On Biomedical Named Entity Recognition: Experiments in Interlingual Transfer for Clinical and Social Media Texts

Zulfat Miftahutdinov[1 [0000−0002−8467−4824]], Ilseyar Alimova[1 [0000−0003−4528−6631]], and Elena Tutubalina[1,2,3 [0000−0001−7936−0284]]

[1] Chemoinformatics and Molecular Modeling Laboratory, Kazan Federal University
[2] Samsung-PDMI Joint AI Center, Steklov Mathematical Institute at St. Petersburg
[3] Insilico Medicine Hong Kong Ltd, Pak Shek Kok, New Territories, Hong Kong
{zulfatme,alimovailseyar,tutubalinaev}@gmail.com

Abstract. Although deep neural networks yield state-of-the-art performance in biomedical named entity recognition (bioNER), much research shares one limitation: models are usually trained and evaluated on English texts from a single domain. In this work, we present a fine-grained evaluation intended to understand the efficiency of multilingual BERT-based models for bioNER of drug and disease mentions across two domains in two languages, namely clinical data and user-generated texts on drug therapy in English and Russian. We investigate the role of transfer learning (TL) strategies between four corpora to reduce the number of examples that have to be manually annotated. Evaluation results demonstrate that multi-BERT shows the best transfer capabilities in the zero-shot setting when training and test sets are either in the same language or in the same domain. TL reduces the amount of labeled data needed to achieve high performance on three out of four corpora: pretrained models reach 98–99% of the full dataset performance on both types of entities after training on 10–25% of sentences. We demonstrate that pretraining on data with one or both types of transfer can be effective.

Keywords: Biomedical entity recognition · BERT · transfer learning

1 Introduction

Drugs and diseases play a central role in many areas of biomedical research and healthcare. A large part of the biomedical research has focused on scientific abstracts in English; see a good overview of the field in [9]. In contrast to the biomedical literature, research into the processing of electronic health records (EHRs) and user-generated texts (UGTs) about drug therapy has not reached the same level of maturity. The bottleneck of modern supervised models for named entity recognition (NER) is the human effort needed to annotate sufficient training examples for each language or domain. Moreover, state of the art text processing models may perform extremely poorly under domain shift [14]. Recent advances in neural networks, especially deep contextualized word representations

via language models [1, 4, 12, 16] and Transformer-based architectures [19], offer new opportunities to improve NER models in the biomedical field.

In this work, we take the task a step further from existing monolingual research in a single domain [2, 3, 6, 12, 13, 20, 22] by exploring multilingual transfer between EHRs and UGTs in different languages. Our goal is not to outperform state of the art models on each dataset separately, but to ask whether we can transfer knowledge from a high-resource language, such as English, to a low-resource one, e.g., Russian, for NER of biomedical entities. Our transfer learning strategy involves pretraining the multilingual cased BERT [4] on one corpus and transferring the learned weights to initialize training on a gold-standard corpus in another language or domain. In this work, we seek to answer the following research questions: **RQ1:** How well does a BERT-based NER model trained on one corpus works for the detection of drugs and diseases from another language or domain in the zero-shot setting? **RQ2:** Given a small number of training examples, can the NER model perform as well as a model trained on much larger datasets? **RQ3:** Will transfer learning help achieve more stable performance on a varying size of training data?

All experiments are carried out on 4 datasets: English corpora CADEC [10] and n2c2 [7], a dataset of EHRs in Russian [17], and our novel dataset of UGTs in Russian. All three existing corpora share an entity of interest with our corpus. To our knowledge, this is the first work exploring the interlingual transfer ability of multilingual BERT on bioNER on two domains in English and Russian. Our experiments are available at `https://github.com/dartrevan/multilingual_experiments`.

2 Data

Each corpus is characterized by two parameters: (i) language: English (EN) or Russian (RU); (ii) domain: electronic health records (EHRs) or user-generated texts (UGTs). A statistical summary of the datasets is presented in Table 1. Since all corpora have different annotation schemes for disease-related entities, and these subtypes are highly imbalanced in the corpus, we join them into a single primary type named *Disease*. Further, we unify the names of four datasets according to their characteristics.

