Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain

Mikhail Tikhomirov(✉), N. Loukachevitch, Anastasiia Sirotina, and Boris Dobrov

Lomonosov Moscow State University, Moscow, Russia
tikhomirov.mm@gmail.com, louk_nat@mail.ru, overnastuhed@yandex.ru, dobrov_bv@mail.ru

Abstract. The paper presents the results of applying the BERT representation model in the named entity recognition task for the cybersecurity domain in Russian. Several variants of the model were investigated. The best results were obtained using the BERT model, trained on the target collection of information security texts. We also explored a new form of data augmentation for the task of named entity recognition.

Keywords: Cybersecurity · Named Entity Recognition · Pretraining · Augmentation

1 Introduction

Automatic named entity recognition (NER) is one of the basic tasks in natural language processing. The majority of well-known NER datasets consist of news documents with three types of named entities labeled: persons, organizations, and locations [1,2]. For these types of named entities, the state-of-the-art NER methods usually give impressive results. However, in specific domains, the performance of NER systems can be much lower due to necessity to introduce new types of entities, to establish the principles of their labeling, and to annotate them consistently.

In this paper we discuss the NER task in the cybersecurity domain [3]. Several additional types of named entities for this domain were annotated if compared to general datasets such as software programs, devices, technologies, hackers, and malicious programs (vulnerabilities). The most important entities for this domain are names of malicious software and hackers. However, the annotated dataset contains a modest number of entities of these types. This could be explained by the fact that usually names of viruses and hackers are not known at the time of an attack and are revealed later.

The research was supported by RSF (project No. 20-11-20166). Computational experiments were carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

E. Métais et al. (Eds.): NLDB 2020, LNCS 12089, pp. 16–24, 2020.
https://doi.org/10.1007/978-3-030-51310-8_2

To improve NER performance in such conditions, we suggest using BERT transformers [4] as well as an automatic dataset augmentation method, by which we mean extending a training dataset with sentences containing automatically labeled named entities. In this paper we study how quality of a NER system changes depending on variants of the BERT model used. We experimented with the following models: a multilingual model, a model fine-tuned on Russian data, and a model fine-tuned on cybersecurity texts. We also introduce a new method of dataset augmentation for NER tasks and study the parameters of the method.

2 Related Work

The information extraction task in cybersecurity domain has been discussed in several works. However, most works consider information extraction only from structured or semi-structured English texts [5]. The training corpus presented in [7] does contain unstructured blog posts, but those comprise less than 10% of the corpus. The proposed NER systems are based on such methods as principle of Maximum Entropy [5], Conditional Random Fields (CRF) [6,7]. Gasmi et al. [8] explored two different NER approaches: the CRF-model and neural network based model LSTM-CRF.

Currently, the state-of-the-art models for named entity recognition utilize various contextualized vector representations such as BERT [4], unlike static vector representations, such as word2vec [9]. BERT is pretrained on a large amount of unlabeled data on the language modeling task, and then it can be fine-tuned for a specific task. The paper [13] describes an approach to further training of the multilingual BERT model on the Russian-language data. The new model, called RuBERT, showed an improvement in quality in three NLP tasks in Russian, including named entity recognition [16].

In 2019, the NER shared task for Slavic languages was organized [14]. Most participants and the winner used BERT as the main model. The data had a significant imbalance among the types of entities. For example, the "product" entity was annotated only for 8% of all entities in the Russian data. The results of extracting this type of entities were significantly lower than for other entities.

As far as methods of data augmentation for natural language processing are concerned, they are mainly discussed for such tasks as machine translation and automatic text classification. The simplest augmentation method is to replace source words with their synonyms from manual thesauri or with similar words according to a distributional model trained on a large text collection [17]. In [18] the replacement words were selected among the most probable words according to a language model. The authors of [19] used four simple augmentation techniques for the classification tasks: replacing words with their synonyms, occasional word insertion, occasional word deletion and occasional word order changing. This method was applied to five datasets, showing average improvement of 0.8% for F-score. All four operations contributed to the obtained improvement.

In this paper we discuss a specialized method of data augmentation for named entity recognition. We obtain additional annotated data by inserting named entities in appropriate sentences and contexts.

3 Data

We use a renewed version of Sec_col[1] corpus [3] as a training dataset for the NER task. The final corpus contains 861 unstructured texts (more than 400 K tokens), which are articles, posts, and comments extracted from several sources on cybersecurity. The set of corpus labels (14K labeled entities) includes four general types: PER (persons excluding hackers), ORG (organizations excluding hacker groups), LOC, and EVENT; and five domain-specific types such as PROGRAM (computer programs excluding malware), DEVICE (for various electronic devices), TECH (for technologies having proper names), VIRUS (for malware and vulnerabilities), and HACKER (for single hackers and hacker groups). The annotation principles are described in detail in [3]. The authors of [3] compared different models of NER including CRF and several variants of neural networks on this corpus.

