Combining Knowledge and CRF-based Approach to Named Entity Recognition in Russian

Mozharova V. A.[1] and Loukachevitch N. V.[2]

[1] Lomonosov Moscow State University, Moscow, Russia,
valerie.mozharova@gmail.com,
[2] Lomonosov Moscow State University, Moscow, Russia,
louk_nat@mail.ru

Abstract. Current machine-learning approaches to information extraction often include features based on large volumes of knowledge in form of gazetteers, word clusters, etc. In this paper we consider a CRF-based approach to Russian named entity recognition based on multiple lexicons. We test our system on the open Russian collection "Persons-1000" labeled with personal names. We additionally annotated this collection with names of organizations, media, locations, and geo-political entities and present the results of our experiments for one type of names (Persons) for comparison purposes, for three types (Persons, Organizations, and Locations), and five types of names. We also compare two types of labeling schemes for Russian: IO-scheme and BIO-scheme

Keywords: CRF, named entity recognition.

1 Introduction

Information extraction is one of the most important tasks in natural language processing. There are several basic types of information to extract. The first type is named entities, such as person names, company names, or locations. The second type is relationships between named entities, for example, a person post in an organization. The third type is events that occur with named entities, for example, company merging, stock purchasing, or business meetings. All this information is used in information retrieval tasks, document annotation tasks, business analytics, and many other areas.

Most papers devoted to named entity recognition [1][2] present studies for Enlish. For Russian, such experiments were carried out, mainly, on proprietary text collections, and the issues on comparison of approaches, the best feature sets still exist. In this paper we present our experiments on named entity extraction using the open Russian text collection "Persons-1000". In our approach we combine statistical and knowledge-based methods. Also we compare two labeling schemes for Russian named entity recognition: IO-scheme and BIO-scheme. We use the CRF method as a machine learning method for this task.

2 Named entity recognition task

A named entity is a word or a word collocation that means a specific object or an event and distinguishes it from other similar objects [1]. Named entities must have a referent and they are usually written with a capital letter, for example:

1. Президент Владимир Путин 17 декабря провел традиционную пресс-конференцию перед Новым годом.
2. Студенты и Татьяны получат эксклюзивный пропуск на Главный каток страны.

In the first sentence, the collocation "Владимир Путин" is a named entity because it means a specific person. In the second sentence, the word "Татьяны" is not a named entity because it does not have a specific referent.

There are many types of named entities, such as persons, organizations, locations, events, and time.

3 Related work

Machine learning methods, such as CRF, maximum-entropy, or SVM, are very popular in the named entity recognition task in many languages, including Slavic languages.

In [3] the authors carried out experiments in Czech with forty-two named entity types. To recognize named entities, they used a maximum-entropy based recognizer. Two-stage prediction was implemented: the second stage used the results of the first stage. To extract the features, the authors used the large number of gazetteers and corpus-based word clusters (Brown clusters [4]).

For the open Polish collections CZER, CEN, and CPR, the authors of [5] applied the CRF method for five-type named entity recognition (first names, surnames, countries, cities, roads). They used such features as orthographic features, wordnet-based features, morphological features, and gazetteer-based features.

There are several works applying CRF in the Russian named-entity recognition.

In [6] the authors presented the results of the CRF method on various tasks, including the named entity recognition. The experiments were carried out on their own Russian text corpus, which contained 71,000 sentences. They used only n-grams and orthographic features of tokens without utilizing any knowledge-based features. They achieved 89.89% of F-score on three named entity types: names (93.15%), geographical objects (92.7%), and organizations (83.83%).

In [7] the experiments were based on the open Russian text collection "Persons-600"[3] for the person name recognition task. The authors also chose the CRF method for recognition. Such features as token features, context features, and

[3] `http://ai-center.botik.ru/Airec/index.php?option=com_content&view=article&id=27:persons-600&catid=15&Itemid=40`

the features based on knowledge about persons (roles, professions, posts, and other) were utilized. They achieved 88.32% of F-score on person names.

