KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue

Zulfat Miftahutdinov and **Ilseyar Alimova**
Kazan Federal University,
Kazan, Russia
zulfatmi@gmail.com
ISAlimova@kpfu.ru

Elena Tutubalina
Kazan Federal University,
Kazan, Russia
Samsung-PDMI Joint AI Center,
PDMI RAS, St. Petersburg, Russia
elvtutubalina@kpfu.ru

Abstract

This paper describes a system developed for the Social Media Mining for Health (SMM4H) 2019 shared tasks. Specifically, we participated in three tasks. The goals of the first two tasks are to classify whether a tweet contains mentions of adverse drug reactions (ADR) and extract these mentions, respectively. The objective of the third task is to build an end-to-end solution: first, detect ADR mentions and then map these entities to concepts in a controlled vocabulary. We investigate the use of a language representation model BERT trained to obtain semantic representations of social media texts. Our experiments on a dataset of user reviews showed that BERT is superior to state-of-the-art models based on recurrent neural networks. The BERT-based system for Task 1 obtained an F1 of 57.38%, with improvements up to +7.19% F1 over a score averaged across all 43 submissions. The ensemble of neural networks with a voting scheme for named entity recognition ranked first among 9 teams at the SMM4H 2019 Task 2 and obtained a relaxed F1 of 65.8%. The end-to-end model based on BERT for ADR normalization ranked first at the SMM4H 2019 Task 3 and obtained a relaxed F1 of 43.2%.

1 Introduction

Short-text communication forms, such as Twitter microblogging, present a wide variety of facts and opinions on numerous topics, and this treasure trove of information is currently severely underexplored. Here we focus on the problem of discovering adverse drug reaction (ADR) concepts in Twitter messages as part of the Social Media Mining for Health (SMM4H) 2019 shared tasks.

This work is based on the participation of our team, named *KFU NLP*, in the first three tasks. Organizers of SMM4H 2019 Tasks 1-3 (Weissenbacher et al., 2019) provided participants with datasets of English tweets annotated at the message level with binary annotation indicating the presence or absence of ADRs, text spans of reported ADRs, and their corresponding medical codes from the Medical Dictionary for Regulatory Activities (MedDRA). The goal of Task 1 is to classify the tweets according to the presence of ADRs. For the second task, named entity recognition (NER) aims to detect the mentions of ADRs. The third and final task is designed as an end-to-end problem, intended to perform full evaluation of a system operating in real conditions: given a set of raw tweets, the system has to find the tweets that are mentioning ADRs, find the spans of the ADRs, and normalize them with respect to a given knowledge base (KB). These tasks are especially challenging due to specific characteristics of user-generated texts from social networks which are noisy, containing misspelled words, abbreviations, emojis, etc.

Motivated by the recent success of deep architectures in general and language representation networks in particular, we explore an application of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and its extension for biomedical domain BioBERT (Lee et al., 2019) to the SMM4H 2019 tasks. For both ADR extraction and medical concept normalization, we conclude that BERT outperforms previous state-of-the-art baselines based on recurrent neural architectures (RNNs), including bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), and Gated Recurrent Units (Cho et al., 2014) paired with *word2vec* word embeddings.

The paper is organized as follows. In Section 2, we present the task description, machine learning baselines, and classification experiments for Task 1. We describe our models for end-to-end extraction of ADR concepts in Sections 3 and 4. Finally,

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 52–57
Florence, Italy, August 2, 2019.

we discuss future directions in Section 5.

2 Task 1: Classification

The goal of this sub-task is to identify the tweets with ADR mentions. This is a necessary filtering step to remove noise, since most of the health-related chatter in the domain does not contain relevant information.

2.1 Dataset

The training set consists of 25,678 tweets with 2,377 labeled as positive examples with ADRs; this statistic shows that the corpus has huge class imbalance. Tweet text lengths vary from 1 to 53 words, the average length is 20 words. The test dataset includes 4,576 tweets. Minimum tweet length is also 1 and the maximum consists of 186 words, which is much longer than in the training set. However, the average amount of words in tweets is on par with the training set and equals 23 words.

2.2 Method

Previous studies have shown the effectiveness of classical machine learning approaches (Ofoghi et al., 2016; Jonnagaddala et al., 2016; Kiritchenko et al., 2018; Alimova and Tutubalina, 2017). We applied the SVM-based model with a set of features as a baseline method. For SVM features, we utilized the bag-of-words representation, drug name, and ADRs from a Diego Lab ADR lexicon (Sarker et al., 2015). The list of drug names was obtained from the Food and Drug Administration (FDA). We've also explored the potential of *sent2vec* tool for tweets representation (Pagliardini et al., 2018). The Twitter unigram pre-trained model was applied for obtaining vectors[1].

