Data Augmentation for Rare Symptoms in Vaccine Side-Effect Detection

Bosung Kim and **Ndapa Nakashole**
Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
bosungkim@ucsd.edu, nnakashole@eng.ucsd.edu

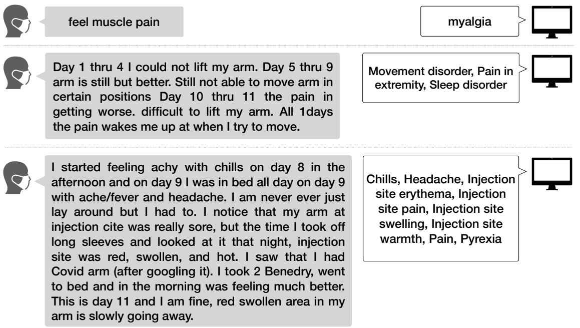
Abstract

We study the problem of entity detection and normalization applied to patient self-reports of symptoms that arise as side-effects of vaccines. Our application domain presents unique challenges that render traditional classification methods ineffective: the number of entity types is large; and many symptoms are rare, resulting in a long-tail distribution of training examples per entity type. We tackle these challenges with an autoregressive model that generates standardized names of symptoms. We introduce a data augmentation technique to increase the number of training examples for rare symptoms. Experiments on real-life patient vaccine symptom self-reports show that our approach outperforms strong baselines, and that additional examples improve performance on the long-tail entities.

1 Introduction

Motivation. Outside of clinical trials of vaccines on a small part of the population, it is important to study symptoms that arise as side effects of vaccines in the broader population. This is particularly crucial when the vaccines have only been granted emergency use permission, as has been the case for the COVID-19 vaccines such as the Pfizer-BioNTech mRNA vaccine, the Oxford-AstraZeneca adenovirus-vectored vaccine, and others. In the United States, the Vaccine Adverse Event Reporting System (VAERS)[1], co-managed by the Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA), is a national system that collects and analyzes reports from patients, about possible side effects after taking a vaccine.

VAERS presents a rich source of data for researchers to analyze. A challenge that arises when trying to analyze patient self-reports such as those in VAERS is that patients are free to use their choice of words to describe the side-effects they have experienced. This necessitates data normalization so that across different patient reports, even in the face of polysemy, abbreviations, spelling errors, or other variations, the same symptom is mapped to the same name. Thus, in this paper, we study entity detection and normalization on the VAERS dataset. The task we are addressing is illustrated with sample reports from VAERS in Figure 1.



feel muscle pain

myalgia

Day 1 thru 4 I could not lift my arm. Day 5 thru 9 arm is still but better. Still not able to move arm in certain positions Day 10 thru 11 the pain in getting worse. difficult to lift my arm. All 1days the pain wakes me up at when I try to move.

Movement disorder, Pain in extremity, Sleep disorder

I started feeling achy with chills on day 8 in the afternoon and on day 9 I was in bed all day on day 9 with ache/fever and headache. I am never ever just lay around but I had to. I notice that my arm at injection cite was really sore, but the time I took off long sleeves and looked at it that night, injection site was red, swollen, and hot. I saw that I had Covid arm (after googling it). I took 2 Benedry, went to bed and in the morning was feeling much better. This is day 11 and I am fine, red swollen area in my arm is slowly going away.

Chills, Headache, Injection site erythema, Injection site pain, Injection site swelling, Injection site warmth, Pain, Pyrexia

Figure 1: Examples of patient self-reports from the VAERS, and their corresponding symptom entities.

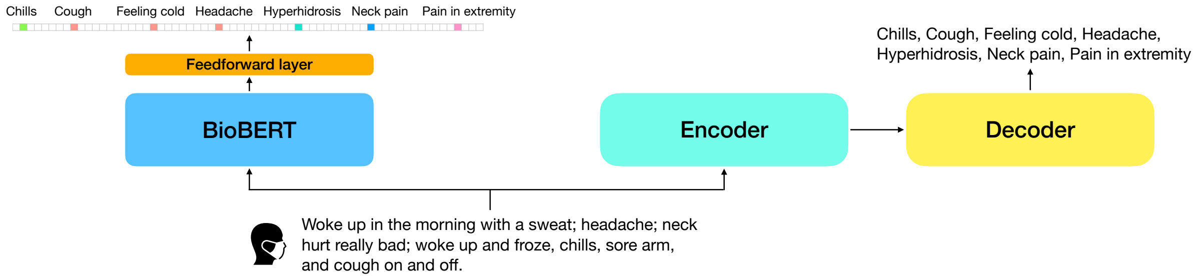
Currently, VAERS self-reports are manually tagged with standardized names of symptoms that are mentioned in them — a time consuming, and imperfect process as our inspection showed cases where not all symptoms were tagged. Automated models could support human effort to speed up the process, and potentially suggest entities a human might miss.