One of the labels, HACKER, is severely underrepresented in the dataset (60 occurrences). The VIRUS label was annotated 400 times, which is lower than for other tags.

4 BERT Models Used in Cybersecurity NER

We explore the use of the BERT model [4] for the NER task in the information-security domain. This model receives a sequence of tokens obtained by tokenization using the WordPiece technique [10] and generates a sequence of contextualized vector representations. BERT training is divided into two stages: pretraining and fine-tuning [12]. At the pretraining stage, the model is trained on the masked language modeling task. At the fine-tuning stage, the task-specific layers are built over BERT; the BERT layers are initialized with the pretrained weights, and further training for the corresponding task takes place.

For Russian, researchers from DeepPavlov [16] trained the model RuBERT on Russian Wikipedia and a news corpus [13]. To do this, they:

- took pre-trained weights from multilingual-bert-base,
- constructed a new vocabulary of tokens of a similar size, better suited for processing Russian texts, thereby reducing the average length of tokenized sequences by 1.6 times, which is critical for the model performance,
- initialized vector representations of new tokens using vectors from multilingual-bert-base in a special way,
- trained the resulting model with a new vocabulary on the Russian Wikipedia and the news corpus.

As part of this study, we evaluated BERT in the NER task in the field of information security with the following pretrained weights: 1) multilingual-bert-base model (BERT), 2) model trained on Russian general data RuBERT, 3) RuCyBERT, which was obtained by additional training RuBERT on information-security texts. Training RuCyBERT was similar to training RuBERT, but

[1] https://github.com/LAIR-RCC/InfSecurityRussianNLP.

without creating a new vocabulary. To do this, the pretraining procedure was launched on 500K cybersecurity texts with the initialization of all weights from RuBERT. The training lasted 500k steps with batch size 6.

All three models have the same architecture: transformer-encoder [15] with 12 transformer blocks, 12 self-attention heads and H = 768 hidden size. The models are fine-tuned for 6 epochs, with B = 16 batch size, with learning rate 5e−5 and T = 128 maximum sequence length. When forming input for the model, only the first token of a word gets a real word label, the remaining tokens get a special label X. At the prediction step, the predicted label of the first token is chosen for the whole word.

5 Augmentation of Training Data

The important classes of named entities in the cybersecurity domain are names of viruses and hackers (including hacker groups). The Sec_col collection, however, includes a quite small number of hackers' names. Many texts related to cybersecurity include only unnamed descriptors (such as *hacker, hacker group, hacker community*).

The core idea of the NER augmentation is as follows: in most contexts where an entity descriptor is mentioned, some other variants of mentions are possible. For Russian, such variants can be: 1) a descriptor followed by a name or 2) just the name alone. The first above-indicated variant of entity mentioning is language-specific, depends on language-specific grammar rules. Consequently, we could augment the collection by adding names after descriptors or by replacing descriptors with names. The following sentences show the examples of the substitution operation for malware.

- **Initial sentence:** Almost 30% are seriously concerned about this issue, another 25% believe that the danger of **spyware** is exaggerated, and more than 15% do not consider this type of threat to be a problem at all.
- **Augmented sentence:** Almost 30% are seriously concerned about this issue, another 25% believe that the danger of **Remcos** is exaggerated, and more than 15% do not consider this type of threat to be a problem at all.

The suggested augmentation includes two subtypes: inner and outer. The inner augmentation involves sentences that contain relevant descriptors within the existing training data. If a sentence meets augmentation restrictions, then the descriptor is replaced with a name or a name is added after the descriptor with equal probability. In both cases, we require that the descriptor must not be followed by a labeled named entity and it must not be preceded by words that agree with the descriptor in gender, number or case, such as adjectives, participles, ordinal numbers, and others.

For the outer augmentation, we look for sentences with relevant descriptors in a collection of unannotated cybersecurity texts. There also must not be any evident named entities (words starting with a capital letter) in a window of certain width around the descriptor. As for this purpose an unannotated collection

is used, we do not know the classes of potential named entities, thus we have to exclude sentences with such entities. Besides, we also require the absence of adjectives before the descriptor. The selected sentences also undergo the procedure of inserting a name after a descriptor or replacing the descriptor with a name with equal probability.

