BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages

Benjamin Heinzerling[1,2] **Michael Strube**[2]

[1]AIPHES [2]Heidelberg Institute for Theoretical Studies

AIPHES
ADAPTIVE PREPARATION OF INFORMATION FROM HETEROGENEOUS SOURCES

Heidelberg Institute for Theoretical Studies
HITS

Myxomatosis

-osis
"state, abnormal condition, or action"

-oma / **-omato**
Forming nouns indicating tumors or masses

myxo
From Ancient Greek *múxa*, "mucus"

Computational Approximations to Morphological Analysis

1. Split into **subwords**
   characters: *m, y, x, o, m, a, t, o, s, i, s*
   ngrams, e.g.: *myx, yxo, xom, oma, mat, ato, tos, osi, sis*
   FastText: *myx + yxo + xom + oma + mat + ato + tos + osi + sis + myxo + yxom + xoma + omat + mato + atos + tosi + osis + myxom + . . . + xomato + omatos + matosi + atosis*
   byte pairs
2. Learn function that infers word meaning from subwords

Byte-Pair Encoding (BPE) (Gage, 1994)

A B A B C A B C D

Most frequent pair: A B

Merge pair A B to X:
X X C X C D

Most frequent pair: X C

Merge pair X C to Y:
X Y Y D

Symbol table:
X: A B
Y: X C

BPE for Text (Sennrich et al., 2016)

