Building Dataset and Morpheme Segmentation Model for Russian Word Forms

Alexander Sapin
Elena Bolshakova

Moscow State Lomonosov University, CMC

Morpheme Segmentation (Parsing)

A kind of morphological analysis:
breaking words into constituent morphs (root and affixes)

душ – *евн* – *ость*
из – *мен* – *я* – *ть*

soul – *ful* – *ness*
chang – *e*

- 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 and cognates
  - machine translation
- Difficulties for languages with rich morphologies:
  - Russian: many affixes (prefixes, suffixes, endings)

Morpheme Segmentation with 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 – *крас* – *н* – *ый* *beauti* – *ful*

  pref root suff end root suff

  - **Quality Metrics:** Precision, Recall, F-score + Accuracy of classification

Previous Segmentation Models

Harris' method (1967, Harris S. Zellig)

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

Previous 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

Previous 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

Pure supervised methods were absent until 2017 because there were no relevant labeled datasets of necessary size

- For Russian, such datasets have appeared in 2017-2018, e.g., dataset from Tikhonov's dictionary (96,046 lemmas)

Segmentation with Classification
SOTA Models for Russian Lemmas

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
- Word level accuracy ≈88% for Tikhonov's dataset

Segmentation with Classification
SOTA Models for Russian Lemmas

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
- Word level accuracy ≈88% for Tikhonov's dataset

Gradient Boosted Decision Trees (2019, Sapin A.S., et al.)

- Classification of letters into 10 classes (simplified BMES)
- 24 features: of letters and the word (4 morphological tags)
- Word accuracy: ≈86% for Tikhonov's dataset,
  ≈94% for CrossLexica's dataset

Segmentation with Classification
SOTA Models for Russian Lemmas

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
- Word level accuracy ≈88% for Tikhonov's dataset

Gradient Boosted Decision Trees (2019, Sapin A.S., et al.)

- Classification of letters into 10 classes (simplified BMES)
- 24 features: of letters and the word (4 morphological tags)
- Word accuracy: ≈86% for Tikhonov's dataset,
  ≈94% for CrossLexica's dataset

Bi-LSTM neural network (2019, Sapin A.S., et al.)

- Ensemble of three LSTM models
- Word accuracy: ≈89% for Tikhonov's dataset,
  ≈94.5% for CrossLexica's dataset

Towards Neural Model for Word Forms

- **Motivation:** All previous studies were conducted only for lemmas, but Russian texts contain significantly varying word forms

Towards Neural Model for Word Forms

- **Motivation:** All previous studies were conducted only for lemmas, but Russian texts contain significantly varying word forms
- A dataset with segmented word forms is required to train a neural model

Towards Neural Model for Word Forms

- **Motivation:** All previous studies were conducted only for lemmas, but Russian texts contain significantly varying word forms
- A dataset with segmented word forms is required to train a neural model
- **Idea:** to enrich Tikhonov's dataset (96,046 lemmas) with segmented word forms

Towards Neural Model for Word Forms

- **Motivation:** All previous studies were conducted only for lemmas, but Russian texts contain significantly varying word forms
- A dataset with segmented word forms is required to train a neural model
- **Idea:** to enrich Tikhonov's dataset (96,046 lemmas) with segmented word forms
- Difficulties of Russian word inflection:
  - Many word formation suffixes for verb forms
  - Alternating consonants and fluent vowels in affixes

  *звер:ROOT/ин:SUFF/**ец**:SUFF*
  *звер:ROOT/ин:SUFF/**ц**:SUFF/а:END*
  *звер:ROOT/ин:SUFF/**ц**:SUFF/у:END*

Procedure for Building Dataset

- Rule based procedure was elaborated, that produces segmentation for word forms, based on segmented lemmas

Procedure for Building Dataset

- Rule based procedure was elaborated, that produces segmentation for word forms, based on segmented lemmas
- Inflectional classes and dictionary from CrossLexica system were used to generate word forms
  - 636 flexion classes for nouns, adjectives, verbs, which take into account alternating consonants and fluent vowels
  - Segmentation labels for classes were added manually

Procedure for Building Dataset

- Rule based procedure was elaborated, that produces segmentation for word forms, based on segmented lemmas
- Inflectional classes and dictionary from CrossLexica system were used to generate word forms
  - 636 flexion classes for nouns, adjectives, verbs, which take into account alternating consonants and fluent vowels
  - Segmentation labels for classes were added manually
- Lemmas absent in CrossLexica’s dictionary were taken from OpenCorpora and their classes were atomatically restored

