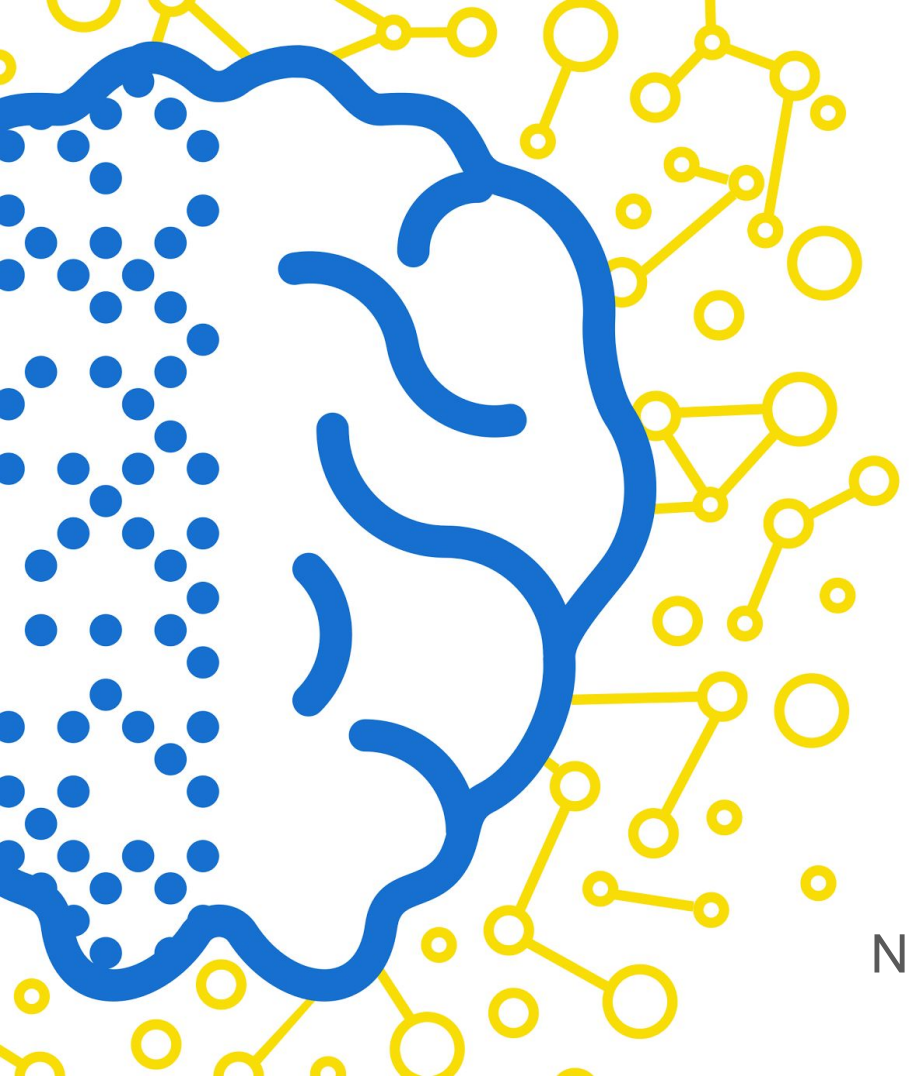
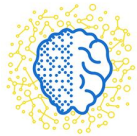


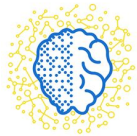
How to solve NLP tasks with DeepPavlov

Dilyara Baymurzina
Neural Networks and Deep Learning Lab



Plan

1. What NLP is?
2. NLP tasks:
    a. Text Vectorization
    b. Text Classification
    c. Sequence Tagging: Named Entity Recognition, Morphological Tagging
    d. Question Answering
    e. Text Ranking
    f. Text Generation (seq2seq)
3. `DeepPavlov` links

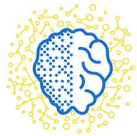


Natural Language Processing (NLP)

Natural language processing is a subfield of Computer Science, Information Engineering, and Artificial Intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

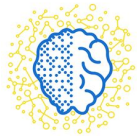
So, NLP is about:

- processing text to machine-readable format
- text understanding and information extraction
- text generation



Text Vectorization

- Embedding is a vector representation of character/word/text.
- Types:
  - one-hot vectors
  - TF-IDF vectors
  - word2vec
  - GloVe
  - fastText
  - ELMo: Embeddings from Language Models
- Embedding models can be trained either for specific language/domain or multi-lingual.
- Embeddings can be trained together with target model or be frozen.