In [8] the experiments were carried out on the Russian text collection, which contained 97 documents. The authors used two approaches for the named entity recognition: knowledge-based and CRF-based approach. In the machine learning framework they utilized such features as the token features and the knowledge features based on word clustering (LDA topics [10], Brown clusters [4], Clark clusters [11]). They achieved 75.05% of F-score on two named entity types: persons (84.84%) and organizations (71.31%).

To extract Russian personal names, in [9] the author used the knowledge-based approach without any machine learning method. This approach was based on regular expressions and gazetteers. The system was tested on the open collection "Persons-1000"[4]. Initially, the system achieved 81.36% of F-score, but after adding the global context feature, it achieved 96.62% of F-score on person names.

4 Text collection and labeling rules

To extract Russian entities, we experiments on the open Russian text collection "Persons-1000", which contains 1000 news documents with person labels. This collection was annotated by Research Center of Artificial Intellegence [12] in a similar way to MUC-7 labeling [13].

We additionally labeled this collection with other named entities:

- Organizations (ORG)
- Media organizations having a specific function of information providing (MEDIA)
- Locations (LOC)
- States and capitals in the role of a state (GEOPOLIT), for example, "Москва аннонсировала ..." ("Moscow announced that ...")

4.1 Labeling rules

Originally, only named entities, related to persons (PER), were labeled in the "Person-1000" text collection. According to the guidelines, only proper personal names were annotated. Roles and posts (for example, "Президент" ["The President"]), and persons, which names were not explicitly declared in the text (for example, "его отец" ["his father"]), were not labeled as named entities [12].

We additionaly labeled the collection with names of organizations, media organizations, locations, and geopolitical entities. We employed the following rules:

1. A descriptor is a word or a phrase indicating a generic type of a named entity. A descriptor is a part of a named entity:

[4] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

(a) If it is an abbreviation
   - [ОАО "Газпром"] ORG ([JSC "Gazprom"] ORG)

(b) If it is the head of a noun group, but it is not the supplement
   - [Санкт-Петербургский государственный университет] ORG ([Saint-Petersburg State University] ORG)
   - город [Тула] LOC (town [Tula] LOC)

2. A person name inside a proper name is not labeled separately
   - [Библиотека имени Ленина] ORG ([State Lenin Library] ORG)

3. A geographical object inside a named entity is labeled separately, if a named entity is not in quotes
   - [Университет при правительстве] ORG [РФ] GEOPOLIT ([University Under The Government of the] ORG [Russian Federation] GEOPOLIT)
   - гостиница ["Москва"] ORG (hotel ["Moscow"] ORG)

Our text collection labeling is similar to the markup standard accepted in MUC7 [13] and CoNLL [14]. Table 1 presents the quantitative characteristics of the labeled named entities in the collection "Persons-1000".

Table 1. The quantitative characteristics of the labeled named entities in the collection "Persons-1000"

PER	10623
ORG	7032
MEDIA	1509
LOC	3141
GEOPOLIT	4103

4.2 Labeling schemes

To represent labeled text segments as features for machine learning, several labeling shemes can be used. In our experiments we considered two schemes: IO-scheme and BIO-scheme.

IO-scheme (Inside-Outside)

In IO-scheme, every token can be labeled by only two types: "it belongs to named entity" (I), "it does not belong to named entity" (O). On the one hand, the classifier can learn easier when the number of label types is smaller, but, on the other hand, it is necessary to introduce additional rules for finding named entity boundaries. Table 2 presents an example of this labeling.

Table 2. IO-scheme example

Владимир	I-PER
Путин	I-PER
посетил	O
Англию	I-GEOPOLIT

Further, for this labeling scheme, the prefix c "I-" is omitted, and the label "O" is replaced by the label "NO". This scheme presupposes the prediction of $|C| + 1$ classes, where $|C|$ is the number of the named entitiy categories.

BIO-scheme (Begin-Inside-Outside)

In BIO-scheme, every text token can be associated with the label from one of three types: "named entity beginning" (B), "named entity continuation" (I), or "not named entity" (O). Thus, the classifier determines the boundaries of named entities by itself, and it should make named entity recognition much easier. Table 3 shows an example of this labeling scheme. The BIO-scheme requires to predict $2|C| + 1$ classes, where $|C|$ is the number of named entity categories.

