Two-stage Approach in Russian Named Entity Recognition

Valerie Mozharova
Lomonosov Moscow State University
Moscow, Russia
valerie.mozharova@gmail.com

Natalia Loukachevitch
Lomonosov Moscow State University,
Bauman Moscow State Technical University
Moscow, Russia
louk_nat@mail.ru

Abstract—**In this article we consider a two-stage prediction approach for named entity recognition in Russian. In the first stage, named entities are extracted by a machine learning method. After that our system collects the statistics of token classes and transforms this statistics to a feature set, which is used for training a new classifier. We consider three types of the two-stage features: the previous history, the whole document statistics, and global statistics of the whole collection. We carry out our experiments on several text collections. We show that the utilizing of the two-stage prediction approach improves the quality of named entity recognition.**

I. Introduction

Named entity recognition is a necessary step for various applications of natural language processing and information retrieval. During natural language processing of news flows and social network messages, names of persons, organizations and geographical objects are usually extracted.

There are two main approaches for this task: rule-based approaches and machine-learning approaches, in which automatic systems use different feature sets, including token features, dictionaries, clusters, etc.

To improve the quality of named entity recognition, it can be important to consider the information about the frequency of a specific token to be classified as a named entity. It can help to recognize named entities met in indefinite contexts. For this goal, a two-stage prediction approach has been proposed: in the first stage, a system classifies tokens, after that the system collects frequency statistics, which is used in the next stage.

Authors of works [1], [2], [3] used two-stage prediction approaches for named entity recognition. But such studies were not carried out for Russian. In this work we consider several versions of two-stage prediction in Russian named entity recognition and explore the influence of the second stage of recognition on several collections.

II. Related work

Several works were devoted to the named entity recognition task for Russian. In [4], [5], [6] the authors created systems based on the Conditional Random Field (CRF) machine learning method. In [7] the author described system based on rules and dictionaries.

In [4] 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: persons (93.15%), geographical objects (92.7%), and organizations (83.83%).

In [6] the experiments were based on the open Russian text collection "Persons-600" [10] for the personal name recognition task. The authors also chose the CRF method for recognition. Such features as token features, context features, and 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 [5] 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. 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 [7] the author used the knowledge-based approach comprising regular expressions and gazetteers. The system was tested on the open collection "Persons-1000". Initially, the system achieved 96.62% of F-score on names of persons.

On the other hand, there exist some other approaches that tackle the named entity recognition problem as a two-stage analysis. However, all this approaches are not explored yet (until this paper) for Russian texts. In the next paragraphs, the most relevant two-stage approaches using for English texts are described.

In [1] the authors offer two versions of two-stage analysis and their combination: the extended prediction history, and the two-stage aggregation. They achieved 90.57% of F-score (influence of the two-stage prediction is 2.88%) on the CoNLL03 text collection [11] [12], where experts labeled four types of named entities (persons, locations, organization and others).

In [2], the authors consider only extended prediction history and they achieved 89.16% of F-score on text collection CoNLL03, where the two-stage feature gave 1.13%, and 79.23% on Czech Named Entity Corpus 1.0, where 42 types of named entities have been labeled.

In [13] the authors present a system for English, where tokens classified as named entities on the first stage are used in

ISBN 978-952-68397-6-9 (paperback), ISBN 978-952-68397-7-6 (PDF)

the second stage. The authors did not find any improvements in recognition with the two-stage approach and achieved 91.02% of F-score on the CoNLL03 text collection.

III. TEXT COLLECTIONS

For our experiments, we use three Russian publicly available text collections. The first collection "Persons-1000" [14] contains 1000 news documents labeled with names of persons.

We additionally annotated this collection with other types of named entities:

- Organizations (ORG)
- Media organizations (MEDIA)
- Locations (LOC)
- Geopolitical entities (GEOPOLIT) such as countries

We annotated the collection using guidelines similar to MUC-7 [15]: there is no inserted named entities, they do not cross and each token has only one class. To label entities, we used Brat annotation tool [16] [17], a web-based tool for collaborative text annotation (Fig. 1).

