Improving neural morpheme segmentation for Russian word forms

Anastasia Kravtsova

Morpheme segmentation as sequence labeling

- BMES scheme:

  у ч и т е л ю
  B E S B M E S

- Morpheme types can be included:

  у ч и т е л ю
  B-ROOT E-ROOT S-SUFF B-SUFF M-SUFF E-SUFF S-END

- Looks similar to Named Entity Recognition

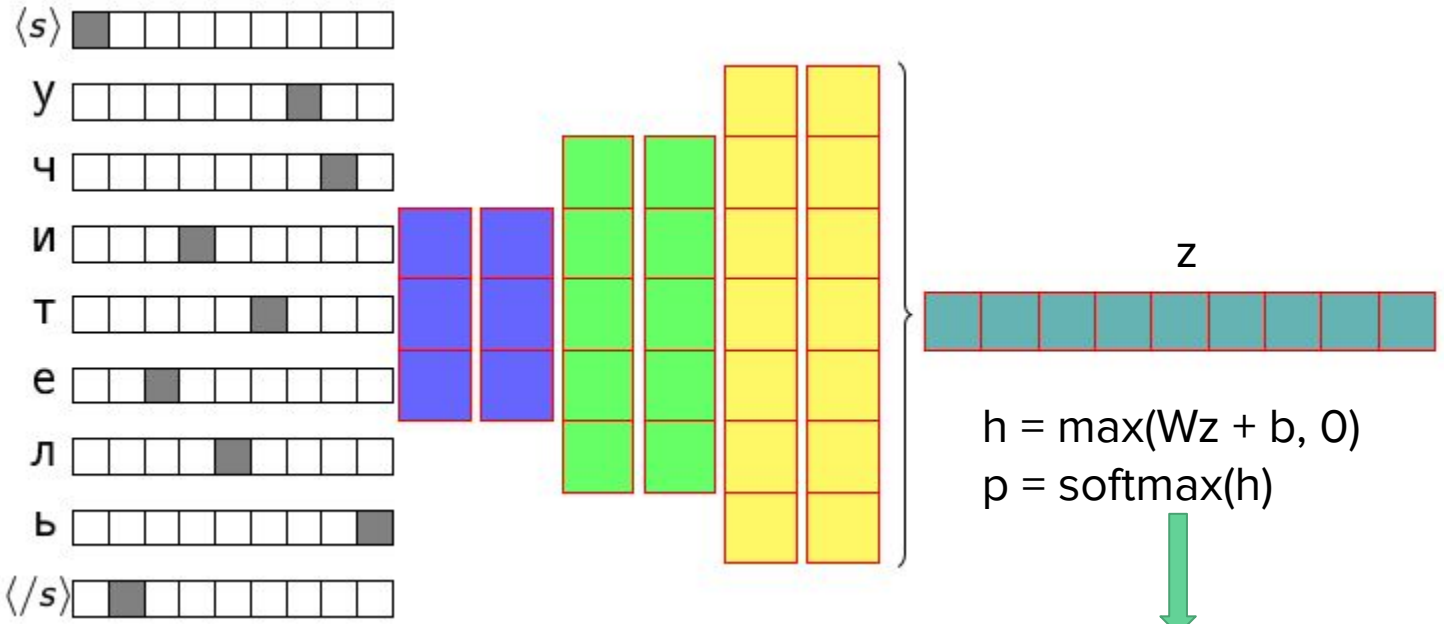
Data

- ru.wiktionary.org-paradigms for words from Morpheme dictionary of Tikhonov
- Labeled on morphemes using regular expressions
- Fixed sequence length mismatches
- Approx. **3/1** train/test partition: **216536**/**78365** word forms
- **4/1** train/validation partition
- **7** morpheme types:

PREFIX **от**тянетесь
SUFFIX оття**н**етесь
POSTFIX оттянете**сь**
HYPHEN сине-зеленому

ROOT от**тян**етесь
ENDING оттян**ете**сь
LINK тепл**о**хода

Model architecture



Results

# layers	# filters with conv windows	Precision	Recall	F1-score	Accuracy	Word accuracy
1	192 (7)	93.54	95.63	94.57	90.17	67.55
2	192 (7)	94.13	96.88	95.48	91.83	72.91
3	96+96 (5,7)	94.27	96.81	95.52	91.84	73.22
	192 (5)	94.31	**97.07**	95.67	92.03	73.89
	192 (7)	**94.38**	**97.05**	**95.70**	**92.11**	**74.00**
DP ner_rus		89.56	89.76	89.66	86.05	48.84

Conclusions

- More layers behave better.
- Dropout between convolutional layers improves performance.
- Adding recurrent layers is harmful.
- Computation on DGX is cool :)