Table 3. BIO-scheme example

Владимир	B-PER
Путин	I-PER
посетил	O
Англию	B-GEOPOLIT

In the well-known Standford named entity recognizer [15], the authors gave preference to the IO-scheme because, in English, named entities of the same type rarely locate beside each other in texts, therefore it is not necessary to use the complicated scheme of labeling.

We found that in Russian, there are a lot of examples of the same type named entities locating beside each other in texts. Table 4 shows the statistics of the co-occurrence of the same-type objects for three types of named entities. It can negatively influence on further named entity token aggregation because of the problem with named entity boundaries. To compare labeling shemes, we carry out the experiments using both schemes of labeling (see Section 6).

5 Features and rules

To extract Russian named entities, we utilize the CRF classifier as a machine learning method because it showed good results in many works devoted to the

Table 4. Statistics of the co-occurrence of the same type named entities

NE type	Statistics
PER	72
ORG	261
LOC	58

named entity recognition. We used CRF++[5], which is an open source implementation for labeling sequential data. This implementation is fast and easy to tune.

5.1 Preprocessing

Before the feature extraction, the text collection was processed with a morphological analyzer. As a result, for each token such features as a part of speech, gender, number, and case were extracted.

5.2 Features

The fixed set of features was computed for every token. We used token features, context features, and features based on lexicons. Below the basic features are listed.

Token features

1. Token initial form (lemma)
2. Number of symbols in a token
3. Letter case. If a token begins with a capital letter, and other letters are small then the value of this feature is "BigSmall". If all letters are capital then the value is "BigBig". If all letters are small then the value is "SmallSmall". In other cases the value is "Fence"
4. Token type. The value of this feature for lexemes is the part of speech, for punctuation marks the value is the type of punctuation
5. The presence of a vowel (a binary feature)
6. If a token ends a sentence (a binary feature)
7. If a token contains a known letter n-gram from a pre-defined set:
   (a) If the last letters match one of the typical last name ends (-енко, -швили, -ова, -ов, etc.)
   (b) If the first letters match one of the typical first name beginnings
   (c) If in a token there is a letter n-gram that usually appears in organization names (-ком-, -орг-, -деп-, etc.)

[5] `https://taku910.github.io/crfpp/`

Features based on lexicons

To improve the results of named entity recognition, we used vocabularies that store lists of useful objects. The object can be expresed a word, or a phrase.

For every token, our system determines if a token is a known word or a token is included in a known phrase. The phrase length was also taken into account. Table 5 presents basic vocabularies and their sizes. The overall size of all vocabularies is more than 335 thousand entities.

Table 5. Vocabulary sizes

Vocabulary	Size, objects	Clarification	Examples
Famous persons	31482	Famous people	Владимр Путин, Ангела Меркель
First names	2773	First names	Василий, Анна, Том
Surnames	66108	Surnames	Кузнецов, Грибоедов
Person roles	9935	Roles, posts	министр, китаевед
Verbs of informing	1729	Verbs that usually occur with persons	высказать, отпроситься, признаться
Companies	33380	Organization names	Сбербанк
Company types	6774	Oranization types	организация, авиафирма
Media	3909	Media	РИА Новости, Первый канал
Geography	8969	Geographical objects	Балтийское море, Владивосток
Geographical adjectives	1739	Geographical adjectives	финский, томский, югославский
Usual words	58432	Frequent Russian words (nouns, verbs, adjectives)	автомобиль, падать, желтый
Equipment	44094	Devices, equipment, tools	устройство, телефон

Features based on context

The values of the above listed features were also calculated for neighbor tokens in two-word window for every token to the left and to the right.

As a result, the same number of the features were computed for every token. Table 6 presents a feature set example.