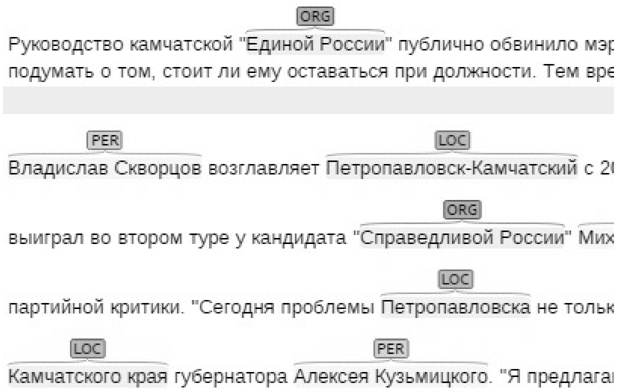

Fig. 1. Brat annotation tool interface

In the experiments presented in this paper, we consider only three types of named entities: persons, locations and organizations.

The second collection "Persons-1111" [18] comprises 1111 news documents containing Eastern names, such as Arabian, Indian, Chinese and Japanese names, which are usually complicated for automatic text analysis.

The third collection is the FactRuEval collection [19], created during the FactRuEval competition on Russian information extraction in 2016 [21]. The collection consists of Russian news and analytical documents concerning social and political issues. The corpus is subdivided into two parts: demonstration and test sets. There are 122 documents in the demonstration set, and the test set includes 133 documents. For our experiments, we use only the test set, because the development set was supposed to be too small to train our system. The collection is labeled with names of persons, locations and organizations by volunteers.

Table I presents information about the number of documents and entities in each experiment collection.

TABLE I. COLLECTION STATISTICS ON DOCUMENTS AND ENTITIES

Collection	Documents	Entities	Person	Location	Organization
Persons-1000	1000	26408	10623	7244	8541
Persons-1111	1111	5683	5683	—	—
FactRuEval:dev	122	2611	741	1082	787
FactRuEval:test	133	5019	1388	1574	2034

IV. METHOD AND FEATURES

Named entity recognition is a task of token classification into specified categories of names. We use CRF-based classifier, which specially developed for classification of sequential data. Conditional Random Fields ([20]) can be considered as a graph, which contains vertexes connected by undirected edges with transition probability. Each vertex corresponds to latent or observed variable. Unlike Hidden Markov Models (Fig. 2), in CRF (Fig. 3) each latent variable conditionally depends not only on the previous latent variable, but depends on other variables in graph including observed variables. In case of the named entity recognition task, text tokens are observed variables, and named entity categories are latent variables.

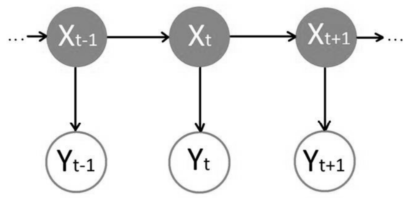

Fig. 2. Hidden Markov Models: x_i — latent variable, y_i — observed variable; a latent variable depends on previous latent variable

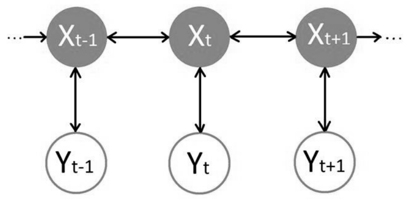

Fig. 3. Conditional Random Fields: x_i — latent variable, y_i — observed variable; a latent variable depends on latent variables and observed variebles from both sides

The fixed set of features was computed for every token. The basic set of features can be subdivided into token features and lexicon-based features. All features are calculated for each token and its context in radius of two words. Below the basic features are listed.

A. Token features

The set of token features includes:

- Token initial form (lemma)

- Number of symbols in a token
- 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”
- Token type. The value of this feature for lexemes is the part of speech, for punctuation marks the value is the type of punctuation
- The presence of a vowel (a binary feature)
- If a token ends a sentence (a binary feature)
- If a token contains a known letter n-gram from a pre-defined set: first name beginnings, last name endings (-ov, -ev, -idze, -enko, etc.), organization n-grams (-com-, -org-, -dep-, etc.)

