

Trick for Improving Model Performance

March 13, 2018

Jindřich Helcl, Jindřich Libovický



Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



Overview

Title slide

- item

BPE algorithm (1)

```
import re, collections

def get_stats(vocab):
    pairs = collections.defaultdict(int)
    for word, freq in vocab.items():
        symbols = word.split()
        for i in range(len(symbols)-1):
            pairs[symbols[i],symbols[i+1]] += freq
    return pairs

def merge_vocab(pair, v_in):
    v_out = {}
    bigram = re.escape(' '.join(pair))
    p = re.compile(r'(?<!\S)' + bigram + r'(?!\S)')
    for word in v_in:
        w_out = p.sub(' '.join(pair), word)
        v_out[w_out] = v_in[word]
    return v_out
```

BPE algorithm (2)

```
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}  
  
num_merges = 10  
for i in range(num_merges):  
    pairs = get_stats(vocab)  
    best = max(pairs, key=pairs.get)  
    vocab = merge_vocab(best, vocab)  
print(best)
```