KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions

Zulfat Miftahutdinov
Kazan Federal University
Kazan, Russia

Andrey Sakhovskiy
Kazan Federal University
Kazan, Russia

Elena Tutubalina
Kazan Federal University
Kazan, Russia

Abstract

This paper describes neural models developed for the Social Media Mining for Health (SMM4H) 2020 shared tasks. Specifically, we participated in two tasks. We investigate the use of a language representation model BERT pretrained on a large-scale corpus of 5 million health-related user reviews in English and Russian. The ensemble of neural networks for extraction and normalization of adverse drug reactions ranked first among 7 teams at the SMM4H 2020 Task 3 and obtained a relaxed F1 of 46%. The BERT-based multilingual model for classification of English and Russian tweets that report adverse reactions ranked second among 16 and 7 teams at two first subtasks of the SMM4H 2019 Task 2 and obtained a relaxed F1 of 58% on English tweets and 51% on Russian tweets.

1 Introduction

Text classification, named entity recognition (NER), and medical concept normalization (MCN) in free-form texts are crucial steps in every text-mining pipeline. Here we focus on discovering adverse drug reaction (ADR) concepts in Twitter messages as part of the Social Media Mining for Health (SMM4H) 2020 shared tasks (Klein et al., 2020).

This work is based on the participation of our team, named *KFU NLP*, in two tasks. Organizers of SMM4H 2020 Task 2 provided participants with train, dev, and test sets of English and Russian tweets annotated at the message level with a binary annotation indicating the presence or absence of ADRs. For Task 3, train, dev, and test sets include tweets with text spans of reported ADRs and their corresponding medical codes from the Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al., 1999).

Neural architectures based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have achieved state-of-the-art results in the biomedical domain. For Task 2, we conduct extensive experiments with two SMM4H 2020 train sets separately and a union of these Russian and English sets. For Task 3, we utilize an additional English training sets and dictionaries. We investigate the following versions of BERT:

(1) BioBERT v.1.1 (Lee et al., 2020), pretrained on English texts from PubMed and PMC;

(2) BERT$_{base}$, Multilingual Cased, pretrained on 104 languages, this model was used for the initialization of two models listed below;

(3) EnDR-BERT (Tutubalina et al., 2020), pretrained on the English corpus of 2.6M health-related comments;

(4) EnRuDR-BERT (Tutubalina et al., 2020), pretrained on English and Russian corpora of 5M health-related texts.

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 51–56
Barcelona, Spain (Online), December 12, 2020.

Pretrained weights of domain-specific EnDR-BERT and EnRuDR-BERT models are available at https://github.com/cimm-kzn/RuDReC. The source code for our model for Task 2 is available at https://github.com/Andoree/smm4h_classification.

The paper is organized as follows. We describe our experiments in the classification of multilingual tweets that report adverse reactions in Section 2. In Section 3, we describe our pipeline for NER and medical concept normalization (MCN) for Task 3. Finally, we discuss future directions in Section 4.

2 Task 2: Classification of Multilingual Tweets

The goal of this task is to detect tweets that report an adverse effect of a medication. We present our results for Russian and English subsets of tweets. The data for Task 2 consists of 3 distinct sets of tweets posted in English, Russian, and French. We present our results on two of them: Russian and English with 20,544 and 6,090 tweets, respectively. The training sets are highly imbalanced as only 1,903 English and 533 Russian tweets are labeled as positive examples. Further, we refer to the union of two train sets as a *bilingual* train. We note that the dev sets are not used for training.

2.1 Models

We utilize three BERT-based models for classification: (i) multilingual BERT_{base}[1], (ii) EnDR-BERT[2], and (iii) EnRuDR-BERT[3]. For BERT models, we compared the following training approaches. First, we fine-tuned the mentioned BERT models on training sets. Second, we combined English and Russian subsets of tweets and trained EnRuDR-BERT using their union as a training set. Third, we tried to improve the Russian subtask results using annotated sentences from the RuDReC (Tutubalina et al., 2020) and PsyTAR (Zolnoori et al., 2019) corpora. Annotations include the following types of entities: adverse drug reaction, drug indication, drug effectiveness/ineffectiveness. We pretrained EnRuDR-BERT on the multilabel sentence classification task. Preprocessing of the dataset included the following steps: (i) replacement of all URLs with word "link"; (ii) masking of all @user mentions by @username tag; (iii) mapping of some emojis to the corresponding words (for example, we replaced the pill and syringe emojis with the words pill and syringe); (iv) fix of "&" ampersand representation.

