# Reproducing Russian NER Baseline Quality without Additional Data

Valentin Malykh[1,2], Alexey Ozerin[2]

[1] Institute for Systems Analysis of Russian Academy of Sciences,
9, pr. 60-letiya Oktyabrya, Moscow, 117312, Russia
http://www.isa.ru/
[2] Laboratory of Neural Systems and Deep learning,
Moscow Institute of Physics and Technology (State University),
9, Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russia
http://www.mipt.ru/

**Abstract.** Baseline solutions for the named entity recognition task in Russian language were published a few years ago. These solutions rely heavily on the addition data, like databases, and different kinds of preprocessing. Here we demonstrate that it is possible to reproduce the quality of existing database-based solution by character-aware neural net trained on corpus itself only.

**Keywords:** named entity recognition, character awareness, neural nets, multitasking

## 1 Introduction

Named entity recognition is a well known task in natural language processing field. It is highly demanded in the industry and has a long history of academic research.

Current approaches are critically dependent on the size and quality of the knowledge-base used. The knowledge base should be kept up to date, which requires additional resources to be constantly involved.

In contrast our solution relies only on the text of the corpus itself without any additional data, except of the training corpus markup.

Contributions of the paper are following:

- We propose an architecture of artificial neural net as an alternative to the knowledge base based approach for the named entity recognition task.
- We provide results of the model tests on publicly available corpus for Russian language.

## 2 Related work

The first results for character-based named entity recognition in English language were presented in early 2000-s [1]. The close idea of character-based named entity tagging was introduced in [2] for the Portuguese and Spanish languages, but our model does not use convolution inside. For the English language text classification (close task for the named entity recognition) character-aware architecture was described in [3], it is also basing on convolutions, so principally differs from our model. Previous research for Russian language hadn't been based on characters, but on words [4]. State of the art solution on the public corpus with named entity markup [5] is also word-level based.

One of the core ideas for our model comes from the character aware neural nets introduced recently in [6], [7]. Another idea, that of matching the sequences to train the artificial neural net to get the text structure is coming from [8]. Our solution is based on the multi-task learning which was introduced for natural language processing tasks in [9].

## 3 Model

The architecture of our recurrent neural network is inspired by [7]. The network consists of long short-term memory units, which were initially proposed in [10]. There are two main differences to the Yoon Kim setup [7]. First one is that our model predicts two things instead of one:

- the next character,
- a markup label for the current character.

Second one is that we do not use convolution, so we not exploiting word concept inside our architecture, only character concept. We suppose that model could learn the concept of word from data, and rely on this assumption while quality measurement. Prediction errors and gradients are calculated, and then weights are updated by truncated back-propagation through time [11].

### 3.1 Mathematical formulation

Let $h_t$ be the state of the last neural net layer before softmax transformations (*hidden state*). The probability is predicted by standard sotfmax over the set of characters $\mathcal{C}$ and the set of markup labels $\mathcal{M}$:

$$Pr(c_{t+1}|c_{1:t}) = \frac{exp(h_t \cdot p_1^j + q_1^j)}{\sum_{j' \in \mathcal{C}} h_t \cdot p_1^{j'} + q_1^{j'}} \tag{1}$$

$$Pr(m_t|c_{1:t}) = \frac{exp(h_t \cdot p_2^i + q_2^i)}{\sum_{i' \in \mathcal{M}} h_t \cdot p_2^{i'} + q_2^{i'}} \tag{2}$$

Here $p_1^j$ is $j$-th column in character output embedding matrix $P_1 \in \mathbb{R}^{k \times |\mathcal{C}|}$, $q_1^j$ is a character bias term. $p_2^i$ is $i$-th column in markup output embedding matrix $P_2 \in \mathbb{R}^{l \times |\mathcal{M}|}$ and $q_2^i$ is markup bias term, $k$ and $l$ are character and markup embedding vector lengths.

The final negative log likelihood ($NLL$) is computed over the test corpus of length $T$:

$$NLL = -\sum_{t=1}^{T} (\log Pr(c_{t+1}|c_{1:t}) + \log Pr(m_t|c_{1:t})) \tag{3}$$

The diagram of our model could be found on the figure 1.

## 4 Experiments

The corpus parameters are presented at table 1, more details on it could be found in [5]. It can be obtained from the authors of the original paper by sending a request to `gareev-rm@yandex.ru` or to any other author of the original paper.

