Recognizing Names in Islam-Related Russian Twitter

© V. Mozharova © N. Loukachevitch

Lomonosov Moscow State University,
Moscow, Russia

valerie.mozharova@gmail.com louk_nat@mail.ru

Abstract. The paper describes an approach to creating a domain-specific tweet collection written by users frequently discussing Islam-related issues in Russian. We use this collection to study specific features of named entity recognition on Twitter. We found that in contrast to tweets collected randomly, our tweet collection contains relatively small number of spelling errors or strange word shortenings. Specific difficulties of our collection for named entity recognition include a large number of Arabic and other Eastern names and frequent use of ALL-CAPS spelling for emphasizing main words in messages. We studied the transfer of NER model trained on a news wire collection to the created tweet collection and approaches to decrease the degradation of the model because of the transfer. We found that for our specialized text collection, the most improvement was based on normalizing of word capitalization. Two-stage approaches to named entity recognition and Word2vec-based clustering were also useful for our task.

Keywords: NER, CRF, Twitter.

1 Introduction

Named entity recognition (NER) is one of the basic natural language processing tasks [17, 20]. Recognition of named entities in texts is used in many other information-processing tasks as relation extraction, entity linking, information retrieval etc. Most studies of NER have been carried out on news collections and shown high quality of named entity extraction. However, the transfer of NER recognizers to other genres of texts demonstrated significant decrease in the performance.

Currently, there is a great interest in information extraction from texts published on social media platforms such as Twitter or Facebook because these platforms can serve as a very useful (fast and/or alternative) source of information [22]. But application of general NER recognizers designed for or trained on news collections can demonstrate the decrease in performance of up to 50% on more in these informal texts [4, 7–9].

Another important direction of social network studies is directed to differences of language and style in specific social media communities [11, 18] or their dependence on social and demographic characteristics of users [12, 21].

In this paper, we consider the transfer of Russian NER recognizer trained on news texts to extracting names from Twitter messages. Our tweet collection is specialized; it is gathered from messages of those users who discuss issues related to Islam in their posts in contrast to other studies where Twitter collections are formed with random sampling of Twitter messages. This allows us to reveal specific features of the tweet language of Islam-oriented and other similar communities. We consider the transfer of CRF-based NER recognizer from a news data to the tweet collection and approaches to decrease the degradation of the model because of the transfer.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

2 Related works

2.1 Named Entity Recognition for Twitter

It is known that extraction of names from Twitter messages is much more difficult task than from other genres of text because of their shortness and informal character.

In [7] the authors review the problems and approaches to named entity recognition and entity linking for tweets. They write that the tweet content is noisy because of incorrect spelling, irregular capitalization, and unusual abbreviations. In their experiments, the main sources of mistakes in named entity recognition in tweets were violations in capitalization especially a large number of names written in lower case. They studied automatic normalization of tweets including spelling and capitalization correction and reported that in their investigation the normalization slightly improved the performance in NER for tweets.

In [24] the authors write that due to unreliable capitalization in tweets, common nouns are often misclassified as proper nouns, and vice versa. Some tweets contain all lowercase words (8%), whereas others are in ALL CAPS (0.6%). In addition to differences in vocabulary, the grammar of tweets differs from news text, for example, tweets often start with a verb. In their experiments, the supervised approach was used to predict correct capitalization of words. The set of features included: the fraction of words in the tweet which are capitalized, the fraction which appear in a dictionary of frequently lowercase/capitalized words but are not

lowercase/capitalized in the tweet, the number of times the word 'I' appears lowercase and whether or not the first word in the tweet is capitalized.

To study NER on Twitter performed with several NER systems, [8] use crowdsourcing to annotate tweets for the NER task. They annotate all @user-names as PER (person name). Annotating tweets for their experiments [24] choose not to annotate @usernames mentioned in tweets as entities because it is trivial to identify them using a simple regular expression, and they would only serve to inate the performance statistics.

In [4] the authors study the transfer of their NER model from news texts to tweets. They create a training set consisting of 1000 tweets. They use a baseline NER model based on token and context features (wordform, lemma, capitalization, prefixes and suffixes) and enhance it with two unsupervised representations (Brown clusters and vector representations) based on a large collection of unannotated tweets. Besides, they propose a technique to combine a relatively small Twitter training set and larger newswire training data. They report that two unsupervised representations work together better than alone, and the combination of training sets further improves the performance of their NER system.

2.2 Named Entity Recognition in Russian

In Russian there is a long tradition of engineering approaches to the named entity recognition task [13, 14, 23].