In addition, we use a classification architecture (Kim, 2014) based on convolutional neural networks (CNN) (LeCun et al., 1998) with FastText embeddings (Bojanowski et al., 2017) trained on 1.4 Russian reviews about health & beauty from the RuDReC corpus. The CNN model consisted of 3 convolutional layers with kernel sizes of 3, 4, 5, and ReLU activation. Each convolutional layer consisted of 128 filters and was followed by a max-pooling layer. We used a dense layer with sigmoid activation as the output layer. We used the Keras (Chollet and others, 2015) implementation[4] of the CNN model. For FastText and CNN experiments, we removed punctuation, lowercased, and tokenized all texts using NLTK (Loper and Bird, 2002).

2.2 Experiments

For the classification task, each BERT model was trained for 3 epochs with the learning rate of $3 * 10^{-5}$ using Adam optimizer. We defined the first 10% of the training steps as warm-up steps and set batch size to 64 and maximum sequence size to 128. We utilized Tensorflow (Abadi et al., 2016) implementation of BERT[5] with softmax output activation for fine-tuning and sigmoid output activation for multilabel classification pretraining. We run multilabel classification pretraining for 1 epoch with the learning rate of $2 * 10^{-5}$ and a batch size of 64. We used a classification threshold of 0.5 for all models. We limited the CNN model to 10 epochs with early stopping after 3 epochs with no accuracy improvement on the dev set and set the batch size to 128.

During the evaluation of BERT models, we encountered an instability of results on the dev set. For the final submissions, we used an ensemble of 10 BERT models with a simple voting scheme to solve this

[1] https://github.com/google-research/bert/blob/master/multilingual.md
[2] https://huggingface.co/cimm-kzn/endr-bert
[3] https://huggingface.co/cimm-kzn/enrudr-bert
[4] https://github.com/ShawnyXiao/TextClassification-Keras
[5] https://github.com/google-research/bert

Model	Dev set			Test set (official results)		
	P	R	F1	P	R	F1
Russian tweets						
Multilingual BERT, train on Russian tweets	0.20	**0.87**	0.32	–	–	–
FastText+CNN, train on Russian tweets	0.28	0.79	0.41	–	–	–
EnRuDR-BERT, train on Russian tweets	0.46	0.37	0.41	–	–	–
EnRuDR-BERT, bilingual train	0.44	0.57	0.50	–	–	–
EnRuDR-BERT, ensemble, bilingual train, pretrained on RuDReC+PsyTAR	**0.55**	0.53	**0.54**	**0.54**	0.48	**0.51**
Average scores provided by organizers	–	–	–	0.36	**0.58**	0.43
English tweets						
EnRuDR-BERT, train on English tweets	**0.65**	0.64	0.64	–	–	–
EnRuDR-BERT, bilingual train	**0.65**	0.65	**0.65**	–	–	–
EnDR-BERT, ensemble	0.60	**0.68**	0.64	**0.63**	0.54	**0.58**
Average scores provided by organizers	–	–	–	0.42	**0.59**	0.46

Table 1: Text classification results on the SMM4H Task 2 dev and test sets.

problem.

Table 1 shows the performance of BERT models and CNN in Task 2 in terms of precision, recall, and F1-score. The following conclusions can be drawn based on the results. First, the FastText + CNN approach outperformed multilingual BERT in terms of F1-score on the Russian subtask. However, EnRuDR-BERT, fine-tuned on Russian tweets, showed a performance comparable to CNN. An explanation for this might be that EnRuDR-BERT and FastText embeddings were pretrained on healthcare domain texts, and domain-specific pretraining considerably increases the quality of fine-tuning. Second, the Russian training set's replacement with the bilingual training set resulted in a significant increase in recall (20%) and F1-score (9%) on the Russian subtask. Third, the addition of the Russian data to the training set does not significantly improve the English subtask's performance. Finally, the combination of multilabel classification pretraining and simple ensembling increased F1-score by 4%.