CADEC (EN UGT) CSIRO Adverse Drug Event Corpus [10] contains medical forum posts taken from *AskaPatient.com* about 12 drugs of two categories: Diclofenac and Lipitor. Medical students and computer scientists annotated the dataset. The agreement between four annotators computed on a set of 55 user posts was approximately 78% for Diclofenac and 95% for Lipitor posts.

Our dataset of UGTs (RU UGT) We utilized and annotated user posts in Russian from a publicly accessible source *Otzovik.com*; we note that we have obtained all reviews without accessing password-protected information. Four annotators from the I.M. Sechenov First Moscow State Medical University and the department of pharmacology of the Kazan Federal University were asked to read the review and highlight all spans of text including drug names and patient's

Table 1. Summary statistics of four datasets. Summary of each dataset includes the number of Drug and Disease entities, the number of documents and sentences, the average length of a document (in sentences), the average length of a sentence (in tokens), the average length of a Drug/Disease entity (in tokens).

Corpus	**Disease subtypes**	**Drug**	**# doc.**	**# sent.**	**Avg. doc. len.**	**Avg. sen. len.**	**Avg. Drug len.**	**Avg. Dis. len.**
CADEC (EN UGT) [10]	ADR, Symptom, Disease, Finding (6590)	Drug (1798)	1249	7670	6.14	8.27	1.11	2.48
n2c2 (EN EHR) [7]	ADE, Reason (7984)	Drug (26797)	503	70960	140.51	11.32	1.18	1.80
Our dataset (RU UGT)	ADR, Disease (2429)	Medication (1195)	400	4230	10.57	6.82	1.26	2.22
RU EHR [17]	Disease, Symptom (7874)	Drug (3479)	159	16835	105.86	6.14	1.27	2.91

health conditions experienced before/during/after the drug use. The agreement between two annotators computed on a set of 100 posts was 72%.

Russian EHRs (RU EHR) Shelmanov et al. [17] created a corpus of Russian clinical notes from a multi-disciplinary pediatric center. The authors extended an annotation scheme from the *CLEF eHealth 2014 Task 2*.

n2c2 (EN EHR) This corpus consists of de-identified EHRs [7]. Two independent annotators annotated each record in the dataset and a third annotator resolved conflicts. For both EHR corpora, the agreement rates were not provided.

3 Models

For NER, we utilize BERT with a softmax layer over all possible tags as the output. Word labels are encoded with the BIO tag scheme. The model was trained on a sentence level. Due to space constraints, we refer to [4, 12] for more details. In particular, we use BERT_{base}, Multilingual Cased (Multi-BERT), which is pretrained on 104 languages and has 12 heads, 12 layers, 768 hidden units per layer, and a total of 110M parameters. All models were trained without fine-tuning or explicit selection of parameters. The loss function became stable (without significant decreases) after 35-40 epochs. We use Adam optimizer with polynomial decay to update the learning rate on each epoch with warm-up steps in the beginning. As baselines, we utilized LSTM-CRF with default settings from the Saber library [5] and BioBERT [12]. For LSTM-CRF, we adopted (i) 200-dim. *word2vec* embeddings trained on 2.5M of health-related posts in English [18] and (ii) 300-dim. *word2vec* embeddings trained on the Russian National Corpus [11].

4 Experiments and Evaluation

We randomly split each of the datasets into 70% training set and 30% test set. We trained a total of 720 models on one machine with 8 NVIDIA P40 GPUs. The training of all models took approximately 96 hours. We compare all models in

terms of precision (P), recall (R), and F1-score (F) on the test sets with exactly matching criteria via a CoNLL script.

Comparison with Baselines Table 2 shows the in-corpus (IC) performance of Multi-BERT with BioBERT and LSTM-CRF when trained and tested on the same corpus. On all datasets, BERT-based models achieve the best scores over LSTM-CRF based on word embeddings. The difference in the performance of BioBERT and Multi-BERT is not statistically significant; we measured significance with the two-tailed t-test ($p \leq 0.05$). All models achieve much higher performance for the detection of drugs rather than diseases; it can be explained by boundary problems in multi-word expressions (see the av. length in Table 1).

