Russian Person Names Recognition Using the Hybrid Approach

Anna Glazkova

University of Tyumen, Tyumen, Russia
a.v.glazkova@utmn.ru

Abstract. Russian Person Name Recognition has been widely discussed in research papers devoted to models based on rules and machine learning. In this paper, the problem of Russian Person Name Recognition is tackled by hybrid using of rule-based models and neural networks trained on vector representations of words. The empirical results indicate that this approach shows results comparable to rules-based models and models trained in the syntactic and semantic text features. The advantage of the presented approach is the absence of the need for deep semantic-syntactic analysis of the text and connecting dictionaries, as well as the simplicity of the architecture of the used networks, which allows to limit the memory and runtime of the model.

Keywords: data extraction, hybrid approach, named entity recognition, natural language processing, neural networks.

1 Introduction

Named Entity Recognition (NER) is the task of detecting and classifying proper names within texts into predefined types, such as Person, Location and Organization names [1]. NER tools are actively used in different Natural Language Processing applications.

Russian is the official language of Russian Federation and several other post-Soviet countries. It has over 150 million native speakers in the world and Russian is the most geographically widespread language of Eurasia [2]. Russian is an inflectional language from the point of view of morphology. The syntax is characterized by a relatively free order of words and an active role of intonation means. The basis of writing is the Cyrillic alphabet.

Although Persons Names Recognition and NER for Russian is a quite widely studied problem, its solution is usually built on models based on templates or syntactic or semantic features extracted from the text [3,4]. The use of such models demonstrates high efficiency, but requires additional research related to the development of rules and templates and the search for effective features for model training. A number of studies for the Russian language are dedicated to the construction of neural network models for NER. The presented models [5,6,7] demonstrate high accuracy rates and show the efficiency of using neural

network technologies for solving this problem. These models, however, may be quite difficult in terms of memory and computation.

In this work, we made an attempt to solve the problem of Russian Person Names Recognition. We tried to combine a neural network and a rule-based approach. At the same time, in our work we tried to avoid using complicated templates and rules for extracting named entities and we decided to use a fairly simple network architecture that can be created and trained in a short time.

The article is structured as follows. In the introduction we announce the purpose of our work and have referred to related works. Further, in the section «Methods» we describe the methodology of our work: datasets and tools, modelling and features, defining the boundaries of personal names in the text. Finally, we compare our results obtained on a textual collection with the results of other researchers.

2 Methods

2.1 Data Collections and Libraries

To train our network, we used a set comprising of a random sample of manually-annotated Persons-1000 texts [8] which includes 1000 news texts and their corresponding xml-files containing initial forms of personal names. In addition, we needed to use word embeddings by RusVectōrēs project [9] based on Russian National Corpus and Wikipedia texts (300-dimension vector corresponding to some word). To lemmatize words, we used pymorphy2. Neural network models are built and trained using TensorFlow.

2.2 Text Preprocessing

The text preprocessing was performed in the following way. We divided the texts into sentences and then sentences into words. We did not set ourselves a separate task of dividing the text into sentences, so the breakdown was simply carried out by punctuation. If punctuation marks are part of a personal name (for example, a dot after the initial), then such punctuation marks are ignored and included in other sentences.

For each word we received the following features:

1. Word embeddings.
2. The serial number of the word in the sentence.
3. Indicator of whether the word begins with a capital letter.
4. Indicator of whether the word contains specific suffixes of surnames and patronymics.

The last two features are binary. In our work, we focused on those features that can be extracted without additional semantic and syntactic analysis of the text. Obtaining mentioned characteristics requires minimal effort to analyse sentences. The sample was divided into training, test and examination samples in the ratio of 70, 20 and 10%.

2.3 The Network Architecture

In the test we used feed forward network architecture. The main reason to choose this type of architecture is a quite large feature set and significant amount of training sample [10]. We focused on «budgeted» (in terms of memory) models and compensated possible loss of accuracy by applying the rules when defining the boundaries of personal names.

The used model has two hidden layers with sigmoid activation. Each hidden layer contains 200 neurons. For optimization we have chosen the Adam optimizer and exponentially decaying learning rate.

The choice of the best model was carried out as follows. We trained models using all input features with the number of hidden layers from 1 to 4 and the number of neurons on the hidden layers from 150 (half the dimension of the input data) to 300 (increment is 10) (Fig. 1).