Our main solution is a classifier based on the BERT architecture. For the BERT-based model, the tweet's representation was obtained with the *Transformer* architecture (Vaswani et al., 2017), and then logistic regression was used as a classifier. We used the implementation from the model's official repository[2].

2.3 Experiments

For the SVM-based classifier, we set class weights to 0.3 and 0.7 for non-ADR and ADR classes respectively and applied a linear kernel. The BERT-based model was trained on 20 epochs with learning rate equal to $5 * 10^{-5}$, maximum sequence size 128, and batch size 32.

Run name	F1	P	R
KFU NLP, BERT	57.38	69.14	49.04
KFU NLP, SVM	51.64	56.2	47.76
Average scores	50.19	53.51	50.54

Table 1: Text classification results on the Task 1 test set.

The official evaluation metrics are precision (P), recall (R), and F1-measure (F1) computed for the positive class. During preprocessing, we removed all URLs, user mentions, and symbols of re-tweets using the *tweet-preprocessor* package[3]. We conducted a set of experiments on the training set with 5-fold cross-validation. Results of these experiments shows that utilizing *sent2vec* as tweet representations did not improve classification quality. Results on the test set are presented in Table 1. Our baseline SVM classifier (run-2) obtained the F1 score of 51.64%, which is on par with average results. The BERT-based classifier (run-1) achieved the F1 score of 57.38 and outperformed by 7.19% the F1 score averaged across 43 submissions.

3 Task 2: Extraction of Adverse Effect Mentions

Following state-of-the-art research (Miftahutdinov et al., 2017; Tutubalina and Nikolenko, 2017; Lee et al., 2019), we view the second task from the perspective of a sequence labeling problem. Sequence labeling refers to the task of learning to predict a label for each token in a sequence of tokens. State-of-the-art methods employ neural architectures based on bidirectional LSTMs and conditional random fields (CRF) (Lample et al., 2016; Tutubalina and Nikolenko, 2017; Giorgi and Bader, 2019). Recent advancements in language representation models such as BERT have opened up new directions of research in sequence labeling.

3.1 Dataset

The data for the second sub-task includes 2,367 tweets that are fully annotated for ADR mentions and Indications. This set contains a subset of (i) 1,212 tweets from Task 1 tagged as 'hasADR' and

[1] https://github.com/epfml/sent2vec

[2] https://github.com/google-research/bert

[3] https://pypi.org/project/tweet-preprocessor/

(ii) 1,155 tweets marked as 'noADR' (1,828 ADR mentions in total).

3.2 Method

Sequence labeling methods view a message as a sequence of tokens labeled using the BIO tagging scheme: B indicates the beginning of the entity mention, I is used for tokens inside the entity mention, and O indicates tokens outside any entities. To solve the sequence labeling task, we utilize and empirically compare several models: (i) bidirectional LSTM-CRF; (ii) BERT; (iii) BERT for Biomedical Text Mining named BioBERT. We have also utilized a CRF tagger on top of BioBERT. A technical explanation of these neural models is omitted due to space constraints; we refer to the studies listed above.

We have also combined deep neural network representations with additional dictionary-based features. 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 extensional semantic information of an entity mention. We adopted the dictionaries from our previous work (Miftahutdinov et al., 2017).

3.3 Experiments

For the NER sub-task each network was trained for 25 epochs with batch size set to 32. We used the Adam algorithm as the optimizer with initial learning rate $5 * 10^{-5}$. We used the publicly available implementation of BioBERT-CRF[4]. Training all 10 networks took 2-3 hours on eight NVIDIA Tesla P40 GPUs. Additionally, we have used the CADEC corpus along with the corpus provided by the organizers.

Since the boundaries of an entity mention in social media texts are hard to define, two types of evaluation were used: *strict* and *relaxed*. Precision, recall, and F-measure are used for performance evaluation.

In order to select the best neural models, we evaluated our models on the CADEC corpus using 5-fold cross-validation at the development stage. BERT showed 5-7% improvement in the strict evaluation over LSTM-CRF, while BioBERT showed slightly better performance over BERT. BioBERT with CRF stayed roughly on par with the model without CRF.

During BioBERT evaluation, we encountered unstable results on development sets. Therefore, for the final submission we combined the results of ten BioBERT-CRF with the same settings using a simple voting scheme with the intent of increasing the robustness of the final system. Table 2 shows a comparison of the ensemble model to the official average scores computed using the participants' submissions. Our model has obtained the highest relaxed F1 score of 65.8% among 9 teams.

