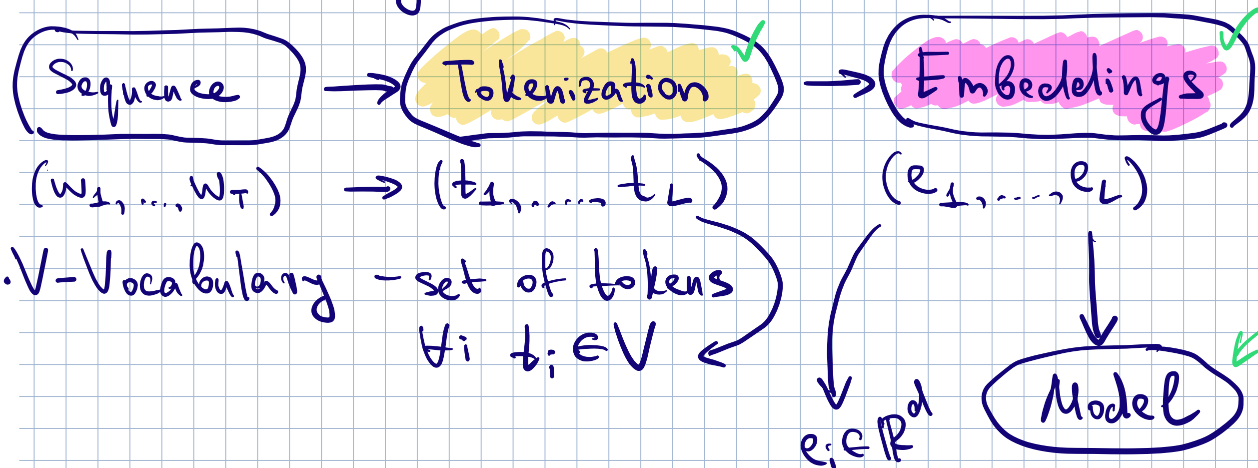
Бобры добры, добротой бобры полны.

Sequence processing

Processing pipeline

- Texts are discrete
- Variable length



- V - Vocabulary - set of tokens
  $\forall i \; t_i \in V$

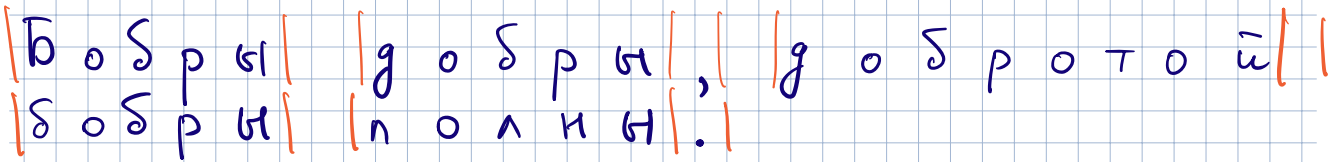
Tokenization

- Character-level

|Б|о|б|р|ы| |д|о|б|р|ы|,| |д|о|б|р|о|т|о|й| |
|б|о|б|р|ы| |п|о|л|н|ы|.|

\+ Very small vocab
\+ Easy to encode new words
− Lose language structure
− Too long sequences

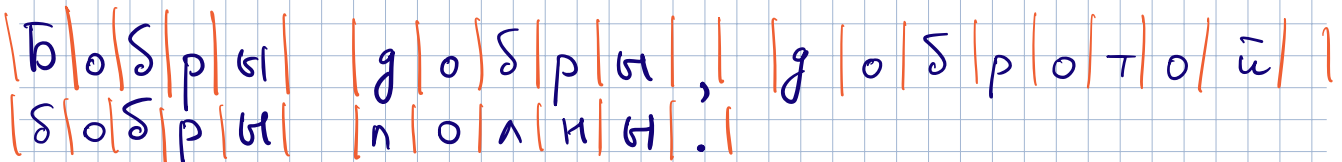
- Word-level



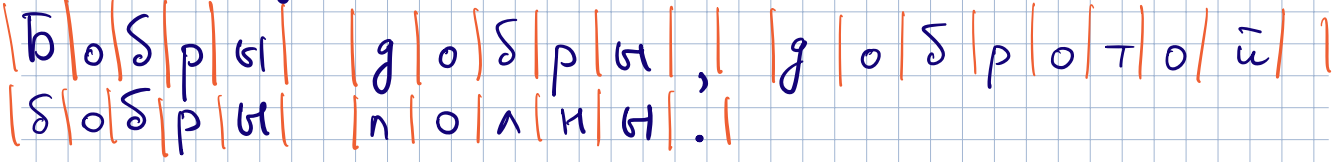
- Very large vocab
- No tokens for new words
  - <UNK> - token for unknown words
+ Preserve language structure
+ Shorter sequences

- Subword-tokenization

Byte-pair encoding (BPE)



1) Start with character-level tokenization
2) while size(V) < vocab_size
   - Find most frequent pair of tokens
   - Merge 2 tokens, add to vocab