Procedure for Building Dataset

- Rule based procedure was elaborated, that produces segmentation for word forms, based on segmented lemmas
- Inflectional classes and dictionary from CrossLexica system were used to generate word forms
  - 636 flexion classes for nouns, adjectives, verbs, which take into account alternating consonants and fluent vowels
  - Segmentation labels for classes were added manually
- Lemmas absent in CrossLexica's dictionary were taken from OpenCorpora and their classes were atomatically restored
- Lemmas absent in CrossLexica and OpenCorpora were processed in semi-automatic manner with morphoprocessor
- Less than 2% lemmas of Tikhonov's dataset were discarded

Resulted Dataset with Word Forms

- More than 1.7 mln of segmented word forms: ≈ 367k nouns, ≈ 358k adjectives, ≈ 320k verbs, ≈ 580k participles, ≈ 75k short adjectives, ≈62k gerund and others
- Structured by inflexional groups (paradigms)
- An example of noun paradigm:

уголек	*угол:ROOT/ек:SUFF*
уголька	*уголь:ROOT/к:SUFF/а:END*
угольку	*уголь:ROOT/к:SUFF/у:END*
угольку	*уголь:ROOT/к:SUFF/у:END*
уголек	*угол:ROOT/ек:SUFF*
угольком	*уголь:ROOT/к:SUFF/ом:END*
угольке	*уголь:ROOT/к:SUFF/е:END*
угольки	*уголь:ROOT/к:SUFF/и:END*
угольков	*уголь:ROOT/к:SUFF/ов:END*
уголькам	*уголь:ROOT/к:SUFF/ам:END*

Resulted Dataset: Verb Paradigm

расшить	*рас:PREF/ши:ROOT/ть:END*
расшил	*рас:PREF/ш:ROOT/и:SUFF/л:SUFF*
расшила	*рас:PREF/ш:ROOT/и:SUFF/л:SUFF/а:END*
расшило	*рас:PREF/ш:ROOT/и:SUFF/л:SUFF/о:END*
расшили	*рас:PREF/ши:ROOT/л:SUFF/и:END*
разошью	*разо:PREF/шь:ROOT/ю:END*
разошьешь	*разо:PREF/шь:ROOT/ешь:END*
разошьет	*разо:PREF/шь:ROOT/ет:END*
разошьем	*разо:PREF/шь:ROOT/ем:END*
разошьете	*разо:PREF/шь:ROOT/ете:END*
разошьют	*разо:PREF/шь:ROOT/ют:END*
расшей	*рас:PREF/ш:ROOT/ей:SUFF*
расшейте	*рас:PREF/ш:ROOT/ей:SUFF/те:END*

Neural Model for Word Form Segmentation

Convolutional architecture was choosen because it much faster to train without loose of quality

Properties:

- 10 output classes of letters (only **B**egginings of morphs)
- 7 input features

Letter features: Window of 5 letters + Vowel or consonant?

Word features: Part of speech

Architecture:

- One-hot encoding input
- Three stacked CNN layers, dropout between layers
- Output: dense layer + softmax

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

Training the Model on Word Forms

- 70% for training, 10% for validation, 20% for testing
  - Implemeted with Keras framework
- Three models for three variants of training
  - Simple mixing: random mixing of all labeled word forms
  - Group mixing: random mixing of inflexional groups
  - Only lemmas: only lemmas are taken for training
- Metrics for Evaluation:
  - **Segmentation:** Precision, Recall, F1-measure
  - **Classification:** Accuracy for letters and for whole word

Evaluation Results: Segmentation

Model: Training set	**Word Forms**		
	Precision	Recall	F-measure
Simple mixing	99.56	99.71	99.63
Group mixing	98.22	99.05	98.63
Only Lemmas	86.56	90.67	88.57

Model: Training set	**Lemmas**		
	Precision	Recall	F-measure
Simple mixing	99.41	99.55	99.48
Group mixing	98.16	98.95	98.55
Only Lemmas	98.06	98.53	98.30

Evaluation Results: Classification Accuracy

Model: Training set	**Word Forms**		**Lemmas**	
	Letters	Words	Letters	Words
Simple mixing	99.42	97.34	99.15	96.40
Group mixing	97.66	91.06	97.54	91.03
Only Lemmas	81.67	41.02	97.13	89.32

- Model trained only on lemmas shows obviously poor quality on word forms
- Group mixing is the more proper way of evaluating
- Models on word forms slightly outperforms SOTA models even for lemmas (91.06% vs 87-88%)

Conclusion and Future Work

- The volume dataset with Russian word forms split into morphs and classified by main morpheme types was built
- The built neural model for word forms outperforms the SOTA results for lemmas, giving about 91% in word-level classification accuracy
- The built dataset and the implemented neural models are of free access
- More than 1500 errors were fixed in Tikhonov's dataset while creating the dataset for word forms
- Further improvement of segmentation models can be achieved only after accounting phonological features of morphs

Thank You for Attention

QA