Challenges. Our application setting presents unique challenges : 1) entity names can be long and contain a lot of common nouns; 2) the number of entity types is large; 3) the number of labels in each example varies widely, e.g., patient reports contain anywhere from a minimum of 1 to a maximum of 131 symptoms; and 4) while a few symptoms are common, many are rare, resulting in a long-tail distribution of labels per entity type.

Contributions. To tackle these challenges, we frame the problem as an entity retrieval (ER)

[1] https://vaers.hhs.gov/data.html

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 310–315
May 26, 2022. ©2022 Association for Computational Linguistics


Chills
Cough
Feeling cold
Headache
Hyperhidrosis
Neck pain
Pain in extremity
Feedforward layer
BioBERT
Encoder
Decoder
Chills, Cough, Feeling cold, Headache, Hyperhidrosis, Neck pain, Pain in extremity
Woke up in the morning with a sweat; headache; neck hurt really bad; woke up and froze, chills, sore arm, and cough on and off.

(a) Multi-label classification model

(b) Autoregressive entity retrieval model

Figure 2: Architecture of a multi-label classification approach (a) and an autoregressive entity retrieval approach (b). The characteristics of our domain render classification approaches ineffective.

task. We leverage an autoregressive entity retrieval model (Cao et al., 2021) that generates standardized names of symptoms from patient self-reports, as opposed to a classification model such as a pre-trained model (Devlin et al., 2019) or BioBERT (Lee et al., 2019) fine-tuned with a classification layer on top. To tackle data sparsity problems of rare symptoms, we propose a data augmentation method that generates training data points through the definition of symptoms. We then obtain symptom definitions in two ways: i) Pre-trained language models: it has been shown that pre-trained language models are good at generating definitions (Shwartz et al., 2020), we therefore use GPT-3 (Brown et al., 2020) to generate symptom definitions. and ii) UMLS: for additional definitions, we consult a medical knowledge graph, the Unified Medical Language System (UMLS) (Bodenreider, 2004). UMLS is the largest and most authoritative knowledge graph of the biomedical domain with over 3 million entities.

Our experiments on the VAERS dataset show that our approach outperforms strong baselines, and that additional examples improve performance on long-tail entities.

2 Autoregressive Entity Retrieval Model

The goal of symptom entity detection is to predict symptom entities $\mathcal{E} = \{e_1, ..., e_n\}$ corresponding to the input description x. Each example is a pair of $(x, \mathcal{E})$ and the number of entities n varies over the dataset. As shown in Figure 2 (a), multi-label classification approaches are trained to minimize cross entropy loss over all symptom classes. In the autoregressive entity retrieval, Figure 2 (b), the model generates a sequence of symptom names as a target sentence instead of classifying each entity class. We adopt GENRE's (Cao et al., 2021) architecture that consists of transformer-based encoder and decoder. However, to retrieve multiple symptoms, GENRE requires annotated spans that refer to each symptom. For example, the source and target sequences should be (*"I have muscle pain and fever", "I have [muscle pain] (Myalgia) and [fever] (Pyrexia)"*). In our setting, a key difference is that the VAERS dataset is not annotated with the mention spans of entities, only whether or not a particular symptom was mentioned by the patient. Therefore, we generate the target sequence as a comma separated list, i.e., the pair of source and target sequences is (*"I have muscle pain and fever", "Myalgia, Pyrexia"*). Then the model is trained to maximize the probability

$$P(y|x, \theta) = \prod_{i=1}^{|y|} p(y_i|y_0, ..., y_{i-1}, x, \theta) \qquad (1)$$

where $y = \{y_1, ..., y_m\}$ is a set of tokens in the target sentence, y_0 is a model specific start token, and θ is the parameters of the model.