0) {S, O, P, Ч, ' ', g, ш, T, Н, Л, n, '.'}

|Б|о|б|р|ы| |д|о|б|р|ы|,| |д|о|б|р|о|т|о|й| |
|б|о|б|р|ы| |п|о|л|н|ы|.|

1) {б, о, р, ы, ' ', д, й, т, н, л, п, '.', бр}

|Б| об |р|ы| |д|о| б |р|ы|,| |д|о| б |р|о|т|о|й| |
|б|о б |р|ы| |п|о|л|н|ы|.|

2) {б, о, р, ы, ' ', д, й, т, н, л, п, '.', бр, обр}

Embeddings

Word2Vec

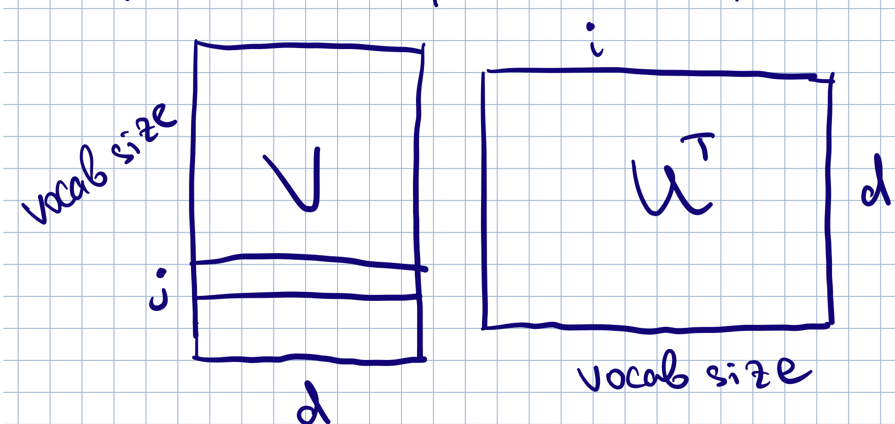
w_1 [w_2 w_3 w_4 w_5 w_6] w_7 ...

- Continuous Bag of Words (CBOW)

  $p(w_4 \mid w_2, w_3, w_5, w_6)$

- Skip-gram

  $p(w_2 \mid w_4)$, $p(w_3 \mid w_4)$, $p(w_5 \mid w_4)$, $p(w_6 \mid w_4)$



d - dim of embeddings

$$p(w_i|w_j) \approx \frac{e^{v_j^T u_i}}{\sum_{k=1}^{\text{vocab size}} e^{v_j^T u_k}}$$

$v_j \in \mathbb{R}^d$

$u_i \in \mathbb{R}^d$

$$-\sum_{\substack{w_j - \text{center} \\ w_i - \text{context}}} \log p(w_i|w_j) \to \min_{u,v}$$

For CBOW $p(w_j | w_{j-H}, w_{j-H+1}, \dots, w_{j-1}, w_{j+1}, \dots, w_{j+H})$

$$(v_{j-H} + v_{j-H+1} + \dots + v_{j+H})^T u_j$$

Multiclass $\to$ Binary classification

w_i – Context word
w_j – center word
} "Do these words appear together"

$p(w_i \text{ from context of } w_j) = \sigma(v_j^T u_i)$

$$-\sum_{\substack{w_j - \text{center} \\ w_i - \text{Context}}} \log \sigma(v_j^T u_i) \to \min_{u,v}$$

Negative sampling

$$-\sum_{\substack{w_j - \text{center} \\ w_i - \text{context}}} \log \sigma(v_j^T u_i) - \sum_{(w_j, w_i) - \text{neg. pair}} \log (1 - \sigma(v_j^T u_i)) \to \min_{u,v}$$

nn. Embedding index $\to$ token

- V
- U^T
- $\frac{1}{2}(V + U^T)$

man ⟷ king

woman ⟷ queen

$V_{man} - V_{king} \approx V_{woman} - V_{queen}$

Fast Text - extended Skip-Gram

[б о б р ы **д о б р ы**, д о б р о т о й
б о б р ы п о л н ы .]

p(бобры | добры) — Skip-gram

добры → <добры>, <до, доб, обр, бры, ры>

p(<бобры> | <добры>)   n = 3

p(<бобры> | <до)

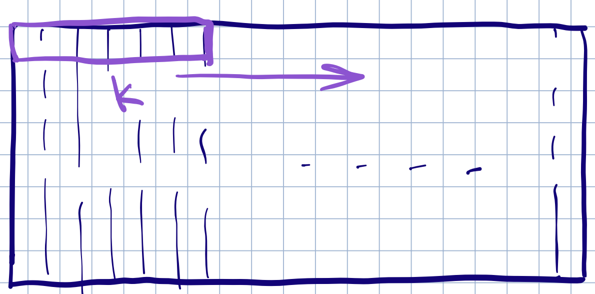
p(<бобры> | доб)

...

GloVe - Global Vectors

Text CNN   1-D convolutions

d



L

d - channels

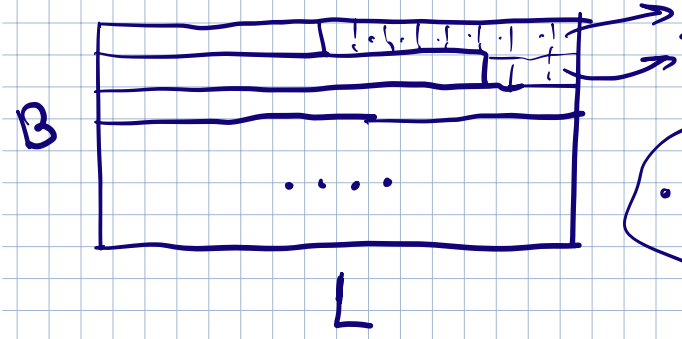
L - similar to height and width

Conv. kernel $K \times 1 \times d \times d'$

$d \times L \rightarrow d' \times L'$

- Variable length

$B \times C \times H \times W$



<PAD> - padding token

- Average pooling