The augmentation has been implemented for two types of named entities: malicious software (VIRUS label) and hackers (HACKER label). 24 virus descriptors and 6 hacker descriptors were used. By means of inner augmentation, 262 additional annotated sentences for viruses and 165 annotated sentences for hackers were created. The outer augmentation can be of an unlimited size.

Inserted named entities are obtained in the following way. We took a large cybersecurity text collection and used it to extract names and sequences of names that follow target descriptors. We created the frequency list of extracted names and chose those names for which frequency was higher than a certain threshold (5). Then we excluded the names that appeared in the annotated training collection and belonged to classes that are different from the target class. The rest of the names were randomly used for insertion into the augmented sentences.

6 Experiments

We compare several variants of the BERT model on the NER task for information security domain. In addition, the results of using augmentation of the labeled data are investigated.

The CRF method was chosen as a baseline model, since in previous experiments with the Sec_col collection, this method showed better results than several variants of neural networks that are usually used for the NER task (BiLSTM with character embeddings) [3]. The CRF model utilizes the following features: token embeddings, lemma, part of speech, vocabularies of names and descriptors, word clusters based on their distributional representation, all these features in window 2 from the current token, tag of the previous word [3].

Table 1 shows the classification results for four models for all labels used, as well as the averaged macro and micro F-measures. It can be seen that the use of the multilingual-bert-base (BERT in the table) gives better results than the CRF model for all types of named entities. The use of the pretrained models on the Russian data (RuBERT) and information security texts (RuCyBERT) gives a significant improvement over previous models.

Since models based on neural networks due to random initialization can give slightly different results from run to run, the results in the tables for all BERT models are given as averaging of four runs. The last row of Table 1 indicates (F-macro std) the standard deviation of the results from the mean. It can be seen that the better the model fits the data, the better the results are, and the standard deviation decreases.

For CRF, all types of the augmentation improved the results of extracting target entities. The best augmentation was inner augmentation, which achieved 43.58 HACKER_VIRUS F-measure, which means an increase in the average

Table 1. Results of basic models

	CRF	BERT	RuBERT	RuCyBERT
DEVICE	31.78	34.04	43.13	**46.77**
EVENT	42.70	60.38	64.49	**67.86**
HACKER	26.58	42.69	52.43	**61.03**
LOC	82.30	90.00	**91.28**	90.01
ORG	68.15	76.10	**78.95**	78.58
PER	67.10	80.99	84.32	**84.56**
PROGRAM	62.15	63.15	64.77	**66.57**
TECH	60.65	67.08	67.60	**69.24**
VIRUS	40.90	40.21	46.92	**54.72**
F-micro	63.95	69.37	71.61	**72.74**
F-macro	53.59	61.63	65.99	**68.82**
F-macro std	–	1.52	0.93	**0.86**

quality of the target named entities by 10% points (almost a third). Macro F1 measure for all types of entities (57.39) was also improved significantly.

Table 2 shows the use of the proposed data augmentation approach to extract two types of named entities HÀCKER and VIRUS with inner and outer augmentations. For the outer augmentation, options for adding 100, 200, 400, 600 augmented sentences for each entity types (HÀCKER and VIRUS) were considered. However, the outer augmentation of 600 sentences gave a stable decrease in the results for all models, and therefore these results are not given in the tables. The "mean F1" column shows the averaging of the values of the F1 measure over all types of entities. The best achieved results are in bold. The results improving the basic results (without augmentation) are underlined.

It can be seen that the multilingual BERT model demonstrates a very high standard deviation on the two types of entities under analysis. Any variant of augmentation reduces the standard deviation, which, however, remains quite high (column F1 std). Two models of outer augmentation increase the quality of extraction of target entities while significantly reducing the standard deviation compared to the original model.

For the RuBERT model, the results are significantly higher than for the previous model, the standard deviation is lower. The augmentation in all cases reduces the standard deviation of F measures for target and all types of entities. The results on the target entities increased with outer augmentation of 200 sentences for both entities. Also, for some reason, the outer augmentation only with viruses positively influenced the extraction of both of them (100 and 200 sentences). The study of this phenomenon is planned to continue.

For RuCyBERT model, the basic performance is much higher, and there is no improvement from the augmentation. The augmentation on average reduces the

standard deviation of F-measure, which leads to the fact that the performance of models with augmentation and the basic model is comparable.

It can be also seen that in almost all experiments the proposed augmentation significantly increases recall, but decreases precision.