Table 6. Features

Token	Lemma	Register	Token Type	Second Name	Geo	Label
В	В	Small	Auxiliary	False	False	NO
России	РОССИЯ	BigSmall	Noun	False	Geo1	GEOPOLIT
Алиев	АЛИЕВ	BigSmall	Noun	True	False	PER
третий	ТРЕТИЙ	Small	Numeral	False	False	NO
раз	РАЗ	Small	Auxiliary	False	False	NO

5.3 IO-labelling: aggregation of tokens into named entities

In result of the classifier work, each token obtains a specific tag. Tokens corresponding to the same named entity should be aggregated. In case of BIO-labeling, named entities are distinguished by label boundaries (Begin-Inside-Outside). To determine named entity boundaries in case of IO-scheme, the special rules were used:

1. A multiword named entity is constructed from the same type tokens located beside each other.
2. If a sequence of tokens with the same label type includes a punctuation mark, then the decision depends on its type:
   (a) If the punctuation mark is a quote, open bracket or dot that is not the sentence ending, then a named entity is not separated.
   (b) Otherwise, a named entity is separated
   (c) Punctuation marks are not included in named entities
3. If the template "<ORG> имени <PER>" is met, the word fragment that matched with this template is joined together into the same organization named entity.
4. If there are more than two words in a person named entity, and two first names with different grammatical cases are met in this named entity, then the named entity is separated. The boundary is the second name.
5. If a token sequence $(p_1, ..., p_n)$, where p_i is a token labeled as a person, contains a known first name p_j, then all tokens p_k $(k \neq j)$ are memorized as possible persons. In cases of missed person labels for p_k, these labels can be restored. For example, if in a text personal name "Анатолий Котляр" ("Anatoliy Kotlyar") was recognized then token "Котляр" ("Kotlyar") will be labeled with the person tag even if the CRF classifier missed it. This is an attempt to utilize the global context of the text (see discussion about global features in [16]).

6 Experiments

The experiments were fulfilled using two labeling schemes: IO-scheme and BIO-scheme. For the IO-scheme, two runs were fulfilled: with and without rules.

The $Fscore$ was used as a target metric. It was calculated as follows:

$$Precision = \frac{intersectionCount}{classifierCount}$$

$$Recall = \frac{intersectionCount}{expertCount}$$

$$Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

where $intersectionCount$ is the number of named entities labeled by both: the classifier and the expert; $classifierCount$ is the number of named entities labeled by only the classifier; $expertCount$ is the number of named entities labeled by only the expert.

To calculate the target metric, we used the 3:1 cross-validation technique. The collection was divided into four parts, and each part was iteratively utilized as a test part and others as train parts. The final value of the target metric was calculated as the average value of the intermediate results.

Table 7 presents the results of the experiments with three types of named entities for all sorts of text labeling are shown. Table 8 shows the results for five named entity types.

Table 7. Results for three types of NE

NE type	F-score, %		
	IO	IO + rules	BIO
PER	94.95	95.09	**96.08**
ORG	80.03	80.23	**83.84**
LOC	92.60	92.60	**94.57**
Average	89.67	89.67	**91.71**

Taking into account the results from tables, we can conclude that, during work with the IO-scheme, the rules of token aggregation positively influence on the target metric, but, on an average, the BIO-scheme gives more significant contribution especially for organization recognition. It means that for the Russian texts BIO-scheme plays an important role because named entities can locate beside each other in texts, and we need to separate them.

The text collection "Persons-1000" is an extension of the collection "Persons-600", used in [7]. In that work, the F-score achieved 88.32% on the person entity type. Our experiments showed the improvement of this metric to 96.08% on the extension of the collection. We consider that it happened because in the previous work the authors used the small number of vocabularies oriented only to personal names. In our approach, we utilize much more various types of knowledge.

Table 8. Results for five types of NE

NE type	F-score, %		
	IO	IO + rules	BIO
PER	94.81	95.01	**95.63**
ORG	75.90	76.16	**80.06**
MEDIA	87.95	87.95	**87.99**
LOC	84.53	84.53	**86.91**
GEOPOLIT	**94.65**	**94.65**	94.50
Average	88.37	88.37	**89.93**

The rule-based system described in [9] was specially tuned on the collection "Persons-1000" and achieved 96.62% of F-sore, which can be considered as the maximum for this collection, and we mostly reached this result.

The same rule-based system [9], applied to another collection "Persons-1111-F"[6], containing mainly Arabian and other eastern personal names, obtained 64.43 F-score. We applied our model trained on the "Person-1000" collection to the "Person-1111-F" and achieved 81.68% (Table 9). This demonstrates more robustness of our system.