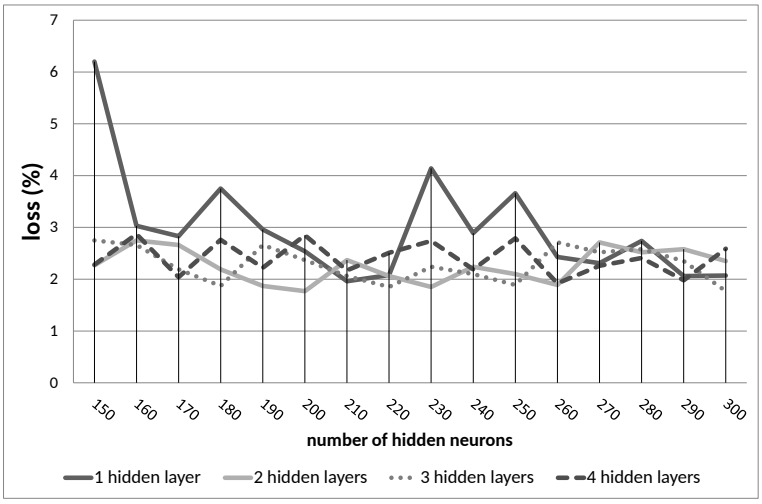

Fig. 1. The results of the models on the test sample, depending on the number and dimension of the hidden layers

The models were trained on a training sample. At each iteration, we calculated the loss on the test sample. The training was interrupted if at some iteration the results of the model on the test sample began to decrease. The choice of the optimal model was carried out in accordance with the results obtained on the test sample (Fig. 2). The examination sample was used to finalize the quality of the selected model.

2.4 Defining Boundaries of Personal Names

All words in the training sample have the index of 1, if they are part of a personal name, and if not, the index of 0. The aim of the training is the correct

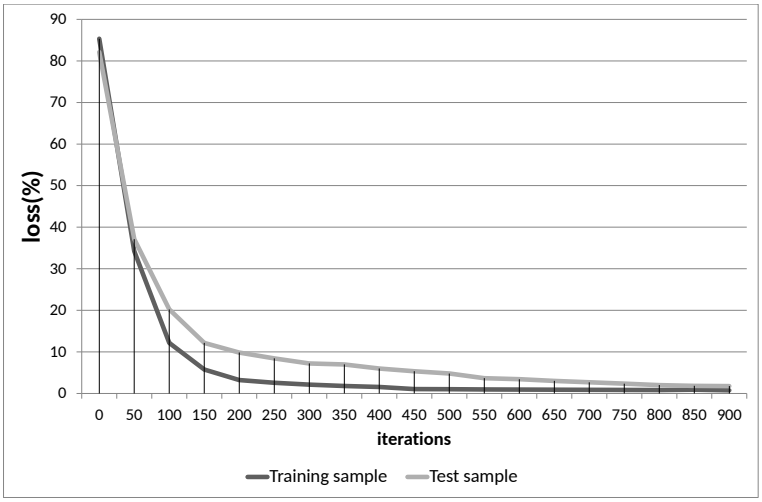

Fig. 2. The dynamics of loss reduction on the test and training samples

prediction of these indices for the elements of the examination sample. Therefore, the trained model associate each word with a number m from 0 to 1, where 1 is the marker for the word being part of a personal name:

$$f(x_i) \to m_i, m_i \in [0, 1],$$

i is the index of word x_i, $i \in [1, n]$, n – the sample size, $f(x_i)$ – the set of features for x_i.

The decision on whether a word is part of a personal name is taken on the basis of the value of m. If the value exceeds the threshold value k, the word can be considered a part of a personal name. In the experiments we use $k = 0.5$.

After processing by the neural network, each word of the text has a number m_i. At this moment these are separate words. Next, we must combine the words into personal names and define the boundaries of personal names. For these purposes, we used a fairly simple rule. First of all, we simply combined those words that look like fragments of personal names (which have $m_i > k$) and which stand side by side in the text and are not separated by punctuation marks. If the word is separate, we consider it a separate personal name.

Further, we check the neighborhood of personal names (words adjacent to personal names and not separated by punctuation marks). We decrease the threshold value k for the words included in the vicinity of personal names. If the value m_i for each word in the vicinity exceeds a new threshold k_v, we include the neighborhood in the personal name and change the boundaries of a name. Next, we estimate the value m_i taking into account the new boundaries. As a result of the experiments, we chose $k_v = 0.35$ and the neighborhood size equal to 1.

3 Experiments and Results

The table 1 contains quality indicators of neural network classification *for the examination sample*, which determines whether a word is part of a personal name. As a target metric, we use F-score:

$$Precision_n = \frac{TP_n}{TP_n+FP_n},$$

$$Recall_n = \frac{TP_n}{TP_n+FN_n},$$

$$F_nscore = 2 * \frac{Precision_n * Recall_n}{Precision_n + Recall_n},$$

TP_n – the number of true positive fragments of personal names, FP_n – the number of false positive fragments of personal names, FN_n – the number of false negative fragments of personal names.

Table 1. The quality of neural model

Features	F_nscore	$Precision_n$	$Recall_n$
Word embeddings	92.94%	92.52%	93.36%
All features	93.12%	92.73%	93.51%

In the table 2 we show the final results of person names recognition *for the examination sample* with $k_v = 0.35$. The F-score is calculated as follows:

$$Precision_p = \frac{TP_p}{TP_p+FP_p},$$

$$Recall_p = \frac{TP_p}{TP_p+FN_p},$$

$$F_pscore = 2 * \frac{Precision_p * Recall_p}{Precision_p + Recall_p},$$

TP_p – the number of true positive personal names, FP_p – the number of false positive personal names, FN_p – the number of false negative personal names.