B. Features based on lexicons

To improve the results of named entity recognition, we used vocabularies storing lists of useful objects. An object can be expressed with 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. For every lexicon, this feature gets next values:

- 0 – if a word was not found in this lexicon,
- {MaxPhraseLength} – if a word was found in this lexicon as a one-word text entry or as a component of a phrase text entry.

If variants of matching a word sequence with a lexicon exist, the system chooses a phrase with maximum length. For example, in the phrase ”role of Barak Obama”, words ”role” and ”of” will have the FamousPersonsLexicon value equal to 0. For words ”Barak” and ”Obama”, this value will be equal to 2.

Table II presents basic vocabularies and their sizes. The overall size of all vocabularies is more than 335 thousand entities. The lexicons were extracted from several sources: phonebooks, Russian Wikipedia, RuThes thesaurus [24], [25], etc.

TABLE II. VOCABULARY SIZES

Vocabulary	Size, objects
Famous persons	31482
First names	2773
Surnames	66108
Person roles and posts	9935
Verbs of informing (ex.: talk, say, ask, etc.)	1729
Companies	33380
Company types	6774
Media	3909
Geography	8969
Geographical adjectives	1739
Frequent Russian words (nouns, verbs, adjectives)	58432
Equipment, devices	44094

V. TWO-STAGE APPROACH

We assume that for better classification, it is useful to utilize previous experience of a classifier and to memorize the statistics of labels for future use.

Thus, on the first stage names are extracted by a trained machine learning method. Further, the system collects the statistics of obtained token labels, and this statistics is transformed into an additional feature set, which is used for training a new classifier together with the baseline set of features. In this work we explore the following types of two-stage features: the previous history, the document statistics, the global statistics.

A. Previous history

We assume that in the beginning of a text, a personal name is usually mentioned in its full form and a classifier can easier recognize it as a named entity. For example, the family name in the context of the first name and patronymic is easier to reveal than if it stands separately. For the text, in which the phrase ”Gennadiy Stolyar” is met before the single word ”Stolyar” mentioning, the second entry of the name is easier to recognize as a named entity because it has been classified as a person before.

For each token from the text, a system finds all its previous entries in the text and counts the statistics of classes that were generated by the classifier. Based on this statistics, the system creates additional features for each class, which can have one of three values: *no_one* (a token was not classified with this class before), *best* (this token was classified with this named entity type more than in 50% cases), *rare* (a token was rarely classified with this class).

For example, if the token ”Russia” was in the text before five times and the classifier determined it as an organization two times and as a location three times, then the previous history for the sixth token ”Russia” will be: PER – *no_one*, ORG – *rare*, LOC – *best*.

B. Utilizing document statistics

This method is very similar to the previous one, and it differs from the previous history only that the system analyzes occurrences of a token in the whole document.

C. Utilizing global statistics

This method differs from the previous ones that the system analyzes not only one document but the whole collection statistics.

VI. EXPERIMENTS

The two-stage prediction approach in named entity recognition was tested on several text collections. We used CRF++ [22] open source realization.

For token labeling, we use BIO-scheme, in which tokens are annotated with three types of labels for each named entity category: named entity beginning, named entity continuation and not a named entity [23]. For example, in the sentence "Vladimir Putin congratulated Russians" , the token "Vladimir" gets the label "Begin-Per" (named entity beginning), the token "Putin" obtains "In-Per" (named entity continuation) and other tokens have "Out" label (not named entity). This scheme helps to determine the named entity borders automatically.

As a target metric, we use F-score, the combination of precision and recall metrics. The results of two-stage prediction approaches are compared with the results of the baseline system based on token features and lexicons.

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

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

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

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.

A. Experiments on "Persons-1000" and "Persons-1111" collections

To test our system, we use cross-validation on the collection "Persons-1000" with relation of train and test parts as 3:1 (Table III). On this collection, the previous history feature shows the best result for person name recognition, the global statistics feature has the biggest influence on recognition of organizations. The whole feature set gives the best results for named entity recognition in whole.