Run name	F1	P	R
Relaxed Evaluation			
KFU NLP Team	65.8	55.4	81.0
Average scores	53.83	51.29	61.74
Strict Evaluation			
KFU NLP Team	46.4	38.9	57.6
Average scores	31.69	30.26	35.81

Table 2: The NER results on the Task 2 test set.

4 Task 3: Medical Concept Normalization

A crucial part of this problem is to translate a text from *social media language* (e.g., "felt sick to my stomach" or "couldn't sleep much") to *formal medical language* (e.g., "nausea" and "insomnia", respectively).

The SMM4H 2019 Task 3 is designed as an end-to-end task. This setup is closer to a real production environment, where the system has free-form text as input and should be able to produce a set of extracted medical concepts. This end-to-end setup is more challenging due to the sequential two-stage pipeline: the system has to (i) first detect ADR mentions and then (ii) map extracted ADRs to knowledge base entries. For the first step, we use the NER model described in Section 3. The system used for concept normalization is based on our previous works (Tutubalina et al., 2018; Miftahutdinov and Tutubalina, 2019) and presented below.

4.1 Dataset

ADR mentions from the SMM4H 2019 dataset are mapped to Preferred Terms (PTs) of the Medical

[4] https://github.com/dmis-lab/biobert

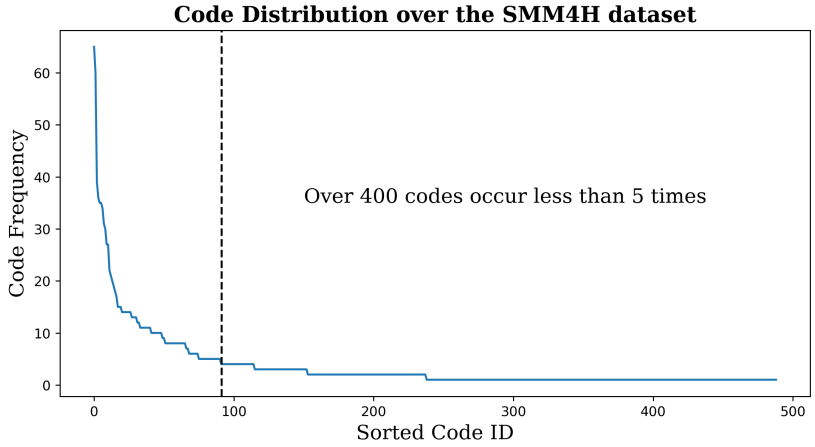

Figure 1: The code frequency distribution of MedDRA codes in the training set (Task 3).

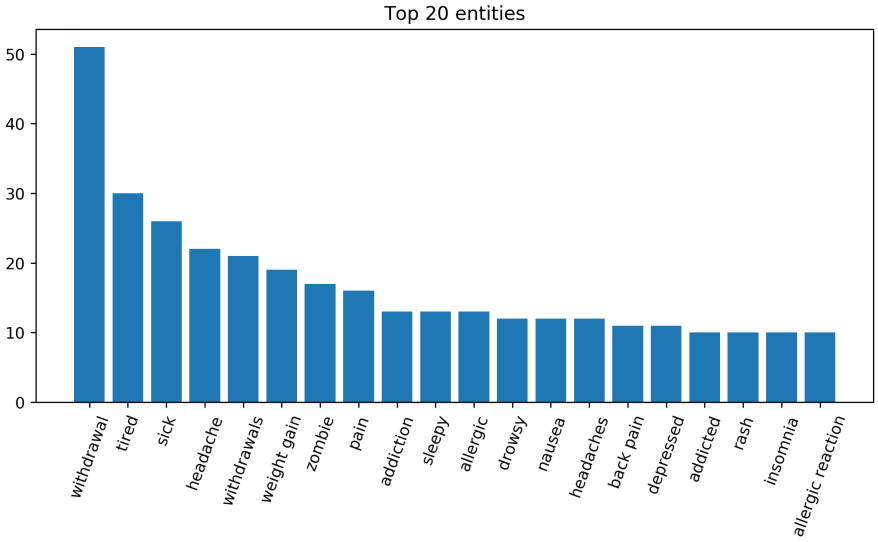

Figure 2: Top 20 entities in the training set (Task 3).

Dictionary for Regulatory Activities (MedDRA). The training SMM4H 2019 set consists of 1,828 phrases mapped to 489 MedDRA codes. The average number of ADR mentions mapped to a given concept is 3.74. The minimum and maximum numbers of queries mapped to a given concept are 1 and 65, respectively. Figure 1 shows a plot of the code frequency distribution of MedDRA concepts presented in the training set. Additionally, we present statistics on the top 20 entity mentions from the training set in Figure 2.