Zero-Shot Transfer To answer **RQ1**, we trained Multi-BERT on one corpus and then applied it to another language/domain in a zero-shot fashion, i.e., without further training. Results of the out-of-corpus (OOC) performance of Multi-BERT are presented in Table 3. For drug recognition, the best generalizability is achieved when training on EHRs and evaluated on UGTs in English. For OOC performance on the EN UGT corpus, the model reaches F1-scores of 77.08% and 36.31% when trained on the EN EHR and RU UGT corpora, respectively, while IC reaches the F1-score of 84.88%. We note that the number of sentences in the EN EHR corpus is nine times higher than in the EN UGT corpus. 78% of Drug tokens in the EN UGT corpus are presented in the EN EHR set (see Table 4). For OOC performance on the RU UGT corpus, the model achieves F1-scores of 26.31% and 34.78% when trained on the EN UGT and EN EHR corpora, respectively, while the IC performance is F1-score of 60.45%.

For disease recognition, Multi-BERT generalizes much worse to corpora other than it was trained on. For OOC performance on the RU UGT corpus, the model achieves F1-scores of 24.12% and 30.86% when trained on the EN UGT and RU EHR corpora, respectively, while the IC performance is F1-score of 49.35%. For OOC performance on the EN UGT corpus, the model obtains F1-scores of 37.94% and 4.32% when trained on the RU UGT and EN EHR corpora, respectively, while the IC performance is F1-score of 67.25%. One possible explanation might be that there are well-known differences in layperson language and professional medical terms.

Few-Shot Transfer Transfer learning aims to solve the problem on a "target" dataset using knowledge learned from a "source" dataset [5,15,21]. In the transfer learning setting, the BERT-based NER model was pretrained on one of three "source" datasets (see Table 2 for the IC performance of these models). To answer **RQ2** and **RQ3**, we begin with a random sampling of 50 sentences from a "target" training set, train the pretrained model on this subsampled dataset, and test it on the "target" test set. Next, we increase the sample size by 50 sentences of the "target" training set and repeat the described procedure, doing so up to 2000 sentences of the training set. In each round, we train from scratch to avoid overfitting, as suggested in [8].

Table 2. In-corpus (IC) performance of multi-BERT with comparison to BiLSTM-CRF and BioBERT, measured by Precision, Recall, and F1-score with an exact matching criteria.

Corpus	Models	Disease			Drug		
		P	R	F	P	R	F
EN EHR (n2c2)	Multi-BERT	55.05	63.91	59.15	92.21	92.58	92.39
	BioBERT	56.33	65.56	60.60	92.39	92.97	92.68
	LSTM-CRF	55.00	56.95	55.96	89.87	89.70	89.79
EN UGT (cadec)	Multi-BERT	65.62	68.96	67.25	79.40	91.18	84.88
	BioBERT	67.14	69.88	68.48	87.27	91.73	89.44
	LSTM-CRF	64.68	62.77	63.71	78.50	70.41	74.23
RU UGT	Multi-BERT	45.93	53.33	49.35	58.85	62.14	60.45
	LSTM-CRF	27.78	17.44	21.43	37.74	40.31	38.98
RU EHR	Multi-BERT	78.61	75.96	77.26	87.18	82.93	85.00
	LSTM-CRF	62.00	61.69	61.85	62.00	79.49	69.66

Table 3. Out-of-corpus (OOC) performance of Multi-BERT in the zero-shot setting. OOC performance is derived by training on one corpus (train) and testing on another (test).

Train	Test	Disease			Drug		
		P	R	F	P	R	F
EN UGT	EN EHR	43.05	7.47	12.73	58.23	71.9	64.35
	RU UGT	20.61	29.07	24.12	23.45	29.97	26.31
	RU EHR	7.73	44.33	13.17	5.81	91.18	10.92
EN EHR	EN UGT	2.25	51.58	4.32	78.09	76.09	77.08
	RU UGT	0.77	12.84	1.44	42.5	29.43	34.78
	RU EHR	3.33	2.85	3.07	5.35	72.65	9.97
RU UGT	EN EHR	11.9	3.23	5.08	14.63	75.3	24.50
	EN UGT	31.98	46.61	37.94	23.22	83.22	36.31
	RU EHR	10.22	41.37	16.4	28.75	44.27	34.86
RU EHR	EN EHR	0.5	20.00	0.97	46.15	37.5	41.38
	RU UGT	24.88	40.65	30.86	17.95	17.95	17.95
	EN UGT	43.78	28.12	34.24	35.90	23.73	28.57

Table 4. Summary statistics of Byte Pair Encoding (BPE) tokens of entities in four datasets. Summary includes the number of unique BPE tokens, intersection between un. tokens, percentage of shared tokens from unique set.