3 Task 3: Extraction and Normalization of Adverse Reactions

This task's objective is to detect ADR mentions and then map these entities to concepts in a controlled vocabulary.

3.1 Named Entity Recognition

Following the best results in SMM4H 2019 Task 2 & 3 (Miftahutdinov et al., 2019), we utilize a BERT-based model for the named entity recognition task.

We experiment with two initialization checkpoints for the language model: BioBERT and EnDR-BERT. We investigate dictionary-based (gazetteer) features and two strategies of extra training data utilization. Dictionary-based features are calculated for each token in a text as follows: first, all the occurrences of predefined vocabulary entries were found in the text, then the first token of the matched part tagged was with B-tag, the last with I-tag, and all other tokens in the text with O-tag. The dictionary-based features are concatenated with the representation learned by the neural network that captures an entity's extensional semantic information. We adopted the dictionaries from previous work (Miftahutdinov et al., 2017).

As extra training data for the NER task, we used the CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015). The first strategy of exploiting the dataset is to train the model on the CADEC data and then fine-tune on the SMM4H train set. The second strategy is to combine these two datasets into one (see '+CADEC' in Table2).

3.2 Normalization

For the normalization task, we applied two models: (i) a classifier (Miftahutdinov et al., 2019; Miftahutdinov and Tutubalina, 2019), (ii) a novel neural model based on similarity distance of BERT vectors of concepts. We also evaluate the combination of the two approaches. Following (Miftahutdinov et al., 2019), we utilize additional data for training. Other corpora are filtered to match a vocabulary of the SMM4H 2020 train set.

Classification over Concepts State-of-the-art studies consider the concept normalization task as a classification problem. We develop a supervised classifier: first, we convert each mention into a vector representation using BERT with mean pooling over layers. Second, we add a softmax layer to convert values to conditional probabilities. The size of the softmax layer is the number of concepts in the terminology.

Metric Learning The second approach is based on metric learning. The intuition here is to map concepts and entities into common embedding space such that entities and their concepts are close to each other. To encode entities, we utilized the BERT model. To encode the concept, first, we extract textual representation from UMLS (Bodenreider, 2004). In particular, UMLS contains concept names of each concept. Each of the concept names is utilized as a textual representation of a concept. Textual representations are then encoded using the BERT model. The entity encoder and the concept name encoder share the weights. Given the embedding u_m of the mention m and u_{c_i} embedding of the i-th concept name c_i, model output is obtained as follows:

$$cid(argmin_{c_i}||u_m, u_{c_i}||), \tag{1}$$

where cid(*) is a function that maps the concept name to the corresponding concept id. The approach included almost no preprocessing steps except lowercasing entities and concept names.

To encode mentions and concepts, we use BERT model fine-tuned using a triplet loss (Reimers and Gurevych, 2019). Given an user-generated entity mention m, a positive concept name c and a randomly sampled concept name n as negative example, triplet loss tunes the network such that the distance between m and c is smaller than the distance between m and n. Mathematically, we minimize the following loss:

$$max(||u_m, u_c|| - ||u_m, u_n|| + \epsilon, 0) \tag{2}$$

where u_m, u_c, u_n are the embeddings for m, c, n, $||u_*, u_*||$ a distance metric, ϵ is margin that ensures that u_c is at least ϵ closer to u_m than u_n. As metric, we use Euclidean distance or cosine similarity and we set $\epsilon = 1$ in our experiments.

Combined Method Since the classification approach can't handle out of vocabulary cases we combined two models based on a threshold. For instance, given (i) prediction c_{bs} from from BERT-based similarity method with the distance equals to d and (ii) prediction c_{clf} from the classification approach, the final prediction is set to c_{bs}, if d is less than a threshold, and to c_{clf}, otherwise. Varying the threshold we can customize the model to make predictions based on classification or metric learning approach. With a small threshold, the combined method will rarely output predictions from a metric learning approach and vice versa. It's reasonable to set low threshold values when the training set covers most of the concepts from the test dataset.