Table 2. The final results (neighborhood size = 1)

	F_pscore	$Precision_p$	$Recall_p$
Results	**93.41%**	93.54%	93.28%

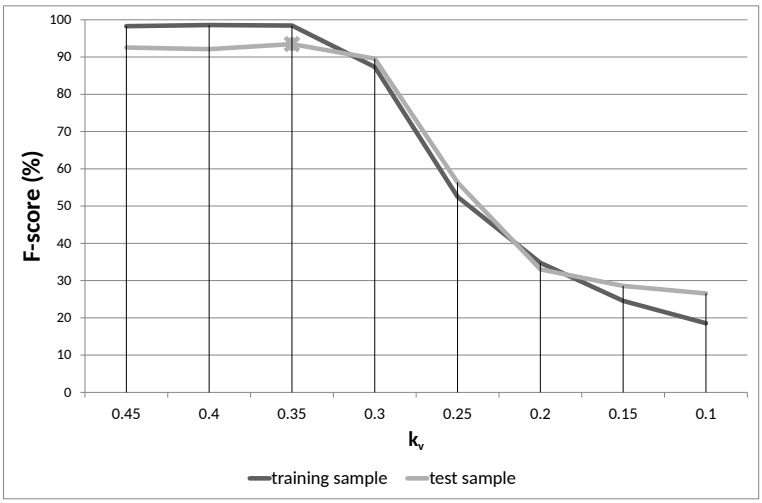

Fig. 3. The influence of the coefficient k_v on the results for the test and training samples. The best result for the training sample is marked with *

Fig. 3 presents the dependence of the results on the value of the coefficient k_v *for the training and test samples.*

In the table 3 we compare our results with the results of other approaches and methods implemented for NER tasks. Each of the studies mentioned below is distinguished by the originality of task and methodology, therefore the comparison can not be absolutely correct. However, we can estimate the overall picture of the results in this field of research.

Table 3. The comparison of results

Authors	F-score
Vlasova N. et al. [11]	91.53%
Blinov P. [12]	92.61%
Our approach	**93.41%**
Trofimov I. [13]	95.57%
Sysoev A. et al. [5]	96.24%
Mozharova V. et al. [3]	96.62%
Anh L. et al. [7]	99.26%

4 Conclusions

The paper presents a hybrid approach to Russian Person Names Recognition that combines the advantages of neural network and rules-based approaches. We

compared our results with the results obtained earlier. Our approach did not show the best result, but it is sufficient to achieve useful F-score.

The main advantage of our approach is its simplicity in terms of resource use and implementation, there is no need to connect dictionaries, create templates and feature set provided that we have ready-made word embeddings.

The results of the article will serve as a basis for further research on information extraction problems.

5 Acknowledgments

The authors would like to acknowledge the valuable comments and suggestions of the reviewers, which have improved the quality of this paper.

The reported study was funded by RFBR according to the research project 18-37-00272.

References

1. *Oudah, M., Shaalan, K.* Person Name Recognition Using the Hybrid Approach // International Conference on Application of Natural Language to Information Systems. 2013. P. 237–248.
2. *Russian Language.* URL: https://www.en.wikipedia.org/Russian_language. Date of access: 29.01.2018.
3. *Mozharova V., Loukachevitch N.* Two-stage approach in Russian named entity recognition. // Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference. 2016.
4. *Rubaylo A. V., Kosenko M. Y.* Software utilities for natural language information retrieval. // Almanac of modern science and education. Volume 12 (114), 2016. P. 87–92.
5. *Sysoev A. A., Andrianov I. A.* Named Entity Recognition in Russian: the Power of Wiki-Based Approach // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». 2016.
6. *Ivanitskiy R., Shipilo A., Kovriguina L.* Russian Named Entities Recognition and Classification Using Distributed Word and Phrase Representations // SIMBig. 2016. P. 150–156.
7. *Anh L. T., Arkhipov M. Y, Burtsev M. S.* Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition // Artificial Intelligence and Natural Language Conference (AINL 2017). 2017.
8. *Vlasova N.A., Sulejmanova E.A., Trofimov I.V.* Message about the Russian-language collection for the task of extracting personal names from texts / in «Proceedings of the conference on computer and cognitive linguistics TEL'2014 «Language semantics: models and technologies» P. 36–40. – Kazan, 2014.
9. *Kutuzov A., Kuzmenko E.* WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models / In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. P. 155–161, vol. 661. – Springer, Cham. 2017.

10. *Botha J.A., Pitler E. et al.* Natural Language Processing with Small Feed-Forward Networks / Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017.
11. *Vlasova N. A., Podobryaev A. V.* Automatic noun phrases extraction using preliminary segmentation and CRF with semantic features // Program Systems: Theory and Applications. Volume 4 (35), 2017. P. 21–30.
12. *Blinov P. D.* Automatic named entity recognition in the Russian text // Scientific and Technical Volga region Bulletin. Volume 3, 2013. P. 91–96.
13. *Trofimov I.V.* Person name recognition in news articles based on the persons-1000/1111-F collections / 16th All-Russian Scientific Conference Digital Libraries: Advanced Methods and Technologies, Digital Collection, RCDL 2014, pp. 217–221.