Transfer learning and Data augmentation for Mail.ru questions

Sorokin Semen

Initial work plan

1. EDA
2. Baseline tf-idf with 3 layer perceptron
3. Advanced model on fast text
4. Embeddings from transformer models
    a. **Bert-base-multilingual-uncased**
    b. **Bert-base-multilingual-cased**
    c. **Xlm-mlm-17-1280**
    d. **Elmo**
5. Data augmentation
6. Result comparing

Baseline - Perceptron

Model features

- Xavier weight initialization
- Batch norm
- 3 layers linear
- Dropout - 0.2

Data representation:

- TF-idf - ngram = (1,1)

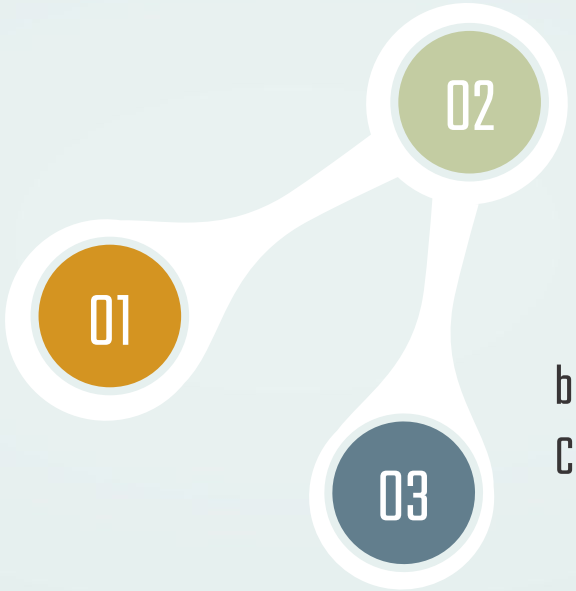
Result: 59.2

Preprocessing

Options:

- As is
- English or Russian words
- Normalized (En\Ru) words

Advanced model



Pad sequences
Batch norm
Xavier distribution

Self-Attention
DropOut (5 times)

biLSTM
CNN(3,4,5 grams)

Embeddings

bERT
(UNcased, multiling)
56,7

eLMO
0

bERT
(cased, multiling)
55,9

xLM
56,9

Answers usage

Tagging answer using question category

IN DEPTH

- Two layer biLSTM
- Two attention heads
- Increase number of model parameters
- Less Dropout value
- Data = 3,3 million, 70% unknown words :(
- Train on extended data ~4 hours
- Adopt to the question domain: 1 epoch on “true” questions

Result: 64,45

Thank you for attention