3.3 Experiments

For the NER sub-task, each network was trained for 50 epochs with batch size set to 32. We used the Adam algorithm (Kingma and Ba, 2015) as the optimizer with an initial learning rate $5 * 10^{-5}$. We use precision, recall, and F-measure for evaluation. Table 2 shows the performance of the models on the development set. It could be seen that using the fine-tuned language model and additional data gives a significant improvement in F-measure. For both submissions, the model trained using EnDR-BERT, gazetteer features, and extra training data was utilized.

Model	P	R	F1
BioBERT	51.09	54.36	52.68
EnDR-BERT	**58.20**	52.99	55.47
EnDR-BERT, gazetteer features	57.65	53.28	55.38
EnDR-BERT, gazetteer features, CADEC pretrain	57.65	55.53	56.57
EnDR-BERT, gazetteer features, +CADEC	57.65	**57.97**	**57.81**

Table 2: NER performance on the dev set.

Model	Acc@1
EnDR-BERT classifier	42.14
EnDR-BERT classifier, +additional data	44.90
EnDR-BERT, triplet loss	36.91
EnDR-BERT, triplet loss, +additional data	37.19
Combined method, threshold 4.5	45.17
Combined method, threshold 8.2	43.25

Table 3: MCN performance in terms of accuracy@1 on the dev set.

Run name	P	R	F1
ADR Detection Evaluation (Relaxed)			
KFU NLP Team, run 1	79.1	72.3	75.5
KFU NLP Team, run 2	79.1	72.3	75.5
Average scores	60.7	55.7	56.4
End-to-End Evaluation (Relaxed)			
KFU NLP Team, run 1	46.1	42.7	44.3
KFU NLP Team, run 2	48.2	44.6	46.3
Average scores	31.2	29	29.2

Table 4: Performance of our models in SMM4H 2020 Task 3 (official results).

For the normalization task, we trained the classifier based on EnDR-BERT. The classifier is trained for 50 epochs with the Adam optimizer. To train BERT with triplet loss, we generated 109,605 triplets. For each positive concept name, 15 negative examples were generated from the MedDRA vocabulary. The model is trained for 10 epochs. Table 3 shows performance of the normalization models on the development set. The combination of the EnDR-BERT classifier and the BERT-similarity approach with a threshold equal to 4.5 outperformed other models. For the final submission, we selected a combined approach with a threshold equal to 8.2 for run 1 and a combined approach with a threshold equal to 4.5 for run 2.

Table 4 shows a comparison of the model to the official average scores computed using the participants' submissions. Our model has obtained the highest relaxed F1 score of 46.3% and 75.5% on ADR detection and end-to-end tasks, respectively.

4 Conclusion

In this work, we have explored an application of domain-specific BERT models pretrained on health-related user reviews in English and Russian to the task of multilingual text classification and extraction and normalization of adverse drug reactions. Experiments have shown that our BERT models outperform general multilingual BERT and BioBERT on two tasks, achieving the best results in Task 3.

There are several directions for future work arising from the results of our experiments. One potential direction is to investigate the impact of multilingual transfer learning from other corpora on the classification of Russian health-related texts. Another future direction is to verify the efficiency of English data for text classification in languages other than Russian. For Task 2, transfer from English to French remain to be explored.

Acknowledgements

This research was supported by the Russian Science Foundation grant #18-11-00284.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations 2015*.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. 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*, pages 393–399.

Z.Sh. Miftahutdinov, E.V. Tutubalina, and A.E. Tropsha. 2017. Identifying disease-related expressions in reviews using conditional random fields. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 1(16):155–166.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. Kfu nlp team at smm4h 2019 tasks: Want to extract adverse drugs reactions from tweets? bert to the rescue. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 52–57.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2020. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 07.

Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, Mike Conway, et al. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for ssri and snri medications. *Journal of biomedical informatics*, 90:103091.