**Table 1.** Russian NER corpus statistics

| | |
|---|---|
| Tokens | 44326 |
| Words & Numbers | 35116 |
| Characters | 263968 |
| Organization annotations | 1317 |
| Org. ann. characters | 14172 |
| Person annotations | 486 |
| Per. ann. characters | 5978 |

Similar to [5] we calculate 5-fold cross-validation with precision (P), recall (R), and F-measure (F) metrics. The results of experiments are presented in table 2. Since we are working with characters we cannot use labelling produced for characters by our system directly, so we parse the produced markup for every token (which is known for us from the corpus) and take the label for the majority of characters in the token as a token label.

**Table 2.** 5-fold cross-validation of the NN-based NER.

| Fold # | Person | | | Organization | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| 1 | 93.09 | 93.32 | 93.20 | 68.75 | 78.57 | 73.33 | 63.25 | 71.94 | 67.32 |
| 2 | 94.85 | 94.16 | 94.51 | 64.29 | 73.90 | 68.76 | 59.38 | 67.86 | 63.33 |
| 3 | 90.91 | 93.37 | 92.12 | 66.22 | 65.52 | 65.87 | 58.45 | 58.76 | 58.60 |
| 4 | 90.45 | 91.74 | 91.09 | 68.02 | 77.48 | 72.45 | 60.12 | 68.56 | 64.06 |
| 5 | 94.03 | 93.06 | 93.54 | 62.15 | 68.81 | 65.31 | 57.06 | 61.40 | 59.15 |
| mean | 92.67 | 93.13 | 92.89 | 65.89 | 72.86 | 69.14 | 59.65 | 65.70 | 62.49 |
| std | 1.92 | 0.88 | 1.32 | 2.70 | 5.60 | 3.67 | 2.31 | 5.44 | 3.63 |

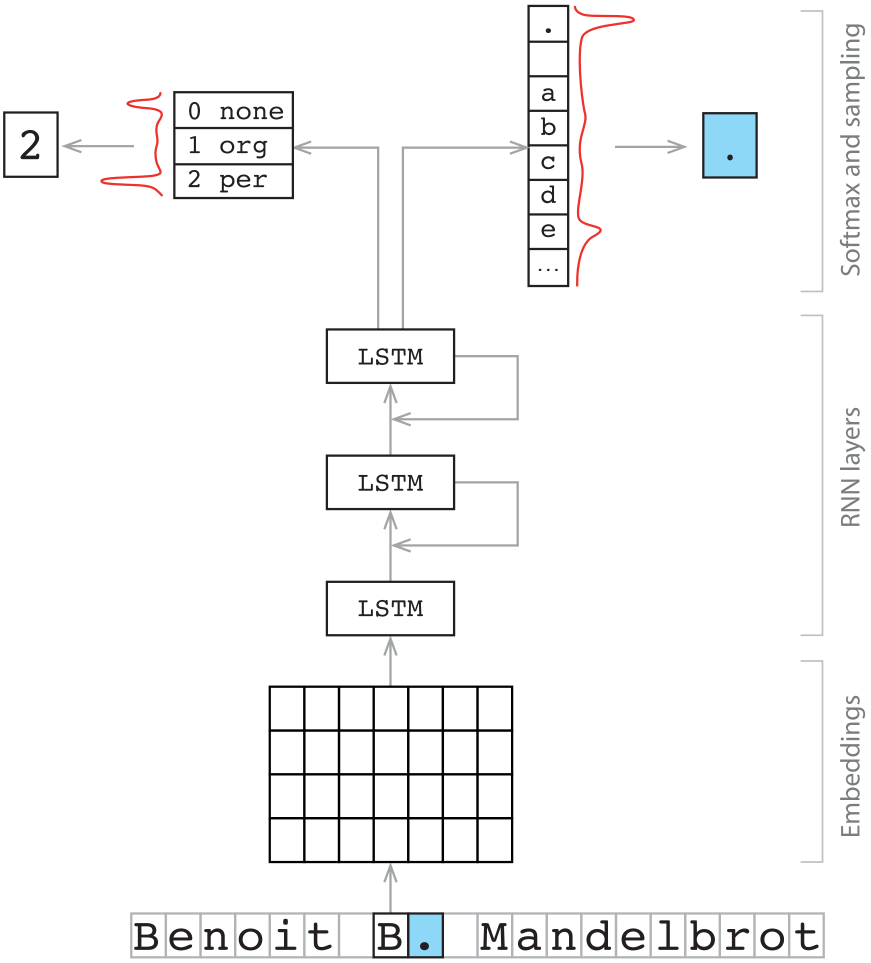

**Fig. 1.** Neural net architecture

## 5 Comparison

The results of comparison are presented on tables 3, 4, 5.

