Additional topics

BPE - Byte Pair Encoding

Initial state:

{'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}

First step:

{'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}

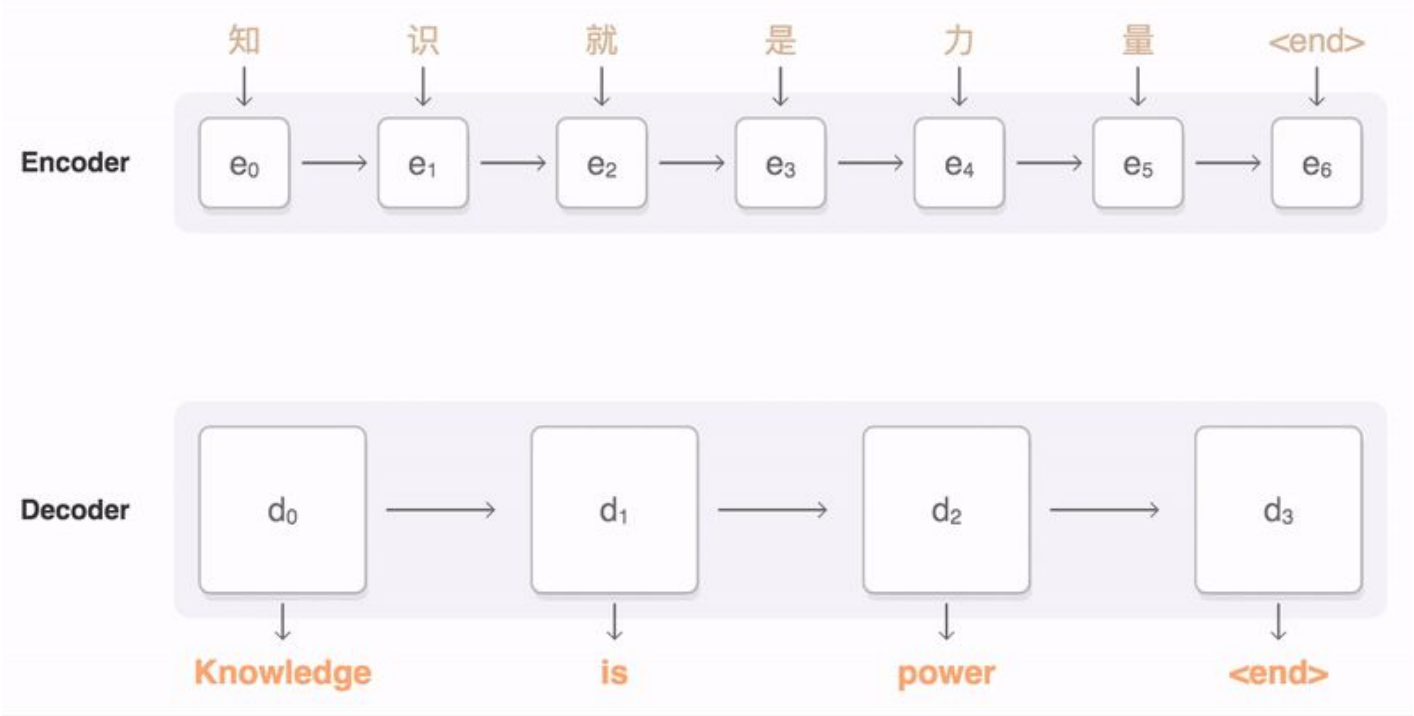
Second step:

{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w **est** </w>': 6, 'w i d **est** </w>': 3}

After each merge, there could be three scenarios, the number of tokens decreases by one, remains the same or increases by one. But in practice, as the number of merges increases, usually the number of tokens first increases then decreases.

More info

Recap


知
识
就
是
力
量
<end>
Encoder
e0
e1
e2
e3
e4
e5
e6
Decoder
d0
d1
d2
d3
Knowledge
is
power
<end>

GS - Greedy search

```
# define a sequence of 10 words over a vocab of 5 words
data = [[0.1, 0.2, 0.3, 0.4, 0.5],
        [0.5, 0.4, 0.3, 0.2, 0.1],
        [0.1, 0.2, 0.3, 0.4, 0.5],
        [0.5, 0.4, 0.3, 0.2, 0.1],
        [0.1, 0.2, 0.3, 0.4, 0.5],
        [0.5, 0.4, 0.3, 0.2, 0.1],
        [0.1, 0.2, 0.3, 0.4, 0.5],
        [0.5, 0.4, 0.3, 0.2, 0.1],
        [0.1, 0.2, 0.3, 0.4, 0.5],
        [0.5, 0.4, 0.3, 0.2, 0.1]]
```

```
Result = [4, 0, 4, 0, 4, 0, 4, 0, 4, 0]
```

More info

BS - Beam search

[[4, 0, 4, 0, 4, 0, 4, 0, 4, 0], 0.0256008632895631 08]

[[4, 0, 4, 0, 4, 0, 4, 0, 4, 1], 0.033842500435843 97]

[[4, 0, 4, 0, 4, 0, 4, 0, 3, 0], 0.033842500435843 97]