TABLE III. "PERSONS-1000" RESULTS

System	PER	LOC	ORG	micro
1) Baseline	96.61	94.94	85.19	92.49
2) (1) + The previous history	97.00	94.53	85.32	92.61
3) (1) + The document statistics	96.78	94.51	85.42	92.56
4) (1) + The global statistics	96.69	94.79	**85.67**	92.77
(1) + (2) + (3) + (4)	**97.21**	**95.21**	85.60	**92.92**
Rule-based system [7]	96.62	–	–	–

To test system portability to the second collection "Persons-1111" , we do not use cross-validation: we train our system on the "Persons-1000" and then test the model on the second collection (Table IV).

It can be seen, that on the "Persons-1111" collection, the previous history feature shows the best result. We can explain a small influence of the global statistics feature by diversity of this collection: it was artificially composed of texts with names from different cultures. But for both collections, the combination of two-stage approaches improves the baseline system.

TABLE IV. "PERSONS-1111" RESULTS

System	Persons
1) Baseline	86.71
2) (1) + The previous history	**88.87**
3) (1) + The document statistics	86.78
4) (1) + The global statistics	86.72
(1) + (2) + (3) + (4)	87.94
Rule-based system [7]	64.43

We can compare our results with the results of the rule-based system presented in [7]. In this work the author describes the system based on rules and dictionaries for person name recognition. This system achieved 96.62% of F-score on the "Persons-1000" collection and 64.43% on the "Persons-1111" . Our system showed better results on both collections. We see large difference of the results on the second collection. We can conclude that our system have demonstrated more portability than the rule-based system in this transfer.

B. Experiments on FactRuEval test collection

The third open collection we used is the FactRuEval test collection. This collection has two levels of markups for the named entity recognition task. The first level contains spans (Fig. 4) — continuous chains of words labeled with one or more special tags. For example, in case of personal names, spans can be subdivided into first names, surnames, patronymics, and nicknames.

The second level consist of object mentions — groups of spans.The types of object mentions correspond to the types of entities involved. There are four types of named entities in this markup: person, location, organization and locOrg. LocOrg is a geography object in the organization context. For example, in the sentence "The Goverment of the Russian Fegderation" the phrase "Russian Federation" has locOrg type. Using a special mode, this type can be transformed into location. So, we use only three types of object mentions.

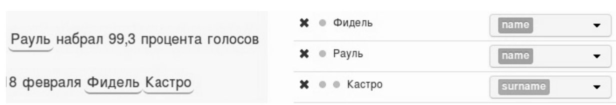

Fig. 4. Spans labeling in FactRuEval collection

The problem we faced is the difference in the labeling of organization names between the FactRuEval and the "Persons-1000" collections. In the FactRuEval collection, a whole organization mention is labeled as one named entity. In the "Persons-1000" collection, each department of an organization is labeled separately. For example, the phrase "Faculty of Journalism of Moscow University" in the first collection would be labeled as one organization named entity, but in the second collection it would be divided on two organization named entities: ”Faculty of Journalism” and ”Moscow University” . So, we adapted our labeling to the FactRuEval collection.

We trained our model on the ”Persons-1000” collection and then applied it to the FactRuEval test collection. We tested the following models: all three features, previous history, global statistics, previous history + global statistics, and the baseline model. The results are shown in Table V. Also, we compared our results with the results of FactRuEval participants and our system achieved the second result (Table VI).

We can see that in recognition of organizations, we achieved smaller F-score than in recognition of other types of named entities. There are two main types of system mistakes: incorrect determination of organization borders (the system tries to divide an organization name into two and more organizations) and wrong labeling words, which do not belong to any type of named entity, but begin with a capital letter (for exampe, artifacts).

