Sub Model

Subword models

- Word embeddings as we have covered them so far assume atomic words and a fixed vocabulary.
- In practical applications, we will often encounter words that we do not have an embedding for.
- One way to deal with this problem is to use models that work at the subword level, such as character-based model

Different types of subword models

- Type 1: Use the same types of architectures that we find in word-based models, but apply them to subword units.
- Type 2: Augment the architectures of word-based models with submodels that compose word representations from characters.
- Type 3: Give up on word-based architectures altogether and process language as a connected sequence of characters.

WordPiece tokenization in BERT

Raw text:

The history of morphological analysis dates back to the ancient Indian linguist Pāṇini, who formulated the 3,959 rules of Sanskrit morphology in the text Aṣṭādhyāyī by using a constituency grammar.

WordPiece tokenization:

The history of m ##or ##phological analysis dates back to the ancient Indian linguist P ##ā ##ṇ ##ini , who formulated the 3 , 95 ##9 rules of Sanskrit morphology in the text A ##ṣ ##ṭ ##ā ##dh ##y ##ā ##y ##ī by using a constituency grammar .

WordPiece tokenization in BERT

Raw text:

The history of morphological analysis dates back to the ancient Indian linguist Pāṇini, who formulated the 3,959 rules of Sanskrit morphology in the text Aṣṭādhyāyī by using a constituency grammar.

WordPiece tokenization:

The history of m ##or ##phological analysis dates back to the ancient Indian linguist P ##ā ##ṇ ##ini , who formulated the 3 , 95 ##9 rules of Sanskrit morphology in the text A ##ṣ ##ṭ ##ā ##dh ##y ##ā ##y ##ī by using a constituency grammar .

Byte Pair Encoding algorithm

- Initialize the word unit vocabulary with all characters. plus a special end-of-word marker, here denoted by $
- Generate a new word unit by combining two units from the current vocabulary, increasing vocabulary size by one. Choose the new unit as the most frequent pair of adjacent units.
- Repeat the previous step as long as the vocabulary size does not exceed a maximal size.

Byte Pair Encoding algorithm

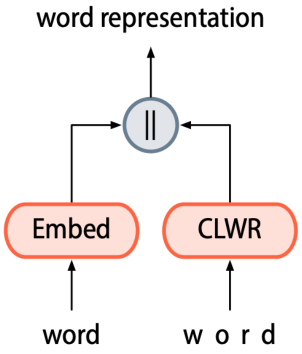
number of occurrences in data

Step	Merged pair	Words	Vocabulary size
0	–	low$/5 lower$/2 new**es**t$/6 wid**es**t$/3	11
1	es/9	low$ lower$ new[**es**]**t**$ wid[**es**]**t**$	12
2	[es]t/9	low$ lower$ new[**est**]**$** wid[**est**]**$**	13
3	[est]$/9	**lo**w$ **lo**wer$ new[est$] wid[est$]	14
4	lo/7	[**lo**]**w**$ [**lo**]**w**er$ new[est$] wid[est$]	15
5	[lo]w/7	[low]$ [low]er$ **new**[est$] wid[est$]	16

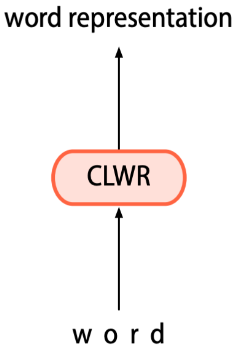
Example from Sennrich et al. (2016)

Composing word representations from characters

- Character-level word representations are typically built using convolutional neural networks or recurrent neural networks.



combined (augmented) model



purely character-based model

Composing word representations using CNNs

<pad>	0.00 **0.50**	0.00 **0.10**	0.00 **0.10**
d	0.08 **1.00**	0.95 **0.20**	0.85 **0.20**
o	0.98 **0.50**	0.78 **0.10**	0.02 **0.10**
c	0.32	0.13	0.82
t	0.64	0.28	0.92
o	0.05	0.25	0.77
r	0.88	0.59	0.66
<pad>	0.00	0.00	0.00

1.010		
1.615		
1.520		
1.262		
1.259		
1.257		

Composing word representations using CNNs

<pad>	0.00 **0.10**	0.00 **0.50**	0.00 **0.10**
d	0.08 **0.20**	0.95 **1.00**	0.85 **0.20**
o	0.98 **0.10**	0.78 **0.50**	0.02 **0.10**
c	0.32	0.13	0.82
t	0.64	0.28	0.92
o	0.05	0.25	0.77
r	0.88	0.59	0.66
<pad>	0.00	0.00	0.00

1.010	1.626	
1.615	1.727	
1.520	1.144	
1.262	0.978	
1.259	1.159	
1.257	1.050	

Composing word representations using CNNs

<pad>	0.00 0.10	0.00 0.10	0.00 0.50
d	0.08 0.20	0.95 0.20	0.85 1.00
o	0.98 0.10	0.78 0.10	0.02 0.50
c	0.32	0.13	0.82
t	0.64	0.28	0.92
o	0.05	0.25	0.77
r	0.88	0.59	0.66
<pad>	0.00	0.00	0.00

1.010	1.626	1.242
1.615	1.727	1.355
1.520	1.144	1.648
1.262	0.978	1.974
1.259	1.159	1.859
1.257	1.050	1.369

Composing word representations using CNNs

<pad>	0.00	0.00	0.00
d	0.08	0.95	0.85
o	0.98	0.78	0.02
c	0.32	0.13	0.82
t	0.64	0.28	0.92
o	0.05	0.25	0.77
r	0.88	0.59	0.66
<pad>	0.00	0.00	0.00

1.010	1.626	1.242
1.615	1.727	1.355
1.520	1.144	1.648
1.262	0.978	1.974
1.259	1.159	1.859
1.257	1.050	1.369
max ↓	*max* ↓	*max* ↓
1.615	1.727	1.974

character-level word representation

Composing word representations using CNNs

Filter specification

Width	Channels
1	32
2	32
3	64
4	128
5	256
6	512
7	1024