Dataset D_1	Dataset D_2	Entity Type	# un. BPE in D_1	# un. BPE in D_2	$D_1 \cap D_2$	% from D_1	% from D_2
EN UGT	EN EHR	Drug	528	2401	410	78%	17%
EN UGT	EN EHR	Disease	2338	2491	1172	50%	47%
RU UGT	RU EHR	Drug	696	896	381	55%	43%
RU UGT	RU EHR	Disease	1487	2427	1011	68%	42%

Table 5. Summary of the number of training sentences needed to achieve 99% of the full dataset performance with Multi-BERT with pretraining.

Pretrain	Entity type	EN UGT	RU UGT	RU EHR
Best pretrain	Drug	500	700	550
Worst pretrain	Drug	900	650	1200
No pretrain	Drug	1050	1000	1500
Best pretrain	Disease	1050	700	1850
Worst pretrain	Disease	1050	900	1850
No pretrain	Disease	1300	1100	1850

For each pretraining setup, we record the size of the subset when the model achieves at least 99% of the F1-measure achieved on the full dataset. Results for the RU UGT, RU EHR, and EN UGT datasets are given in Table 5 and Fig. 1. Multi-BERT pretrained on the EN UGT set and trained on 2000 sentences from the EN EHR corpus (2.81% of the full corpus) obtains 92% F1 and 76% F1 of the full dataset performance on drugs and diseases respectively. As shown in Table 5 and Fig. 1, models with transfer knowledge outperform the models without the pretraining phase even in cases when both domain and language shifts between "source" and "target" sets. Using the transfer learning strategy could require up to 550 sentences less than training from scratch. In particular, models require only 10% and 23% of the EN UGT and RU URT corpora respectively to achieve results as good as full dataset performances. We believe that this observation is very crucial for low resource languages and new domains (e.g., social media, clinical trials). We observe that the performance of models with pretraining setup trained on the different numbers of sentences becomes more stable in terms of deviations between F1-scores (see Fig. 1).

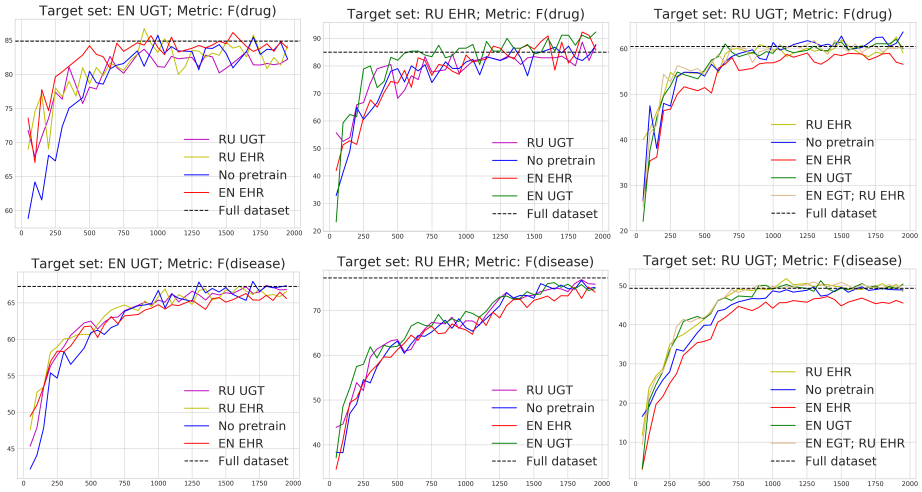

Fig. 1. Performance of Multi-BERT models with pre-training on the source dataset (a corpus's name in a legend) or without pre-training ("No pretrain" line) for the EN UGT, RU UGT, RU EHR datasets. Y-axis: F1-scores for detection of Drug or Disease mentions, X-axis: the number of sentences used for training.