Machine-learning approaches for Russian NER usually employ the CRF machine learning method. In [1] the authors presented the results of the CRF-based 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 [19] the experiments utilized the open Russian text collection "Persons-600"[1] for the person name recognition task. The CRF-based classifier employed such features as token features, context features, and the features based on knowledge about persons (roles, professions, posts, and other). They achieved 88.32% of F-score on person names.

In [10] 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 [17], Brown clusters [3], Clark clusters [6]). They achieved 75.05% of F-score on two named entity types: persons (84.84%) and organizations (71.31%).

In 2016 the FactRuEval competition for the Russian language was organized. The FactRuEvaltasks included recognition of names in Russian news texts, recognition of specific attributes of names (family name, first name, etc), and extraction of several types of facts [2].

So far, named entity recognition in tweets did not have studied for Russian. Also, the dependence of NER performance on the language of specific Twitter user communities has not been studied before.

3 Text collections

3.1 News Text Collection

We study the transfer of CRF-based NER classifier trained on newswire data to the tweet collection. For training our system, we chose open Russian text collection "Persons-1000", which contains 1000 news documents labeled with three types of named entities: persons, organizations and locations[2]. The labeling rules are detailed in [16]. The counts of each named entity type in the collection are listed in Table 1.

Table 1 The quantitative characteristics of the labeled named entities in text collections

Type	News collection	Twitter collection
PER	10623	1546
ORG	8541	1144
LOC	7244	2836
OVERALL	26408	5526

3.2 Tweet Text Collection

We are interested in study of the language of Islam-related Twitter users in Russian. To extract tweets from users discussing Islam-related issues, we created a list of 2700 Islam terms. Then we extracted Russian tweets mentioning these terms using Search Twitter API, got users' accounts containing extracted tweets and ordered the accounts in the decreased number of extracted tweets from these accounts. We found that a lot of words from our list practically are not mentioned in tweets, other words (for example, "mosque" or "Muslim") are often used by very different people, not only Muslims.

After studying tweets from extracted accounts we created a very small list of the main Islam words ("Allah", "Quran", "Prophet", in various forms of Russian morphology). We also added the names of several known Islamist organizations to find their possible non-Muslim proponents. Then we repeated the whole procedure of tweet and account extraction, and found that the extracted collections can be considered as an appropriate approximation of messages generated by Islam-related users.

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

[2]http://labinform.ru/pub/named_entities/descr_ne.htm

We selected 100 users with the largest number of the extracted tweets, downloaded all their tweets and obtained tweet collection consisting of 300 thousand tweets (further FullTweetCollection). Then we randomly extracted tweets from different users, removed non-Russian or senseless tweets and at last obtained the tweet collection of 4192 tweets (further TestTweetCollection). The created collection contains messages with Quran quotes, religious and political argumentation, news-related messages mainly about Near and Middle East events (Syria, Iraq, Afghanistan etc) and Islamist organizations (Syrian opposition groups, ISIL, etc.) and also other types of messages (for example, advertisements).

The obtained collection was labeled similar to "Persons-1000". To annotate numerous mentions of Allah, we added the Deity type to the annotation scheme, but in the current study we consider the Deity type as a subtype of the Person type.

Analyzing the created collection from the point of view of NER difficulties we found that violations in capitalization mainly include all-caps words for the whole tweet and its fragment. Such capitalization is used for emphasizing important words in the text or words related to Allah as in the following example: За все потери ОН дает нам большую награду" ("For all the losses He gives us a great reward"). Also the tweets mention a lot of Eastern names of persons, organizations ("Фастаким Кама умирт" (Fastakim Kama Umirt group), Джабхатфатхаш-Шам (Jabhat Fateh al-Sham)), or local places difficult for correct recognition.

The fraction of tweets with spelling mistakes, unusual shortenings is relatively low. We suppose that this is because the selected users are well-educated, they are professional writers in some sense, in most cases they are leaders of opinions, whose messages are retweeted by many other people. Therefore it is especially useful to study the specific features of their tweet language.

4 Description of NER Model

In our study, we employ the baseline CRF-classifier that utilizes token features, context features, and lexicon features for NER. Then we consider the ways to improve the baseline model adapting it to the Twitter language. The adaptation techniques include the use of two stage-processing and unsupervised word clustering. Besides, we test the impact of tweet normalization on the NER performance.

4.1 Baseline model

Before named entity recognition with CRF, tweets are processed with a morphological analyzer for determining the part of speech, gender, lemma, grammatical number, case and characteristics of words. This information is used to form features of each word for classifying. In the baseline model we consider three types of features: local token features, context features, and features based on lexicons.