A B A B C A B C D

Most frequent pair: A B

Merge symbol pair A B into AB:
AB AB C AB C D

Most frequent pair: AB C

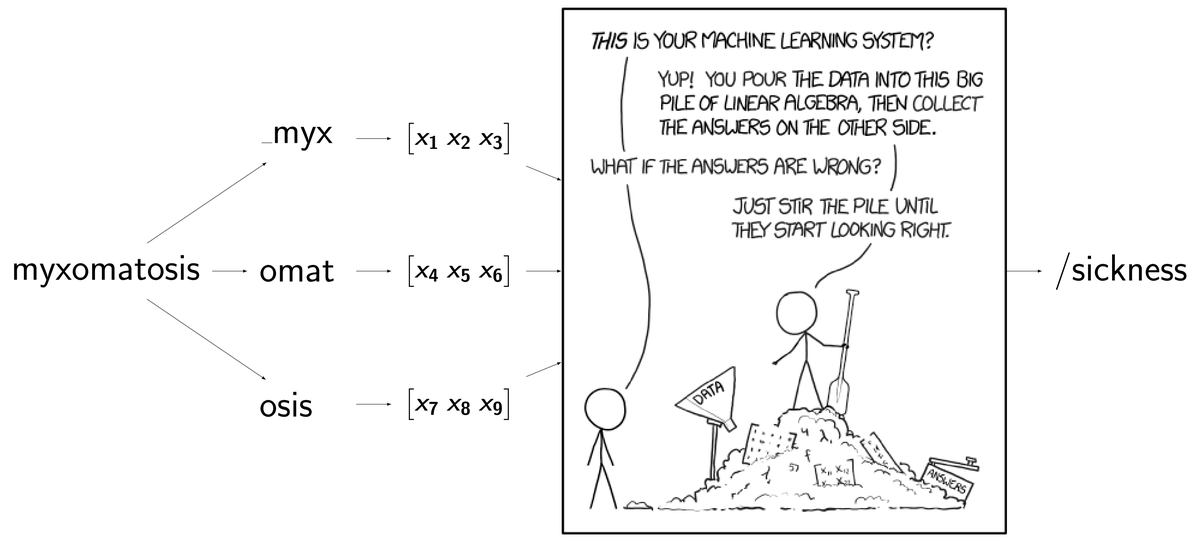
Merge pair AB C into ABC:
AB ABC ABC D

Symbol table:
A B: AB
AB C: ABC

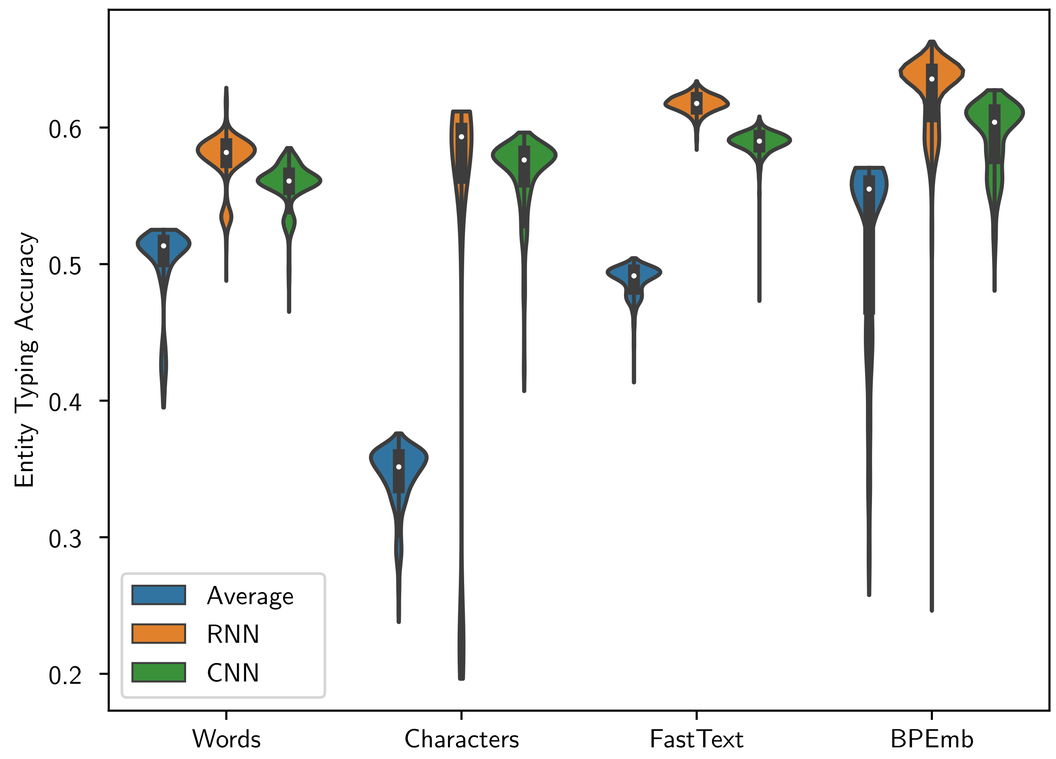
BPE Applied to English Wikipedia

_ t → _t	_o	stit
_ a → _a	_w	ery
h e → he	is	ister
i n → in	an	. . .
_t he → _the	_in	igo
er	. . .	osis
_s	ough	_jose
on	_series	. . .
_c	int	omatosis
re	ai	. . .

Subword-based Entity Typing



Results: English Entity Typing



Unsupervised Segmentation with BPE

Merge ops	Byte-pair encoded text
1000	to y od a _station _is _a _r ail way _station _on _the _ch ū ō _main _l ine
3000	to y od a _station _is _a _railway _station _on _the _ch ū ō _main _line
10000	toy oda _station _is _a _railway _station _on _the _ch ū ō _main _line
50000	toy oda _station _is _a _railway _station _on _the _chū ō _main _line
100000	toy oda _station _is _a _railway _station _on _the _chūō _main _line
Tokenized	toyoda station is a railway station on the chūō main line
10000	豐 田 站 是 東 日本 旅 客 鐵 道 (JR 東 日本) 中央 本 線 的 鐵路 車站
25000	豐田 站是 東日本旅客鐵道 (JR 東日本) 中央 本 線的鐵路車站
50000	豐田 站是 東日本旅客鐵道 (JR 東日本) 中央 本線的鐵路車站
Tokenized	豐田站 是 東日本 旅客 鐵道 （ JR 東日本 ） 中央本線 的 鐵路車站
5000	豊 田 駅 (と よ だ え き) は 、 東京都 日 野 市 豊 田 四 丁目 にある
10000	豊 田 駅 (と よ だ えき) は 、 東京都 日 野市 豊 田 四 丁目にある
25000	豊 田駅 (とよ だ えき) は 、 東京都 日 野市 豊田 四 丁目にある
50000	豊 田駅 (とよ だ えき) は 、 東京都 日 野市 豊田 四丁目にある
Tokenized	豊田 駅 （ とよ だ え き ） は 、 東京 都 日野 市 豊田 四 丁目 に ある

Download Embeddings and BPE Models in 275 Languages

https://github.com/bheinzerling/bpemb

Acknowledgements: This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.