5 Conclusion and Future Work

We studied the task of recognition of drug and disease mentions in English and a low-resource language in the biomedical area, using a newly collected Russian corpus of user reviews about drugs (RU UGT) with 3,624 manually annotated entities. We ask: can additional pretraining on an existing dataset be helpful for bioNER performance of multilingual BERT-based NER model on a new dataset with a small number of labeled examples if the domain, the language, or both shift between these datasets? Our study consisted of over 720 models trained on different subsets of two corpora in English and two corpora in Russian. For each language, we experimented with the clinical domain, i.e., electronic health records, and the social media domain, i.e., reviews about drug therapy. As expected, models with pretraining on data in the same language or the same domain obtain better results in zero-shot or few-shot settings. To our surprise, we found that pretraining on data with two shifts can be effective. The model with the best pretraining achieves 99% of the full dataset performance using only 23.56% of the training data on our RU URT corpus, while the model with pretraining on data with two shifts (the EN EHR set) used 26.1% of the training data. The model without pretraining achieves similar results on the RU URT corpus using 31.97% of the training set.

We foresee three directions for future work. First, transfer learning and multi-task strategies on three and more domains remain to be explored. Second, a promising research direction is the evaluation of multilingual BERT on a broad set of entities. Third, future research will focus on the creation of fine-grained

entity types in our corpus of Russian reviews that can help in finding associations between drugs and adverse drug reactions.

Acknowledgments We thank Sergey Nikolenko for helpful discussions. This research was supported by the Russian Science Foundation grant # 18-11-00284.

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78 (2019)
2. Crichton, G., Pyysalo, S., Chiu, B., Korhonen, A.: A neural network multi-task learning approach to biomedical named entity recognition. BMC bioinformatics **18**(1), 368 (2017)
3. Dang, T.H., Le, H.Q., Nguyen, T.M., Vu, S.T.: D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information. Bioinformatics **34**(20), 3539–3546 (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
5. Giorgi, J., Bader, G.: Towards reliable named entity recognition in the biomedical domain. Bioinformatics (Oxford, England) (2019)
6. Gupta, A., Goyal, P., Sarkar, S., Gattu, M.: Fully contextualized biomedical ner. In: European Conference on Information Retrieval. pp. 117–124. Springer (2019)
7. Henry, S., Buchan, K., Filannino, M., Stubbs, A., Uzuner, O.: 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association (2019)
8. Hu, P., Lipton, Z.C., Anandkumar, A., Ramanan, D.: Active learning with partial feedback. arXiv preprint arXiv:1802.07427 (2018)
9. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics **17**(1), 132–144 (2015)
10. Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C.: Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics **55**, 73–81 (2015)
11. Kutuzov, A., Kunilovskaya, M.: Size vs. structure in training corpora for word embedding models: araneum russicum maximum and russian national corpus. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 47–58. Springer (2017)
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (09 2019)
13. Miftahutdinov, Z., Tutubalina, E.: Deep neural models for medical concept normalization in user-generated texts. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 393–399 (2019)
14. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669 (2019)

15. Pan, S., Yang, Q.: A survey on transfer learning. ieee transaction on knowledge discovery and data engineering, 22 (10) (2010)
16. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). vol. 1, pp. 2227–2237 (2018)
17. Shelmanov, A., Smirnov, I., Vishneva, E.: Information extraction from clinical texts in russian. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue. vol. 14, pp. 537–549 (2015)
18. Tutubalina, E., Miftahutdinov, Z.S., Nugmanov, R., Madzhidov, T., Nikolenko, S., Alimova, I., Tropsha, A.: Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. Russian Chemical Bulletin **66**(11), 2180–2189 (2017)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
20. Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M., Leser, U.: Huner: Improving biomedical ner with pretraining. Bioinformatics (2019)
21. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**(1), 9 (2016)
22. Zhao, S., Liu, T., Zhao, S., Wang, F.: A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 817–824 (2019)