TABLE V. FACTRUEVAL RESULTS

System	PER	LOC	ORG	F-score
1) Baseline	91.28	92.71	74.71	85.40
2) (1) + The previous history	92.13	92.99	74.68	85.79
3) (1) + The global statistics	92.08	93.09	74.68	85.80
4) (2) + (3)	**92.92**	92.97	74.97	86.16
5) (2) + (3) + The document statistics	92.91	**93.24**	**75.00**	**86.23**

TABLE VI. FACTRUEVAL PARTICIPANT’S BEST RESULTS

System	Person	Location	Organization	F-score
Violet	93.00	89.71	78.58	86.72
Our system	92.91	93.24	75.00	**86.23**
Pink	91.32	89.25	77.64	85.75
Beige	90.16	90.87	76.97	85.58
Crimson	92.08	88.13	75.63	84.93
Aquamarine	91.27	87.72	76.36	84.75
Black	91.75	90.44	73.61	84.60
Orange	89.95	86.99	71.52	82.43
Purple	88.88	87.95	71.50	82.31
Green	88.29	89.10	65.57	80.82
Brown	89.61	80.24	63.96	77.71
White	83.33	78.68	51.04	71.47
Ruby	81.10	76.13	38.69	66.51
Grey	91.22	00.00	00.00	43.56

C. Transfer two-stage model to large text collection

In the last experiment, we tried to assess the contribution of the global statistics feature calculated from a larger collection to the two-stage approach. Also we wanted to check the possibility of the two-stage model to be transferred to another collection. With this aim, we took 10,000 news texts (News collection) relating to another time interval than ”Persons-1000” . We applied the first-stage model to this News collection, calculated the global statistics feature on this collection, and then applied the second-stage model.

To evaluate a possible improvement of the two-stage model over the one-stage model, we randomly extracted texts from the News collection, where the two-stage model and the one-stage model had differences in revealed named entities. Then we analyzed the differences and counted improvements or errors of the two-stage model compared to the one-stage model. Then these counts were divided by the number of named entities met in these texts. In such a way, we could assess the change of precision and recall metrics between the models.

For example, if in sentence ”Professor Sergey Kuznetsov met friend” the baseline system labels ”Professor Sergey” as a person, and a two-stage system labels ”Sergey Kuznetsov” as a person, when we consider that the second system improves precision and recall with one named entity.

Tables VII and VIII present differ between basic and two-stage system results in objects. The **Better** label is the number of cases in that the two-stage prediction was better than the one-stage system, the **Worse** label means that the one-stage system was better. The Table IX shows the growth of the quality characteristics of the two-stage model. It can be also seen that, the two-stage model utilizing only the global statistics feature was better than the two-stage model using all three features.The results show us that utilizing only global statistics feature improves precision on this text collection.

TABLE VII. TWO-STAGE PREDICTION IMPROVEMENTS (TOTAL COUNT OF NAMED ENTITIES: 3122)

Better: 177	PER	LOC	ORG	Total
Precision	50	8	54	112
Recall	24	24	17	65
Worse: 81	PER	LOC	ORG	Total
Precision	14	8	25	47
Recall	14	3	17	34

TABLE VIII. GLOBLAL STATISTICS IMPROVEMENTS (TOTAL COUNT OF NAMED ENTITIES: 3122)

Better: 171	PER	LOC	ORG	Total
Precision	41	12	54	107
Recall	29	17	18	64
Worse: 61	PER	LOC	ORG	Total
Precision	9	1	11	21
Recall	29	2	11	42

TABLE IX. ”NEWS COLLECTION” RESULTS

System	Precision	Recall	Total
Baseline +The global statistics	+2%	+1.4%	+3.4%
Baseline + 3 two-stage features	+2%	+1%	+3%

So, we can make the conclusion that an influence of each feature depends on a text collection, but all experiments showed that the applying of two-stage approach improves the quality of named entity recognition on all collections.

VII. Conclusion

In this article we explored the two-stage prediction approach in Russian named entity recognition. Three features of the two-stage approach were considered and tested: the previous history, the document statistics and the global statistics obtained on the whole text collection. We applied the two-stage approach to several text collections and showed that the combination of two-stage approach features improves the quality of named entity recognition.

Recently, neural-network approaches demonstrated the state-of-the-art results in the named entity recognition task [26]. We plan to study those approaches in combination with our features.

Acknowledgment

This work is partially supported by RFBR grants No. 15-07-09306, 16-29-09606.

References

[1] L. Ratinov, D. Roth, "Design challenges and misconceptions in named entity recognition" , *in Proc. CoNLL Conf.*, 2009, pp. 147-155.