Table 2. Models with augmentation

		HACKER_VIRUS				Macro	
		P	R	F1	F1 std	F1	F1 std
BERT	Base (no augmentation)	**46.43**	38.14	41.45	7.23	61.63	1.52
	Inner	36.81	45.44	39.92	3.53	61.26	0.86
	Outer 100	39.13	44.96	41.04	**2.18**	62.02	**0.55**
	Outer 200	39.32	48.24	**42.51**	4.33	**62.21**	0.74
	Outer 400	40.23	45.97	**42.53**	4.59	62.12	1.08
RuBERT	Base (no augmentation)	53.65	47.38	49.67	4.65	65.99	0.93
	Inner	45.01	**55.74**	48.87	3.48	65.92	0.68
	Outer 100	47.46	53.29	49.38	3.1	65.88	0.79
	Outer 200	47.83	55.34	50.71	2.96	66.24	**0.59**
	Outer 400	45.57	53.45	48.46	**2.36**	65.77	0.67
	Outer viruses 100	**57.14**	51.67	**53.79**	3.05	**66.85**	0.64
RuCyBERT	Base (no augmentation)	**61.33**	55.89	57.87	3.75	68.82	0.86
	Inner	52.51	**62.57**	56.03	2.54	68.61	0.53
	Outer 100	50.78	59.69	53.79	2.36	67.78	**0.43**
	Outer 200	52.82	59.61	54.82	3.94	68.06	0.74
	Outer 400	52.42	61.31	55.64	**2.16**	67.93	0.71

7 Conclusion

In this paper we present the results of applying BERT to named entity recognition for cybersecurity Russian texts. We compare three BERT models: multilingual, Russian (RuBERT), and cybersecurity model trained on specialized text collection (RuCyBERT). The highest macro F-score is shown by the domain-specific RuCyBERT model.

For each model, we have also presented a new form of augmentation of labeled data for the NER task, that is adding names after or instead of a descriptor of a certain type. The adding procedure is language-specific. In our case it is based on the Russian grammar. In practically all cases, the augmentation increases recall, but decreases precision of NER. A significant improvement from the augmentation was revealed for relatively weak CRF and multilingual BERT models. For the fine-tuned models, the quality has barely grown. Nevertheless, if in some

cases it is impossible to fine-tune BERT on a specialized collection, the presented augmentation for named entities could be of great use while extracting named entities of non-standard types. The described Sec_col collection and the trained RuCyBERT model can be obtained from the repository[2].

References

1. Sang, E., Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003, vol. 4, pp. 142–147 (2003)
2. Mozharova, V.A., Loukachevitch, N.V.: Combining knowledge and CRF-based approach to named entity recognition in Russian. In: Ignatov, D.I., et al. (eds.) AIST 2016. CCIS, vol. 661, pp. 185–195. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52920-2_18
3. Sirotina, A., Loukachevitch, N.: Named entity recognition in information security domain for Russian. In: Proceedings of RANLP-2019, pp. 1115–1122 (2019)
4. Devlin, J., et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Bridges, R., Jones, C., Iannacone, M., Testa, K., Goodall, J.: Automatic labeling for entity extraction in cyber security. arXiv preprint arXiv:1308.4941 (2013)
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic: models for segmenting and labeling sequence data. In: International Conference on Machine Learning ICML-2001 (2001)
7. Joshi, A., Lal, R., Finin, T., Joshi, A.: Extracting cybersecurity related linked data from text. In: 2013 IEEE Seventh International Conference on Semantic Computing, pp. 252–259. IEEE (2013). https://doi.org/10.1109/ICSC.2013.50
8. Gasmi, H., Bouras, A., Laval, J.: LSTM recurrent neural networks for cybersecurity named entity recognition. In: ICSEA-2018, vol. 11 (2018)
9. Mikolov, T., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
12. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
13. Kuratov, Y., Arkhipov, M.: Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv:1905.07213 (2019)
14. Piskorski, J., Laskova, L., Marcinczuk M., Pivovarova, L., Priban P., et al.: The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages. In: 7th Workshop on Balto-Slavic Natural Language Processing BSNLP-2019, pp. 63–74 (2019)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
16. DeepPavlov documentation. http://docs.deeppavlov.ai/en/master/. Accessed 25 Dec 2019

[2] https://github.com/LAIR-RCC/InfSecurityRussianNLP.

17. Yang Wang, W., Yang, D.: That's so annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In: 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2557–2563 (2015)
18. Kobayashi, S.: Contextual augmentation: data augmentation by words with paradigmatic relations. In: 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-2018, pp. 452–457 (2018)
19. Wei, J.W., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2019, pp. 6381–6387 (2019)