4.1.1 Token features

The token features include:

- 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 "Big Small". If all letters are capital then the value is "Big Big". If all letters are small then the value is "Small Small". 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;
- Symbol n-grams. The presence of prefixes, suffixes and other n-grams from the predefined sets in a token.

4.1.2 Context-based features

The group of context features includes two feature types. The first type is local context features. It takes into account all mentioned token feature values of nearby words in two-word window to the right and to the left from the current word.

The second type is the bigram context feature. It contains information about the determined named entity type of the previous word. It helps to find named entity borders more precisely. For example, if the person second name is difficult for recognition, the presence of the first name before this word makes the classification easier.

4.1.3 Features based on lexicons

To improve the quality of recognition, we added special lexicons with lists of useful objects. An object can be a word or a phrase. The lexicons had been created before the current work and were not changed during the study.

To calculate the lexicon features, the system matches the text and lexicon entries. If a token is met in a matched lexicon entry then it obtains the lexicon feature value equal to the length of the found entry. The use of the entry length as a feature helps to diminish the affect of lexical ambiguity. For example, in the list of organizations there is "Apple" as the name of a company. But this word does not necessarily mean a company because it has the second sense of a fruit. In the opposite, if we found in the text the phrase "Lomonosov Moscow State University", which is also included in the organization lexicon, the probability of the organization sense is higher than in the first case. The lexicon feature containing the matched entry length helps the system to distinguish these two cases.

The biggest lexicons are listed in Table 2. The overall size of all vocabularies is more than 335 thousand entities. These lexicons were collected from from several sources: phonebooks, Russian Wikipedia, RuThes thesaurus [15].

Table 2 Vocabularysizes

Vocabulary	Size, objects	Clarification
Famous persons	31482	Famous people
First names	2773	First names
Surnames	66108	Surnames
Person roles	9935	Roles, posts
Verbs of informing	1729	Verbs that usually occur with persons
Companies	33380	Organization names
Company types	6774	Oranization types
Media	3909	Media
Geography	8969	Geographical objects
Geographical adjectives	1739	Geographical adjectives
Usual words	58432	Frequent Russian words (nouns, verbs, adjectives)
Equipment	44094	Devices, equipment, tools

4.2 Adaptation of NER Model to Tweets

4.2.1 Unsupervised word clustering

In previous studies it was shown that unsupervised word clustering on the basis of a large text collection improves the NER performance. In our case we compare the impact of word clusters calculated on a large news collection and large tweet collection. For clustering we use the Word2vec package[1]. It represents words with distributional vectors (word embeddings), computed with the neural network. The semantic similarity between two words is calculated as the cosine measure between two corresponding vectors. The package allows automatic clustering of words according to their calculated similarity. We used the c-bow model with vector sizes equal to 300. Thus, each word has an additional feature – the number of a cluster in that it appears. The news collection utilized for clustering contains two million news documents. For tweet-based clustering we use a tweet collection consisting of randomly extracted Russian tweets and including 8.3 million tweets.

4.1.1 Two-stage prediction

We suppose that for adapting a classifier to a text collection it can be useful to take into account the entities already labeled by the classifier and to memorize the named entity type statistics for future use.

On the first stage the classifier extracts named entities. Then the system collects the class statistics determined in the first stage for each word and used it for features of the second stage. After that, new features together with old ones participate in final classification. These statistics can be collected from the current text (the whole text or its part preceding to the word analysis) or from a large text collection (collection statistics). In case of tweet processing, texts are small therefore only the collection statistics can be used. In our experiments this statistics can be obtained from the FullTweetCollection gathered from the selected user accounts or the labeled TestTweetCollection as described in Section 3.2.

For each word, the system finds all mentions of this word in the processed collection and counts frequencies of determined named entity types for this word. Using these frequencies for each entity type, the system creates additional features, which have one of three values: no_one (if the word has not been recognized as a named entity of the chosen type), best (if the word has been as signed to the chosen named entity type more than in 50% of cases), and rare (if the word has been assigned to the chosen named entity type less than in 50% of cases).

For example, if the word “Russia” was met 500 times in a collection, and the classifier assigned it 200 times to organizations and 300 times to locations, then the values of the global statistics feature for the word “Russia” will be as following: PER –no_one, ORG – rare, LOC – best.

4.2 Normalization of Word Capitalization

As we found that in our tweet collection the share of misprints is not very high we did normalization only for word capitalization. The normalization was based on the large news collection described in Section 4.2. For each word in this collection, we counted how many times the word was written in letter case or capital case when it stands not in the beginning of a sentence. The more frequent case was considered as normal for this word.