Table 9. Comparison of the rule-based system [9] and our system on two collections

Collection	F-score, %	
	Rule-based system [9]	Our system
Persons-1000	**96.62**	95.63
Persons-1111-F	64.43	**81.68**

7 Conclusion

In this work we present our experiments devoted to the Russian named entity recognition task on the open text collection "Person-1000". We used knowledge-based approach together with the CRF classifier. The knowledge was expressed in gazetteers and rules. We described our results for three types (persons, organizations, and locations) and five types of names (persons, organizations, media, locations, and geopolitical objects). We compared our study with previous works on the same and similar collections.

[6] `http://ai-center.botik.ru/Airec/index.php/ru/collections/29-persons-1111-f`

References

1. *Nadeau, D., Sekine, S.* A survey of named entity recognition and classification // Lingvisticae Investigationes. 2007. V. 30, № 1, pp.3–26.
2. *Tkachenko, M., Simanovsky, A.* Named Entity Recognition: Exploring Features / «Empirical Methods in Natural Language Processin», Proceedings of the 11th Conference on Natural Lanuage Processing, KONVENS 2012. pp. 118-127. — Eigenverlag ÖGAI, 2012.
3. *Straková, J., Straka, M., Hajič, J.* A New State-Of-The-Art. Czech Named Entity Recognizer / «Text, Speech, and Dialogue» , Proceedings of the 16th International Conference, TSD 2013. pp. 68-75. — Springer Berlin Heidelberg, 2013.
4. *Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., Mercer, R. L.* Class-based n-gram models of natural language // Computational Linguistics. 1992. V. 18, № 4, pp. 467–479.
5. *Marcińczuk, M., Stanek, M, Piasecki, M., Musiał, A.* Rich Set of Features for Proper Name Recognition in Polish Texts // «Security and Intelligent Information Systems» , Proceedings of the International Joint Conferences, SIIS 2011. pp.332—344. — Springer Berlin Heidelberg. 2012.
6. *Antonova, A. Y., Soloviev, A. N.* Conditional random field models for the processing of Russian / «Computational Linguistics and Intellectual Technologies», Proceedings of the International Conference "Dialog 2013", pp. 27-44. — RGGU, 2013.
7. *Podobryaev, A. V.* Persons recognition using CRF model // Proceedings of 15th All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collection», RCDL-2013, pp. 255—258. — Demidov Yaroslavl State University, 2013.
8. *Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., Ivanov, V.* Introducing baselines for russian named entity recognition / «Computational Linguistics and Intelligent Text Processing», Proceedings of 14th International Conference, CICLing 2013, pp. 329–342. — Springer Berlin Heidelberg, 2013.
9. *Trofimov, I. V.* Person Name Recognition in News Articles Based on the Persons-1000/1111-F Collections / Proceedings of the 16th All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections», RCDL 2014, pp. 217-221. — 2014.s
10. *Chrupala, G.* Efficient induction of probabilistic word classes with LDA / Proceedings of the 5th International Joint Conference on Natural Language Processing, IJCNLP 2011, pp. 363–372. — Asian Federation of Natural Language Processing, 2011
11. *Clark, A.* Combining distributional and morphological information for part of speech induction / Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics, EACL 2003, v. 1, pp. 59–66. — ACL, 2003.
12. *Vlasova, N. A., Suleimanova, E. A., Trofimov, I. V.* The message about Russian collection for named entity recognition task / Proceedings of TEL'2014, pp. 36–40. — Fan, 2014.
13. *Chinchor, N., Robinson, P.* MUC-7 named entity task definition / In Proceedings of the 7th Conference on Message Understanding, p.29. — 1997.
14. *Tjon Kim Sang, Erik, F., Fien De Meulder.* Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition / Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, v. 4, pp. 142-147. — ACL, 2003.

15. *Finkel, J. R., Grenager, T., Manning, C.* Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling / Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, pp. 363-370. — ACL, 2005.
16. *Ratinov, L., Roth, D.* Design challenges and misconceptions in named entity recognition / Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL, pp. 147–155. — ACL, 2009.