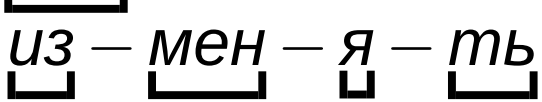
Comparing Models of Morpheme Analysis for Russian Words based on Machine Learning

Elena Bolshakova
Alexander Sapin

Moscow State Lomonosov University, CMC

Morpheme Analysis (Parsing)

Breaking words into constituent morphs (root and affixes)



- Morphs are the smallest meaningful units of texts
- NLP applications for morpheme parsing:
  - constructing word embeddings
  - handling rare and out-of-vocabulary words
  - recognition of paronyms
  - machine translation
- Difficulties for languages with rich morphologies:
  - Russian: many suffixes
- Various approaches to the task including machine learning

Morpheme Segmentation and Classification

Two variants of morpheme parsing:

- **segmentation** – splitting a word into morphs or morpheme-like units:

  пре – *крас* – *н* – *ый* *beauti* – *ful*

  - **Quality Metrics:** Precision, Recall, F-score on morpheme boundaries

- **segmentation with classification** of segmented morphs:

  пре/pref – *крас*/root – *н*/suff – *ый*/end *beauti*/root – *ful*/suff

  - **Quality Metrics:** Accuracy of classification

Segmentation Models

Harris method (1967, Harris S. Zellig)

- Letter variety statistics counted on dictionary words
- ≈61% of precision

Segmentation Models

Harris method (1967, Harris S. Zellig)

- Letter variety statistics counted on dictionary words
- ≈61% of precision

Morfessor (2003-2014, Creutz M., et al.)

- Semi-supervised machine learning
- Training on a text collection with help of segmented data
- ≈70% of F-measure for Finnish and Turkish

Segmentation Models

Harris method (1967, Harris S. Zellig)

- Letter variety statistics counted on dictionary words
- ≈61% of precision

Morfessor (2003-2014, Creutz M., et al.)

- Semi-supervised machine learning
- Training on a text collection with help of segmented data
- ≈70% of F-measure for Finnish and Turkish

seq2seq (2018, Arefyev N.V., et al.)

- Encoder-decoder neural network
- Supervised method, Tikhonov's dictionary for Russian
- ≈93% of F-measure for Russian.

Segmentation and Classification for Russian

CrossMorphy (2017, Sapin A.S., et al.)

- Conditional random fields (CRF)
- Classification of letters to the main types of morphs (4)

з	в	е	р	и	н	е	ц
R	R	R	R	S	S	S	S

- ≈79.5% of accuracy for CrossLexica's dictionary

Segmentation and Classification for Russian

CrossMorphy (2017, Sapin A.S., et al.)

- Conditional random fields (CRF)
- Classification of letters to the main types of morphs (4)

з	в	е	р	и	н	е	ц
R	R	R	R	S	S	S	S

- ≈79.5% of accuracy for CrossLexica's dictionary

Convolutional neural network (2018, Sorokin A., et al.)

- Classification of letters into 24 classes
- BMES labels for **B**egin, **M**iddle, **E**nd and **S**ingle morphs
- Postprocessing: correcting algorithm
- Word accuracy ≈88% for Tikhonov's dictionary

Segmentation and Classification for Russian

CrossMorphy (2017, Sapin A.S., et al.)

- Conditional random fields (CRF)
- Classification of letters to the main types of morphs (4)

з	в	е	р	и	н	е	ц
R	R	R	R	S	S	S	S

- ≈79.5% of accuracy for CrossLexica's dictionary

Convolutional neural network (2018, Sorokin A., et al.)

- Classification of letters into 24 classes
- BMES labels for **B**egin, **M**iddle, **E**nd and **S**ingle morphs
- Postprocessing: correcting algorithm
- Word accuracy ≈88% for Tikhonov's dictionary

Gradient Boosted Decision Trees (2019)

New Model: Gradient Boosted Decision Trees

Properties:

- Classification of letters into 10 classes (**E**nd is reduced)
- Simple postprocessing correcting algorithm
- 24 features for learning

Letter features:

- Window of 5 letters from both sides
- Vowel or consonant?
- Frequency of letter in training data
- Harris' values (letter variety statistics)

Word features:

- Morphological features of word: POS, case, gender, number
- Length of word and length of stem

URL: `https://github.com/alesapin/GBDTMorphParsing`

New Model: Details

Hyperparameters:

- Depth of trees: 10
- Loss function: Categorial CrossEntropy
- Iterations: 10000

Training:

- Catboost implementation
- CPU (Intel Xeon E5-2660v4, 256 GB RAM) $\approx$ 20 hours
- GPU (NVIDIA Tesla V100, 16 GB) $\approx$ 2.5 minutes
- Quality of GPU model is worse than CPU about 5%

Comparison of the Models

Motivation:

- To compare approaches and models on the same data
- To evaluate separately segmentation and segmentation with classification

Two different datasets:

- Tikhonov's dictionary
- CrossLexica's dictionary

Models under comparison:

- Morfessor
- CrossMorphy (CRF)
- seq2seq
- Convolutional Neural Network (CNN)
- Gradient Boosted Dicision Trees (GBDT)

Datasets

Tikhonov's dictionary

- A derivational dictionary, 96,046 words
- Not all prefixes and suffixes are separated:

продукт – *ивн* – *ый*
root suff end

забы – *ть*
root suff

- Many obsolete words: *снитковый*, *лядунка*

CrossLexica's dictionary:

- 23,426 words including words from modern lexicon:

эксклюзивный, целлюлит

- Dictionary for constructing morpheme paronyms:

болотный – болотистый

- Prefixes and suffixes are seprated:

про – *дукт* – *ив* – *н* – *ый*
pref root suff suff end

за – *бы* – *ть*
pref root suff

- Multi-rooted and hyphenated words are absent

Trained Models under Evaluation

Pre-trained Models:

- seq2seq and CNN models for Tikhonov's dictionary
- CRF model for CrossLexica dictionary

We have trained:

- Morfessor's model, on *lib.rus.ec* corpus
- Supervised models on both datasets:
  - seq2seq
  - CNN
  - GBDT
- Training datasets were randomly divided into 80:20

In total, we have 7 models

Experimental Evaluation

For segmentation: All the models

For segmentation and classification: CRF, CNN, GBDT

Metrics:

- Precision, Recall, F-score on boundaries for segmentation
- Accuracy for letter classification and for whole-word classification

Experiments:

- CNN and GBDT were also evaluated without correcting algorithm

Results for Segmentation

	CrossLexica's Dictionary			Tikhonov's Dictionary		
Model	Prec	Rec	F-measure	Prec	Rec	F-measure
Morfessor	93.3	75.4	83.4	94.7	73.7	82.9
CRF	96.05	70.93	81.60	–	–	–
seq2seq	94.62	93.92	94.27	94.07	93.83	93.95
CNN	98.68	98.75	98.72	**97.86**	**98.35**	**98.19**
GBDT	**98.84**	**99.26**	**99.05**	97.76	98.26	98.01

- CNN and GBDT are the best models (98-99% F-score)
- Morfessor and CRF lose in recall (82% F-score)
- seq2seq shows average quality (94% F-score)

Results for Segmentation with Classification

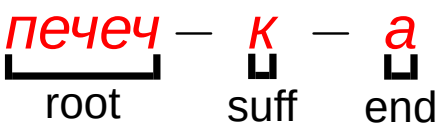
Comparison of the models that were the best in segmentation:

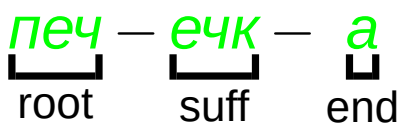
	CrossLexica's Dictionary			Tikhonov's Dictionary		
Model	Letters	Words		Letters	Words	
		Corrected	Uncorrected		Correted	Uncorrected
CNN	97.88	93.23	90.48	**96.64**	**88.71**	82.62
GBDT	**98.39**	**94.20**	**93.85**	96.40	86.54	**86.24**

- The models have close scores (state-of-the-art)
- GBDT is better on CrossLexica's data
- CNN is better for Tikhonov's data
- Correction is more crucial for CNN than GBDT: prefixes before roots, suffixes after roots and so on

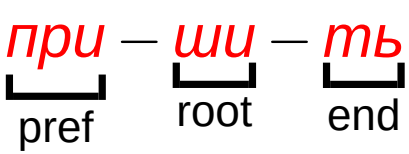
Models Comparison: Analysis of Errors

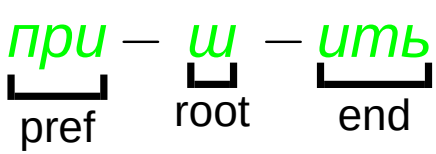
Wrong boundary between root and suffix (both datasets):



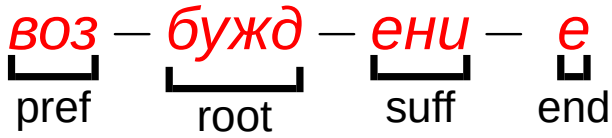


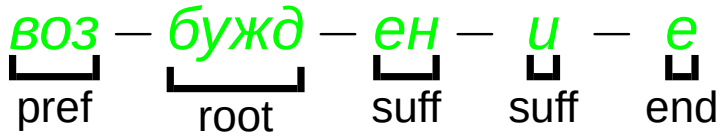
Wrong boundary between root and ending (CrossLexica):