3 Data Augmentation

While the data of common symptoms, such as Headache and Pyrexia, are abundant to train the model, examples of long-tail symptoms are rare, and therefore have fewer reported instances in the dataset. The median of the number of symptoms in our train set is 5 and over 80% of entities occur less than 50 times while Headache and Pyrexia have over 100K examples, see Figure 3.

To overcome the problem posed by this very skewed training data distribution, we propose to generate additional labeled data in the form of definitions. The idea is that we can treat a symptom definition as a synthetic patient report (input sequence), and the symptom name as the corresponding label. We obtained definitions of symptoms in

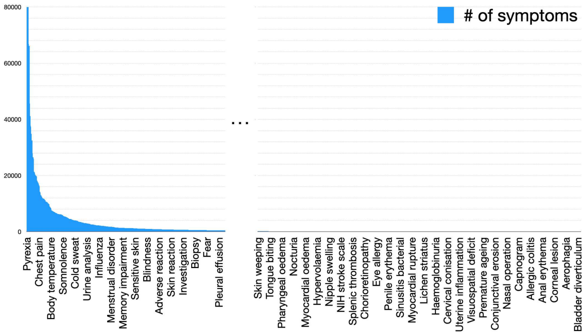


Figure 3: The distribution of symptom entities in the VAERS dataset has a very long tail.

two ways: using a pre-trained language model, and using the UMLS biomedical dictionary.

Pre-trained Language Model. We use GPT-3 (Brown et al., 2020) to generate definitions of long-tail symptoms. We use the prompt: "*The definition of [symptom name] is*". We then add the generated sentence as a synthetic patient report and the symptom name as a label, to our augmented data.

UMLS medical dictionary. For UMLS, we search terms with symptom names and then choose the first top result definition.

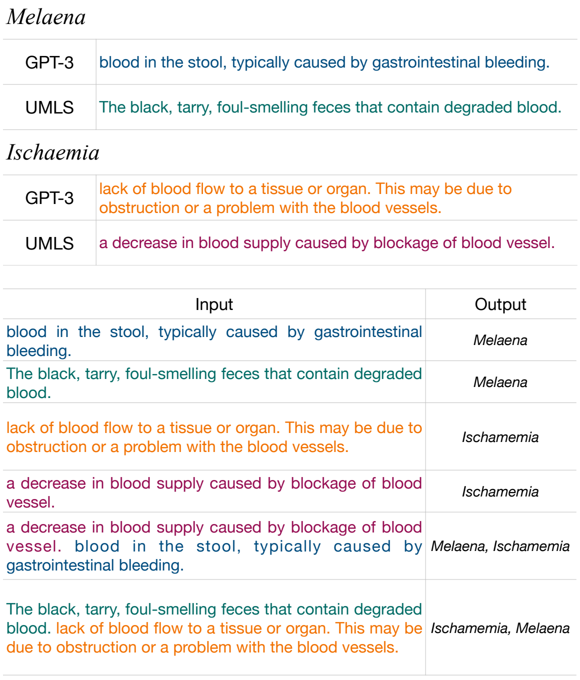
One limitation of this approach is that each symptom definition only corresponds to a single symptom whereas real patients often experience more than one symptom. To mimic the more realistic scenario of multiple symptoms, we also generate synthetic reports with up to two symptoms by concatenating definitions. Examples of such hallucinated data points are shown in Figure 4.

4 Experiments

Dataset. From VAERS, we consider data from the last three years (2019 to 2021), and randomly split it into train, validation, and test sets of 534, 516; 66, 814; and 66, 814 (80%/10%/10%).

Long-tail Symptoms. The VAERS dataset contains 10, 507 symptom entities. We define the symptoms with a frequency of less than 50 as long-tail entities. As a result, 8, 755 entities are classified as long-tail, which are 83.3% of the total entity set.

Data Augmentation. We obtained 10, 507 generated definitions from GPT-3 and 3, 480 definitions through the UMLS dictionary API. For the experiments, we used a single definition to mimic a patient with a single symptom, in addition, we created 50K combination examples of two definitions and two symptoms.



Melaena	
GPT-3	blood in the stool, typically caused by gastrointestinal bleeding.
UMLS	The black, tarry, foul-smelling feces that contain degraded blood.
Ischaemia	
GPT-3	lack of blood flow to a tissue or organ. This may be due to obstruction or a problem with the blood vessels.
UMLS	a decrease in blood supply caused by blockage of blood vessel.