4.2 Method

Following state-of-the-art research (Tutubalina et al., 2018; Sarker et al., 2018; Miftahutdinov and Tutubalina, 2019), we view concept normalization as a classification task. Following (Miftahutdinov and Tutubalina, 2019), we convert each ADR mention into a vector representation using BERT or RNN. Next, we employ the standard softmax activation for the output layer. The softmax layer over all possible medical codes from the training set yields a probability for the sequence.

In order to train the classification model, we utilized training sets from five different sources: SMM4H 2019 dataset, SMM4H 2017 dataset (Sarker et al., 2018), CADEC dataset (Karimi et al., 2015), PsyTAR dataset (Zolnoori et al., 2019), and TwADR-L (Limsopatham and Collier, 2016). SMM4H datasets and CADEC were manually mapped to MedDRA codes. PsyTAR and TwADR-L were mapped to the MedDRA coding system using the UMLS metathesaurus (version 2017AA).

Run name	F1	P	R
Relaxed Evaluation			
KFU NLP Team	43.2	36.2	53.5
Average scores	29.72	29.06	31.15
Strict Evaluation			
KFU NLP Team	34.4	28.8	42.7
Average scores	21.18	20.53	22.41

Table 3: The concept normalization results on the Task 3 test set.

4.3 Experiments

We trained the BERT model for 40 epochs, using batch size 96 and learning rate $5*10^{-5}$. In order to prevent neural networks from overfitting, we used a dropout of 0.2 to control the inputs and the softmax layer. We used the publicly available implementation of BERT[5].

The strict and relaxed evaluations proposed for Task 2 were also adopted for Task 3. As in previous work, we evaluated our models on the CADEC corpus at the development stage using 5-fold cross-validation. The BERT model consistently outperformed attention-based bidirectional LSTM and GRU paired with pre-trained word embeddings in this set of experiments, showing a 6-9% improvement. We did not experiment with BioBERT for this task.

For the final submission, we used the two-stage pipeline based on the ensemble of BioBERT-CRF for NER and BERT for normalization. Table 3 shows a comparison of our best model to the official average scores computed using the participants' submissions. The end-to-end model ranked first at SMM4H 2019 Task 3 and obtained a relaxed F1 of 43.2%. The strict recall of the end-to-end system is 15% lower than the recall of the NER system: 42.7 vs 57.6. Results in Tables 2 and 3 indicate that more than 80% of extracted ADR mentions have been correctly mapped to MedDRA concepts.

5 Conclusion

In this work, we have explored an application of Bidirectional Encoder Representations from Transformers (BERT) to the task of text classification, extraction of adverse drug reactions, and concept normalization. We have evaluated BERT and BioBERT empirically against bidirectional LSTM and GRU. Experiments have shown that BERT outperforms LSTM and GRU on all three tasks, achieving new state-of-the-art results in ADR extraction and normalization.

We foresee three directions for future work. One potential direction is to investigate neural architectures including BERT and RNNs in the end-to-end setup on other existing corpora. Another future direction is to explore how to effectively use of contextual information to map entity mentions to medical concepts. Additionally, the effect of data imbalance can be investigated for BERT-based models.

[5] https://github.com/huggingface/pytorch-pretrained-BERT

Acknowledgments

We thank Sergey Nikolenko for helpful discussions. This research was supported by the Russian Science Foundation grant no. 18-11-00284.

References

Ilseyar Alimova and Elena Tutubalina. 2017. Automated detection of adverse drug reactions from social media posts with machine learning. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 3–15. Springer.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John Giorgi and Gary Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *BioRxiv*, page 526244.

S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. Based on TR FKI-207-95, TUM (1995).

Jitendra Jonnagaddala, Toni Rose Jue, and Hong-Jie Dai. 2016. Binary classification of twitter posts for adverse drug reactions. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, HI, USA*, pages 4–8.

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.

Svetlana Kiritchenko, Saif M Mohammad, Jason Morin, and Berry de Bruijn. 2018. Nrc-canada at

smm4h shared task: classifying tweets mentioning adverse drug reactions and medication intake. *arXiv preprint arXiv:1805.04558*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Nut Limsopatham and Nigel Collier. 2016. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *ACL*.

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 and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of ACL 2019, Student Research Workshop*, Florence, Italy. Association for Computational Linguistics.

Bahadorreza Ofoghi, Samin Siddiqui, and Karin Verspoor. 2016. Read-biomed-ss: Adverse drug reaction classification of microblogs using emotional and conceptual enrichment. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.

E. Tutubalina and S. Nikolenko. 2017. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *Journal of Healthcare Engineering*, 2017.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth social media mining for health (smm4h) shared task at acl 2019. In *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics.

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.