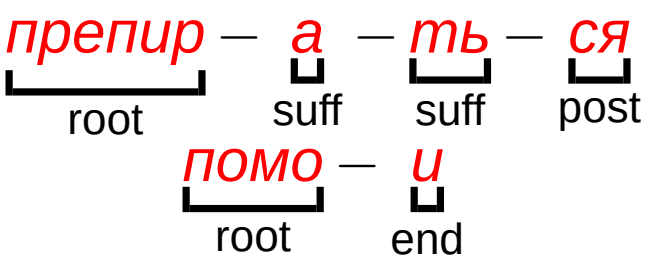


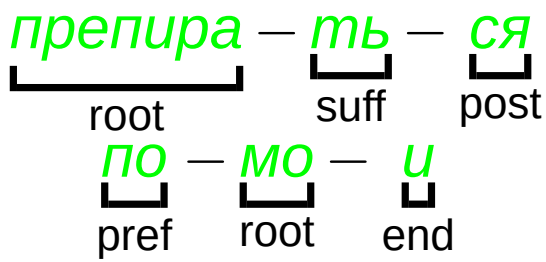
Wrong segmentation of suffixes (CrossLexica):





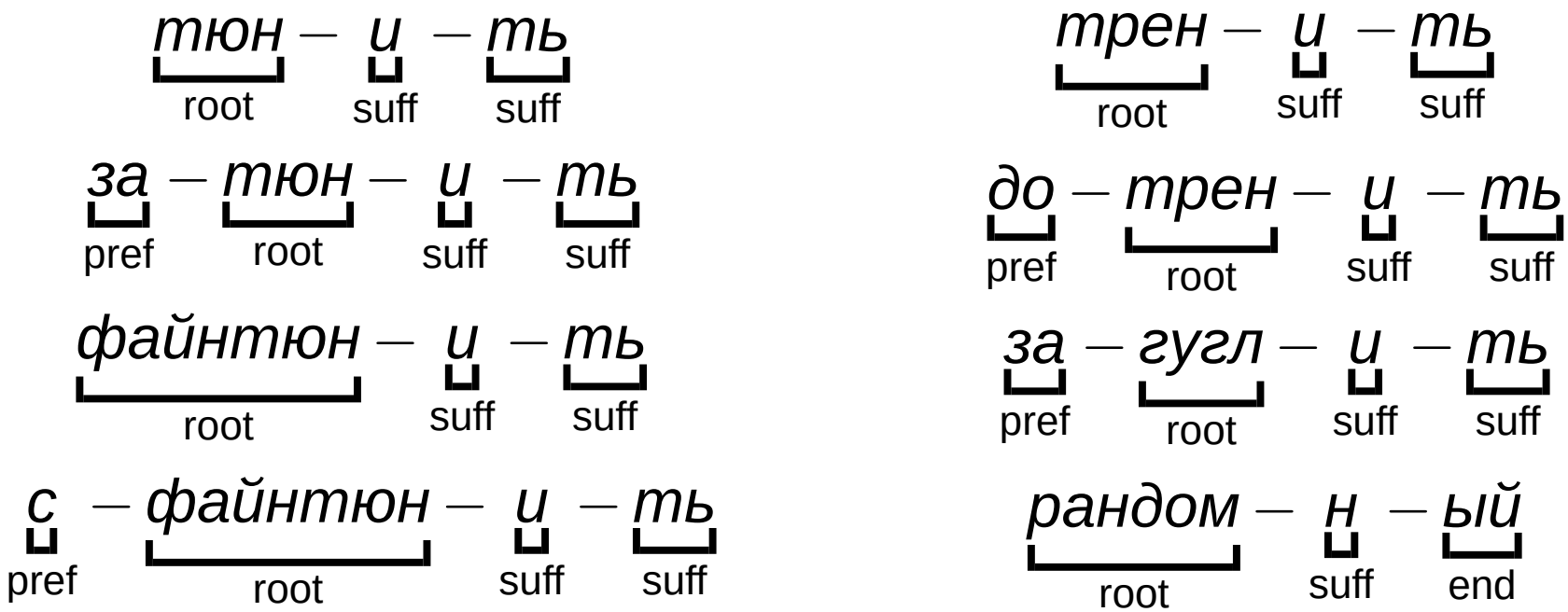
Complex errors examples (Tikhonov):





Examples of Parsing Words with New Roots

GBDT model trained on Tikhonov's dictionary:



Conclusion and Future Work

Conclusion:

- 5 various ML models for morpheme parsing has been evaluated and compared
- CNN and GBDT show the best and comparable results

Future work:

- Is it possible to improve current results?
- What is a gold standard for Russian morpheme segmentation?
- Is morpheme parsing useful for word embeddings?
- How can we use information about importance of features in GBDT?

Importance of GBDT Features

Letter	10.89
Letter[i-5]	3.5
Letter[i-4]	3.55
Letter[i-3]	5.74
Letter[i-2]	**9.19**
Letter[i-1]	**11.41**
Letter[i+1]	**11.79**
Letter[i+2]	**7.32**
Letter[i+3]	4.77
Letter[i+4]	3.35
Letter[i+5]	1.67

Vowelty	4.18
Position in word	1.77
Letter frequency	4.30
Harris (prefix)	4.19
Harris (suffix)	3.12
Part os speech	2.44
Case	1.92
Gender	1.27
Tense	0.52
Word length	1.01
Stem length	1.13

Thank you for attention

QA