Input	Output
blood in the stool, typically caused by gastrointestinal bleeding.	*Melaena*
The black, tarry, foul-smelling feces that contain degraded blood.	*Melaena*
lack of blood flow to a tissue or organ. This may be due to obstruction or a problem with the blood vessels.	*Ischamemia*
a decrease in blood supply caused by blockage of blood vessel.	*Ischamemia*
a decrease in blood supply caused by blockage of blood vessel. blood in the stool, typically caused by gastrointestinal bleeding.	*Melaena, Ischamemia*
The black, tarry, foul-smelling feces that contain degraded blood. lack of blood flow to a tissue or organ. This may be due to obstruction or a problem with the blood vessels.	*Ischamemia, Melaena*

Figure 4: Examples of symptom definitions and generated data for augmentation. To build examples with multiple symptoms, we combine two definitions as one input sentence.

Test sets. We evaluated our approach on three test sets:

- 1) **Full**: Full test set with 66, 814 examples.
- 2) **CUI-mapped**: Many symptoms in our dataset can be mapped to Concept Unique Identifiers (CUI) in UMLS. To compare with previous work that can detect UMLS CUIs, we built a test set with entities mapped to UMLS. 6, 564 out of 10,507 entities are mapped to UMLS by exact string match.
- 3) **Long-tail**: A set of test examples including only long-tail entities.

4.1 Experiments Setup

We adopted GENRE's (Cao et al., 2021) experimental settings with 256 of maximum input length, 128 of maximum output length, 64 of batch size, 2e-5 learning rate and 4 of beam search size. We used the pre-trained BART (Lewis et al., 2020) model and fine-tuned 5 epochs on our training set. In the experiments with BERT, BioBERT and BART, we followed a multi-label classification setting with a feed-forward layer on the top of pre-trained models

Model	Type	Test set	Macro			Micro		
			Precision	Recall	F1	Precision	Recall	F1
String match		Full	**1**	0.1684	0.2883	**1**	0.1849	0.3121
BERT-base (Devlin et al., 2019)	C	Full	0.1453	0.1497	0.1474	0.1453	0.1735	0.1581
BioBERT-base (Lee et al., 2019)	C	Full	0.1321	0.1663	0.1472	0.1382	0.1989	0.1631
BART-base (Lewis et al., 2020)	C	Full	0.1378	0.1695	0.1520	0.1378	0.1976	0.1624
GENRE (Cao et al., 2021)	G	Full	0.8305	**0.7688**	**0.7984**	0.8196	**0.7193**	**0.7662**
GENRE + UMLS + GPT-3	G	Full	0.8305	0.7682	0.7981	0.8189	0.7187	0.7655
MetaMap	C	CUI-mapped	0.1630	0.3232	0.2167	0.0671	0.3169	0.1108
BioBERT-base (Lee et al., 2019)	C	CUI-mapped	0.1453	0.1929	0.1657	0.1453	0.2665	0.1880
GENRE (Cao et al., 2021)	G	CUI-mapped	0.8273	**0.7857**	0.8060	0.8498	**0.7719**	**0.8090**
GENRE + UMLS + GPT-3	G	CUI-mapped	**0.8278**	0.7853	**0.8060**	**0.8502**	0.7712	0.8088
GENRE (Cao et al., 2021)	G	Long-tail	0.1662	0.1391	0.1515	0.7061	0.1229	0.2094
GENRE + UMLS	G	Long-tail	0.1833	0.1541	0.1674	0.6973	0.1381	0.2305
GENRE + GPT-3	G	Long-tail	0.1902	0.1604	0.1741	**0.7106**	0.1436	0.2389
GENRE + UMLS + GPT-3	G	Long-tail	**0.1955**	**0.1629**	**0.1777**	0.6861	**0.1473**	**0.2425**

Table 1: Results of symptom entity detection on the VAERS dataset. C (Classification) and G (Generation) denote the type of each model. The generative models are more effective. Our data augmentation with UMLS and GPT-3 improves upon the generative model, GENRE, on long tail entities (last three rows).

and also we trained 5 epochs for each. All hyperparameters are set on the best validation scores.