[2] J. Straková, M. Straka, J. Hajiĉ, "A New State-Of-The-Art. Czech Named Entity Recognizer" , *in Proc. TSD Conf.*, 2013, pp. 68-75.

[3] V. Krishnan, C. D. Manning, "An effective two-stage model for exploiting non-local dependencies in named entity recognition" *In Proc: ACL Conf.*, 2006.

[4] A. Y. Antonova, A. N. Soloviev, "Conditional random field models for the processing of Russian" , *in Proc. Dialog Conf.*, 2013, pp. 27-44.

[5] R. Gareev, M. Tkachenko, V. Solovyev, A. Simanovsky, V. Ivanov, "Introducing baselines for Russian named entity recognition" , *in Proc. CICLing Conf.*, 2013, pp. 329–342.

[6] A.V. Podobryaev, "Persons recognition using CRF model" , *in Proc. RCDL Conf.*, 2013, pp. 255–258.

[7] I. V. Trofimov, "Person Name Recognition in News Articles Based on the Persons-1000/1111-F Collections" , *in Proc. RCDL Conf.*, 2014, pp. 217-221.

[8] Marcińczuk, M., Stanek, M, Piasecki, M., A. Musiał, "Rich Set of Features for Proper Name Recognition in Polish Texts" , emphin Proc. SIIS Conf., 2012, pp.332-344.

[9] D. Nadeau, S. Sekine, "A survey of named entity recognition and classification" , *Lingvisticae Investigationes*. V. 30, 2007, pp. 3-26.

[10] Official site of Artificial Research Center, collection "Persons-600" , Web: http://ai-center.botik.ru/Airec/index.php/ru/collections/27-persons-600

[11] The shared task of CoNLL-2003, Web: http://www.cnts.ua.ac.be/conll2003/ner/

[12] Tjon Kim Sang, F. Erik, Fien De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition" , *in Proc.: 7th HLT-NAACL Conf.*, 2003, v. 4, pp. 142-147.

[13] M. Tkachenko, A. Simanovsky, "Named Entity Recognition: Exploring Features" , *in Proc. KONVENS Conf.*, 2012, pp. 118-127.

[14] Official site of Artificial Research Center, collection "Persons-1000" , Web: http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

[15] N. Chinchor, "MUC-7 Named Entity Task Definition" , *In Proc. MUC Conf.*, 2007.

[16] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, "BRAT: a web-based tool for NLP-assisted text annotation" , *in Proc.: EACL Conf.*, 2013.

[17] Brat annotation tool, Web: http://brat.nlplab.org/

[18] Official site of Artificial Research Center, collection "Persons-1111" , Web: http://ai-center.botik.ru/Airec/index.php/ru/collections/29-persons-1111-f

[19] Dialog Evaluation, FactRuEvalTextCollection, Web: https://github.com/dialogue-evaluation/factRuEval-2016

[20] J. Lafferty, A. McCallum, F. C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" , *in Proc.: ICML Conf.*, 2011.

[21] V. V. Bocharov, A. S. Starostin, S. V. Alexeeva, A. A. Bodrova, A. S. Chunchunkov, S. S. Dzhumaev, I. V. Efimenko, D. V. Granovsky, V. F. Khoroshevsky, I. V. Krylova, M. A. Nikolaeva, I. M. Smurov, S. Y. Toldova, "FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian" , *In Proc. Dialog Conf.*, 2016.

[22] CRF++, Web: https://taku910.github.io/crfpp/

[23] V. Mozharova, N. Loukachevitch, "Combining Knowledge and CRF-based Approach to Named Entity Recognition in Russian", *In Proc. the 5th International Conference on Analysis of Images, Social Networks, and Texts, AIST2016*, 2016.

[24] Ruthes Thesaurus, Web: http://www.labinform.ru/pub/ruthes/

[25] N. Loukachevitch, B. Dobrov, "RuThes linguistic ontology vs. Russian wordnets" . *in Proc. GWC Conf.*, 2014.

[26] G. Lample, M.Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, "Neural Architectures for Named Entity Recognition" , *In Proc. NAACL Conf.*, 2016.