새로운 자모의 단위로 생성.

↳ (2)를 BPE 처럼 반복하여 실행.

이렇게 분할된 단어는 새로운 단어가 입력되어도 OOV 문제를 피할 수 있음.

3. Word Piece Embedding.

· 한개의 단어 (word)를 sub-word로 분할하여 OOV 문제를 해결

· 기본적 개념은 BPE와 동일.

· 자주 사용되는 단어의 경우 단어 그대로 사용이 가능.

- 자주 사용되지 않는 단어는 여러개의 subwords 로 분할하여 사용가능.