We considered the normalization in two variants:

- Variant A. All words in a tweet, except the first one, are changed to a normal form of capitalization;
- Variant B. All words in a tweet including the first one are changed to a normal form of capitalization.

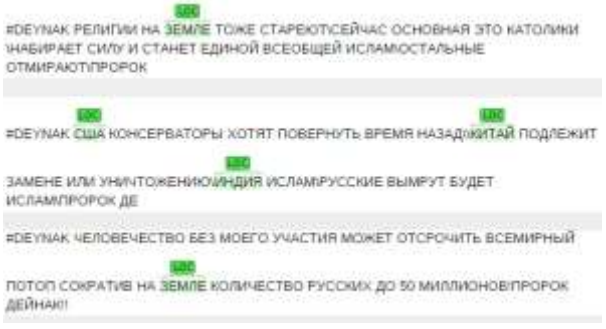

Figure 1 Tweets before normalization

We found that the variant B produces better results and later experimented only with this variant.

Fig. 1 presents several tweets with the manual annotation before normalization. Fig. 2 shows the same tweets after normalization.

[1]https://github.com/dav/word2vec

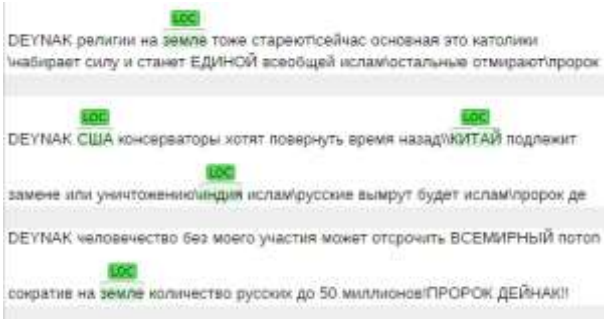

Figure 2 Tweets after normalization

Also the hashtag symbols were removed from a word if this word was found in the news collection to improve its matching with the lexicons.

4 Experiments

In prepocessing we remove mentioned user accounts with "@" and urls in the end of tweets. We consider these data as additional, as meta-information, from which we should not extract names.

We train the described variants of our NER model on the news collection. Table 3 shows results of named entity recognition on the "Persons-1000" collection (cross-validation 3:4). It can be seen that our baseline model is quite good on the news collection and slightly improved after adding clustering features and the two-step approach. In this case the collection statistics is obtained from the same "Persons-1000" collection.

Table 3 News Collection NER Performance

Model	F-measure, %
Baseline	92.49
Baseline+ News clusters	93.48
Baseline+ News clusters + Collection statistics	**93.53**

Then we apply the trained model to the test tweet collection in initial capitalization and normalized capitalization. Table 4 presents the performance of NER models trained on the "Persons-1000" collection for the tweet data. One can see that all models significantly degrade on the tweet collection.

The normalization significantly improves the performance of NER (in contrast to other studies [7]). Word clustering and the collection statistics improve both NER for initial and normalized text collections. Their impact is larger than for the news collection (Table 3). The combination of tweet and news clusters was better than only tweet clusters possibly because of the political and religious character of the gathered collection. In total, the NER performance improves more than 10% on tweet data.

Analyzing mistakes of the best model on the normalized collection we can see still significant share of mistakes because of incorrectly normalized capitalization. We can enumerate the following main subtypes of such problems:

- ambiguous words with different capitalization ("Earth", "Rose"),
- words that should be capitalized in this specific collection. For example, "Paradise" and "Hell" seem to be specific entities in this genre of texts,
- multiword expressions in which each word is usually written in letter case, but together the multiword expression denotes a name and al least the first word should be capitalized. For example, the expression "Московский регион" (Moscow region) is normalized incorrectly because the word "московский" is written in letter case more frequently in the Russian news collection.

Table 4 TweetPerformance

Model	F-measure, TestTweet-Collection	F-measure, Normalized TestTweet-Collection
1) Baseline	64.44%	69.88%
2) Baseline + Collection statistics (TestTweetCollection)	64.99%	70.32%
3) Baseline + Collection statistics (FullTweetCollection)	65.78%	70.44%
4) Baseline + news clusters	66.03%	70.88%
5) Baseline + tweet clusters	66.08%	70.36%
6) Baseline + tweet and news clusters	66.23%	70.89%
7) (2) + tweet and news clusters	**67.27%**	**71.20%**
8) (3) + tweet and news clusters	66.46%	69.73%

5 Conclusion

The paper describes an approach to creating a domain-specific tweet collection written by users frequently discussing Islam-related issues in Russian. We use this collection to study specific features of named entity recognition on Twitter. We found that in contrast to tweets collected randomly, our tweet collection contains relatively small number of spelling errors or strange word shortenings. Specific difficulties of our collection for named entity recognition include a large number of Arabic and other Eastern names (persons, locations, organizations) and frequent use of ALL-CAPS writing for emphasizing main words in messages.