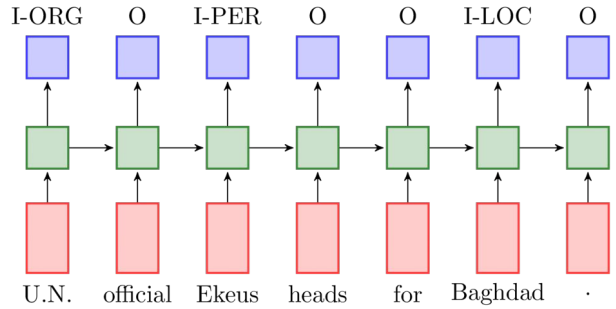
Text Classification

- The task is to assign each document *d* from collection *D* with label(s) *y*.
- Classification can be:
  - binary
  - multi-class
  - multi-label
- Text classification can be used for:
  - sentiment analysis, moderation
  - intent recognition
  - topic labeling
  - language detection
- Text classifiers in `DeepPavlov` are:
  - ML models from `sklearn`
  - neural networks: `Keras` models with `TensorFlow` backend, BERT-based on `TensorFlow`

`docs.DeepPavlov.ai`

Sequence tagging

- The task is to assign each word (token) *w* from document *d* with tag *y*.
- Sequence tagging can be used for:
  - Named Entity Recognition (NER)
  - Morphological Tagging
- Sequence taggers in `DeepPavov` are:
  - NER: standard RNN approach on `TensorFlow`, BERT-based model on `TensorFlow`
  - Morpho-tagger: char-level RNN on `Keras` with `TensorFlow` backend



docs.DeepPavlov.ai

https://www.depends-on-the-definition.com/guide-sequence-tagging-neural-networks-python/

Question Answering with Context

- Question Answering with Context is a task to find an answer on question in a given context, where the answer to each question is a segment of the context.
- Questions Answering models in `DeepPavov` are:
  - R-Net on `TensorFlow`
  - BERT-based on `TensorFlow`

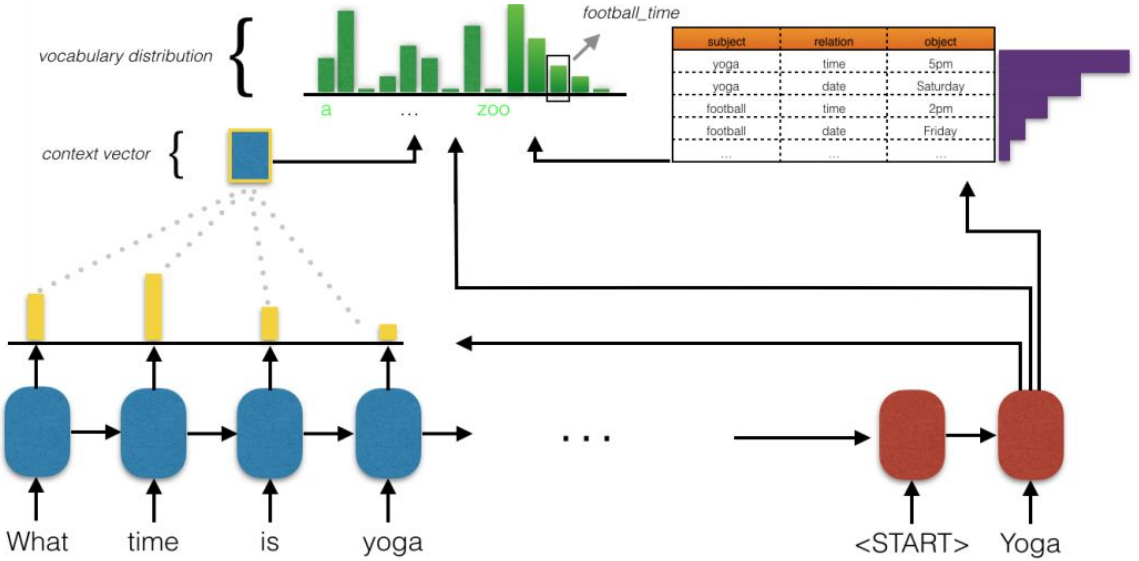
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	XLNet (single model) *XLNet Team*	**89.898**	**95.080**
2 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
6 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835

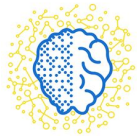
Text Ranking

- Paraphrase identification is a task of labeling pair of sentences whether the second one is a paraphrase of the first one or not.
- Text Ranking is a task of choosing the response closest semantically to a given context from some database.
- Text Ranking models in DeepPavlov are:
  - siamese neural network on Keras with TensorFlow backend
  - BERT-based neural network

Text Generation

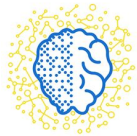
For each time-step of decoding, the cell state is used to compute an attention over the encoder states and a separate attention over the key of each entry in the KB. The attentions over the encoder are used to generate a context vector which is combined with the cell state to get a distribution over the normal vocabulary.





DeepPavlov Components

- Text Vectorization: http://docs.deeppavlov.ai/en/master/components/data_processors.html
- Pretrained ELMo: http://docs.deeppavlov.ai/en/master/intro/pretrained_vectors.html?highlight=elmo
- Text Classification: http://docs.deeppavlov.ai/en/master/components/classifiers.html
- Named Entity Recognition: http://docs.deeppavlov.ai/en/master/components/ner.html
- Morphological Tagging: http://docs.deeppavlov.ai/en/master/components/morphotagger.html
- Question Answering: http://docs.deeppavlov.ai/en/master/components/squad.html
- Text Ranking: http://docs.deeppavlov.ai/en/master/components/neural_ranking.html
- Text Generation (seq2seq): http://docs.deeppavlov.ai/en/master/skills/seq2seq_go_bot.html



Thank you for attention!

docs.DeepPavlov.ai