**Table 3.** Person class performance comparison.

| System | Person | | | | | |
|---|---|---|---|---|---|---|
| | Precision | | Recall | | F-measure | |
| | mean | std | mean | std | mean | std |
| Best KB-based [5] | 79.38 | N/A | 79.22 | N/A | 79.30 | N/A |
| CRF-based [5] | 90.94 | 4.04 | 79.52 | 2.91 | 84.84 | 3.33 |
| NN-based | **92.67** | 1.92 | **93.13** | 0.88 | **92.89** | 1.32 |

**Table 4.** Organization class performance comparison.

| System | Organization | | | | | |
|---|---|---|---|---|---|---|
| | Precision | | Recall | | F-measure | |
| | mean | std | mean | std | mean | std |
| Best KB-based [5] | 59.04 | N/A | 52.32 | N/A | 55.48 | N/A |
| CRF-based [5] | **81.31** | 7.44 | 63.88 | 6.54 | **71.31** | 5.38 |
| NN-based | 65.89 | 2.70 | **72.86** | 5.60 | 69.14 | 3.67 |

**Table 5.** Overall performance comparison.

| System | Overall | | | | | |
|---|---|---|---|---|---|---|
| | Precision | | Recall | | F-measure | |
| | mean | std | mean | std | mean | std |
| Best KB-based [5] | 65.01 | N/A | 59.57 | N/A | 62.17 | N/A |
| CRF-based [5] | **84.10** | 6.22 | **67.98** | 5.57 | **75.05** | 4.82 |
| NN-based | 59.65 | 2.31 | 65.70 | 5.44 | 62.49 | 3.63 |

On the person token class our system performed better than CRF-based one by all the metrics by the mean value and standard deviation. On the organisation class our system is better by recall and comparable by F-measure with CRF-model. In overall case our system was on par with knowledge-base approach performance in F-measure and in recall with CRF-model.

## 6 Conclusion

We applied character aware RNN model with LSTM units to the problem of the named entity recognition in Russian language. Even without any preprocessing and supplementary data from external knowledge-base the model was able to learn solution end-to-end from the corpus with markup. Results demonstrated by our approach are on the level of existing state of the art in the field.

The main weakness of proposed model is differentiation between person and organization tokens. This is due to the small size of the corpus. A possible solution is pre-training on a large corpus such as Wikipedia, without any markup, just to train internal distributed representation of a language model. We presume that such pre-training would allow RNN to beat CRF-model.

Another direction of our future work is addition of attention as it was demonstrated to improve performance on character-level sequence tasks [12].

## References

1. Klein, D., Smarr, J., Nguyen, H., Manning, C.D.: Named entity recognition with character-level models. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics (2003) 180–183
2. dos Santos, C., Guimaraes, V., Niterói, R., de Janeiro, R.: Boosting named entity recognition with neural character embeddings. In: Proceedings of NEWS 2015 The Fifth Named Entities Workshop. (2015) 25
3. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems. (2015) 649–657
4. Popov, B., Kirilov, A., Maynard, D., Manov, D.: Creation of reusable components and language resources for named entity recognition in russian. In: Conference on Language Resources and Evaluation. (2004)
5. Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., Ivanov, V.: Introducing baselines for russian named entity recognition. In: Computational Linguistics and Intelligent Text Processing. Springer (2013) 329–342
6. Bojanowski, P., Joulin, A., Mikolov, T.: Alternative structures for character-level rnns. arXiv preprint arXiv:1511.06303 (2015)
7. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. arXiv preprint arXiv:1508.06615 (2015)
8. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112
9. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning, ACM (2008) 160–167
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
11. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
12. Golub, D., He, X.: Character-level question answering with attention. arXiv preprint arXiv:1604.00727 (2016)