Morpheme Segmentation for Russian: Evaluation of Convolutional Neural Network Models

Lyudmila Maltina (lpmaltina@gmail.com), Alexey Malafeev (aumalafeev@hse.ru)

National Research University Higher School of Economics
Nizhny Novgorod

1. TASKS

- Evaluating CNN models from [1] trained on a relatively small annotated dataset
- Hyperparameter tuning
- Creating a sample of words with previously unseen roots and evaluation on this dataset
- Error analysis

2. DATASET CHARACTERISTICS

Training, validation and test samples in the ratio of 40/30/30 (38,368/28,777/28,777 words) on the basis of Tikhonov's dictionary

Sample	Prefixes	Roots	Suffixes	Endings	Linking morphs	Postfixes	Average number of morphs per word
Train	0.114	0.319	0.367	0.137	0.036	0.028	3.824
Validation	0.116	0.318	0.367	0.135	0.036	0.029	3.836
Test	0.116	0.318	0.366	0.136	0.036	0.028	3.829
Previously unseen roots	0.022	0.436	0.377	0.145	0.012	0.006	2.726

3. CREATING A DATASET WITH PREVIOUSLY UNSEEN ROOTS

with at least one root absent in the training set (dictionaries [2, 3, 4]) → Adding cognate words using the service [5] → Manually decomposing the words

The sample (300 words) includes:
- **loan words** (*буккроссинг*)
- **terms** (*аденозинтрифосфорный*)
- **neologisms** (*загуглиться*)
- **words derived from proper names** (*неогумбольдтианство*)

4. HYPERPARAMETER TUNING

15 combinations [6], including two combinations proposed in [1] (#1 and # 4)

Model	Hyperparameters	Precision	Recall	F1-score	Word accuracy
# 13	convolutional layers: 4 width of filters: [5] filters: 192 dense output units: 64 dropout rate: 0.1 ensembled models: 3	0.962/ 0.963/ 0.784	0.956/ 0.956/ 0.809	0.959/ 0.959/ 0.796	0.823/ 0.824/ 0.544
# 15	convolutional layers: 4 width of filters: [5] filters: 192 dense output units: 64 dropout rate: 0.1 ensembled models: 5	0.962/ 0.962/ 0.792	0.956/ 0.956/ 0.804	0.959/ 0.959/ 0.798	0.822/ 0.823/ 0.536

What improves the performance?
- increasing the number of convolutional layers
- reducing the dropout rate
- using ensembles of 3 or 5 neural networks

5. ERROR ANALYSIS

From the words in the **test sample** that our best model made mistakes in, 100 words were randomly sampled.

Cause of the error, number of such errors (in parentheses)	Example (the correct segmentation is shown in parentheses)	Comment
Influence of more frequent morphs (34)	*том/ат* (*томат*)	The frequency of morphs *-том-* and *-ат-* is greater than that of *-томат-*
Unseen or low-frequency morphs (under 15 entries) (28)	*спринтер* (*спринт/ер*)	The root *-спринт-* is not found in the training set
De-etymologization (16)	*о/град/и/ть/ся* (*оград/и/ть/ся*)	Historically, this word used to have the root *-град-*, but now it is *-оград-*
Roots are abbreviations (5)	*тюз/ов/ец* (*т/ю/з/ов/ец*)	The word is derived from *ТЮЗ*, which is an abbreviation, so each letter represents a separate root
Morphological alternation (3)	*лине/еч/н/ый* (*линееч/н/ый*)	The morph *-лин-* (*раз**лин**овать*) has allomorphs *-лине-* and *-лини-*, which confuses the model
Other (14)	*про/гулоч/н/ый* (*про/гул/оч/н/ый*)	The morphs *-гул-* and *-оч-* have high frequency, yet the model fails to segment them

For words with unseen roots:

High performance if affixes have high frequency:
- postfix *-ся*
- suffixes *-ть-, -вш-, -и-, -изм-, -ист-, -ова-*
- prefixes *рас-, за-*

Lower performance if affixes have low frequency:
- prefix *ре-*
- suffix *-инг*

6. CONCLUSION AND FURTHER WORK

- the existing CNN models with new parameter values are quite effective for an almost twice smaller amount of labeled training data
- the results are worse on a sample with 'unfamiliar' roots

Prospects for research:
- using new architectures of neural networks
- applying automatic morphemic analysis (as well as morpheme-based embeddings) to more general NLP problems such as various text classification tasks

References

1. Sorokin, A., Kravtsova, A.: Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language. In: Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science, vol. 930. Springer, Cham, pp 3-10 (2018)
2. Morpheme Segmentation for the Russian language. https://github.com/kpopov94/morpheme_seq2seq
3. The Dictionary of Neologisms. Neologisms of the century [Slovar' neologizmov. Neologizmy veka]. https://russkiiyazyk.ru/leksika/slovar-neologizmov.html
4. Dictionaries and encyclopedias. Orthographic dictionary by V. V. Lopatin [Slovari I entsiklopedii. Orfograficheskiy slovar' V. V. Lopatina]. https://gufo.me/dict/orthography_lopatin
5. Cognate words [Odnokorennye slova] https://wordroot.ru
6. Morpheme Segmentation for Russian: Evaluation of Convolutional Neural Network Model. https://yadi.sk/d/L3YrwGZAmW3Cug