Baselines. 1) **String match**: String match refers to an approach that relies on exact same string matches with symptom entities.
2) **BERT/BioBERT/BART**: Pre-trained LMs with a multi-label classification setup.
3) **MetaMap** (Aronson and Lang, 2010): MetaMap is a medical entity detection model provided by the National Library of Medicine.[2] Given the input text, MetaMap returns entities mapped to UMLS with confidence scores. We experimented with thresholds {0.05, 0.1, 0.15, 0.2, 0.25, 0.3} and regarded entities as positives over the threshold. The threshold of 0.1 was determined on the best validation score.

4.2 Results

Table 1 shows the results of our experiments. In multi-label classification models, we observe that pre-trained LMs do not outperform even the simple string match algorithm; this is likely due to the challenges outlined in the Introduction. On the other hand, the generative methods significantly boosts the F1 score, achieving over 79.8% and 76.6% of Macro and Micro F1 scores. Similarly, compared to MetaMap, the proposed approach shows substantial gains across all metrics.

In the experiments on the long-tail test set, the models show low performances as we expected because long-tail entities are scarce in the training set. However, when we train the model with each augmented set, we find that our synthetic data can help improve performance. Augmenting with both UMLS and GPT-3 definitions increases scores by 2.62% and 3.31% in Macro and Micro F1 on the long-tail test set. However, augmentation does not change performance for common symptoms that already have sufficient training data, as seen on the Full and CUI-mapped test sets.

5 Related Work

In biomedical entity retrieval or entity linking, BERT-based models, such as BioBERT (Lee et al., 2019) or EnRuDR-BERT (Tutubalina et al., 2020), are often used to classify or re-rank candidate entities (Ujiie et al., 2021; Angell et al., 2021; Sung et al., 2020; Sakhovskiy et al., 2021). In contrast to previous work, we took a generative approach. The Social Media Mining for Health Applications (SMM4H) Workshop (Magge et al., 2021) has introduced various shared tasks including normalization of adverse drug effects (Miftahutdinov et al., 2020) and detection of disease mentions in social media.

Approaches to overcome the problem of data sparsity and long-tail training data distributions include: data sampling (Li et al., 2019; Akhbardeh et al., 2021), cost-sensitive loss function (Lin et al., 2018), regularization (Kim et al., 2022), semi-supervised learning (Hangya et al., 2018), and word/sentence level attention mechanism (Qing et al., 2019).

The success of the few-shot generation demonstrated by GPT-3 (Brown et al., 2020) has resulted

[2] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/run-locally/MetaMap.html

in several studies that leverage GPT-3 for this purpose (Gao et al., 2021; Schick and Schütze, 2020). Kim et al. (2021) explores ways of leveraging external resources such as dictionaries or medical documents. We use both a language model, in addition to a dictionary whose coverage is limited.

6 Conclusion

We studied the problem of vaccine side-effect detection on real-world patient data. The characteristics of this domain render traditional classification approaches ineffective. Our experiments demonstrated that combining a generative approach with synthetic data from symptom definitions obtained from a pre-trained LM and a medical dictionary can help improve performance on rare symptoms. Exploring other approaches for learning with limited data, is an avenue for future work.

References

Farhad Akhbardeh, Cecilia Ovesdotter Alm, Marcos Zampieri, and Travis Desell. 2021. Handling extreme class imbalance in technical logbook datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4034–4045, Online. Association for Computational Linguistics.

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based inference for biomedical entity linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608, Online. Association for Computational Linguistics.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze. 2018. Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–820, Melbourne, Australia. Association for Computational Linguistics.

Bosung Kim, Hyewon Choi, Haeun Yu, and Youngjoong Ko. 2021. *Query Reformulation for Descriptive Queries of Jargon Words Using a Knowledge Graph Based on a Dictionary*, page 854–862. Association for Computing Machinery, New York, NY, USA.

Bosung Kim, Youngjoong Ko, and Jungyun Seo. 2022. Novel regularization method for the class imbalance problem. *Expert Systems with Applications*, 188:115974.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jia Li, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Sampling matters! an empirical study of negative sampling strategies for learning of matching models in retrieval-based dialogue systems. 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 1291–1296, Hong Kong, China. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (#SMM4H) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. KFU NLP team at SMM4H 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet*, 11(12).

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 39–43, Mexico City, Mexico. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few-shot text classification and natural language inference. *CoRR*, abs/2001.07676.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4615–4629. Association for Computational Linguistics.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

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*. Btaa675.

Shogo Ujiie, Hayate Iso, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. End-to-end biomedical entity linking with span-based dictionary matching. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 162–167, Online. Association for Computational Linguistics.