We have studied the transfer of NER model trained on a newswire collection to the created tweet collection and approaches to decrease the degradation of the model because of the transfer. We found that for our specialized text collection, the most improvement was based on normalizing of word capitalization. Two-stage approaches to named entity recognition and word2vec-based clustering were also useful for our task.

In future we plan to improve techniques of tweet normalization and study NER for tweets of followers of the selected users.

References

[1] Antonova, A.Y., Soloviev, A.N.: Conditional Random Field Models for the Processing of Russian. In: Int. Conf. "Dialog 2013", pp. 27- 44. RGGU (2013)

[2] Bocharov, V.V. et al.: "FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian". In: Dialog Conference. (2016)

[3] 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, 18 (4), pp. 467-479 (1992)

[4] Cherry, C., Guo, H.: The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. In: NAACL-2015. pp. 735-745 (2015)

[5] Chrupala, G.: Efficient Induction of Probabilistic Word Classes with LDA. In: 5th Int. Joint Conf. on Natural Language Processing, IJCNLP 2011, pp. 363-372. Asian Federation of Natural Language Processing (2011)

[6] Clark, A.: Combining Distributional and Morphological Information for part of Speech Induction. In: 10th Conf. on European Chapter of the Association for Computational Linguistics, EACL, 1, pp. 59-66. ACL (2003)

[7] Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Bontcheva, K.: Analysis of Named Entity Recognition and Linking for Tweets. Information Processing & Management, 51 (2), pp. 32-49 (2015)

[8] Finin, T., Murnane, W., Karandikar, A, Keller, N., Martineau, J., Dredze, M.: Annotating Named Entities in Twitter Data with Crowdsourcing. In: the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk, pp. 80-88 (2010)

[9] Fromreide, H., Hovy, D., Sogaard, A. Crowdsourcing and Annotating NER for Twitter #drift. In LREC-2014, pp. 2544-2547 (2014)

[10] Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., Ivanov, V.: Introducing Baselines for Russian Named Entity Recognition. In: 14th Int. Conf. CICLing 2013, pp. 329-342. Springer (2013)

[11] Hidayatullah, A.F.: Language Tweet Characteristics of Indonesian Citizens. In: Int. Conf. IEEE-2015. pp. 397-401 (2015)

[12] Hovy, D.: Demographic Factors Improve Classification Performance. In: ACL-2015, pp. 752-762 (2015)

[13] Khoroshevsky, V.F.: Ontology Driven Multilingual Information Extraction and Intelligent Analytics. Web Intelligence and Security. pp. 237-262 (2010)

[14] Kuznetsov, I.P., Kozerenko, E.B., Kuznetsov, K.I., Timonina, N.O.: Intelligent System for Entities Extraction (ISEE) from Natural Language Texts. In: Int. Workshop on Conceptual Structures for Extracting Natural Language Semantics-Sense, (9), pp. 17-25 (2009)

[15] Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. LingvisticaeInvestigationes, 30 (1), pp. 3-26 (2007)

[16] Paris, C., Thomas, P., Wan, S.: Differences in Language and Style Between Two Social Media Communities. In: the 6th AAAI Int. Conf. on Weblogs and Social Media, ICWSM (2012)

[17] Podobryaev, A.V.: Persons Recognition Using CRF Model. In: 15th All-Russian Scientific Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collection", RCDL-2013, pp. 255-258. Demidov Yaroslavl State University (2013)

[18] Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: 13th Conf. on Computational Natural Language Learning, CoNLL, pp. 147-155. ACL (2009)

[19] Ritter, A., Clark, S., Mausam, Etzioni, O.: Named Entity Recognition in Tweets: An Experimental Study. In: EMNLP, pp.1524-1534 (2011)

[20] Ritter, A, Etzioni, O, Clark, S. et al: Open Domain Event Extraction from Twitter. In: Conf. on Knowledge Discovery and Data Mining, KDD, pp. 1104-1112 (2012)

[21] Cherry, C., Guo, H.: The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. In: NAACL-2015. pp. 735-745 (2015)

[22] Trofimov, I.V.: Person Name Recognition in News Articles Based on the Persons-1000/1111-F Collections. In: 16th All-Russian Scientific Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", RCDL 2014, pp. 217-221 (2014)

[23] Yang, Y., Eisenstein, J.: Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis. arXiv preprint arXiv:1511.06052 (2015)