Symptom Extraction and Linking from COVID-19 Related Vaccine Adverse Event Reports

Matthew Hullstrung

1. Introduction

This project focuses on the automated extraction and linking of symptoms from Vaccine Adverse Event Reports (VAERS). VAERS serves as a crucial early warning system co-managed by the Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA), monitoring potential safety concerns following vaccine administration. I will be specifically focusing on COVID-19 related reports. My primary goal is to develop a system that can automatically identify symptoms within the narrative text of VAERS reports and link them to standard symptom terms using named entity recognition packages and entity linking methods.

The task is challenging due to the varying nature of VAERS reports and the absence of ground truth annotations. This project addresses these challenges through innovative natural language processing techniques to enhance vaccine safety monitoring.

At this time, during the midterm report, I have made substantial progress. I have adhered to the outlined schedule and accomplished key milestones. Specifically, I have successfully downloaded and processed the data and implemented Named Entity Recognition (NER) techniques to extract symptom-related entities. This achievement marks a significant step towards the ultimate goal of establishing a robust system for symptom extraction and linking.

These accomplishments did not come without obstacles. Importing the data posed a challenge due to unconventional formatting. To address this, certain unnecessary fields were excluded to ensure a smooth and accurate import process. This decision streamlined data import, overcoming issues associated with irregular data formatting. All other minor challenges encountered during data processing and NER have been addressed efficiently, contributing to the overall progress of the project. Remaining challenges are outlined with proposed solutions, ensuring a proactive approach to potential obstacles.

2. Problem Formulation

Formally, this project is a sequence labeling task. Given VAERS reports as input, the objective is to extract and link symptom-related entities within the text. The task can be defined as follows:

Step 1:

Input: Descriptions of COVID-19 related vaccine adverse events (SYMPTOM TEXT in VAERS DATA table).

Output: A list of symptom-related entities

Task Type: Named Entity Recognition

Step 2:

Input: A symptom detected in STEP 1

Output: Its standard symptom

Task type: Entity Linking

3. Methods

In this project, I will employ a combination of methods and tools to achieve the defined task:

Data Collection: VAERS dataset, focusing on the COVID-19 vaccine. Importing the data presented a hurdle due to unconventional formatting in the .csv data file. By excluding unnecessary fields like hospital visit dates, this challenge was addressed, resulting in a streamlined data import process. This not only resolved the formatting issue but also optimized memory usage for program execution.

Data Preprocessing: Selection of a representative dataset, statistical analysis, and vocabulary creation. We consider only COVID-19 vaccine data from 2022, sampling 10,000 reports. Statistical analysis consists of getting distributions of the counts of different symptoms related to the COVID-19 vaccine as well as getting the distribution of the length of the different reports. This analysis will help us better understand the data we will work with.

Symptom-Related Entity Extraction: Utilizing the Stanza package and biomedical/clinical NER models. To accomplish NER, we use a pipeline with the MIMIC III syntactic analysis model with the i2b2 clinical NER processor. By considering items labeled PROBLEM as symptoms, we successfully accomplish our task of symptom-related entity extraction.

Entity Linking: Implementation of rule-based and similarity-based methods. This has yet to be implemented but is next in the project timeline. I will experiment with several different methods of linking, including exact and fuzzy rule-based matching and similarity-based matching using GloVe embeddings or more task-specific clinical word embeddings.

Evaluation: Systematic evaluation, including manual validation and metrics such as precision, recall, and F1-score.

4. Dataset and Experiments

The project will primarily use the VAERS dataset, focusing on 2022 COVID-19 vaccine-related reports. Data preprocessing includes subset selection (10,000 samples) and standard symptom

list creation. Experiments involve Named Entity Recognition (NER) and Entity Linking (EL) tasks, evaluated using precision, recall, and F1-score metrics.

To analyze the statistics of the dataset, I plotted distributions of the standard symptom counts and report lengths. Each of these plots can be seen as *Figure 1* and *Figure 2* below (respectively):

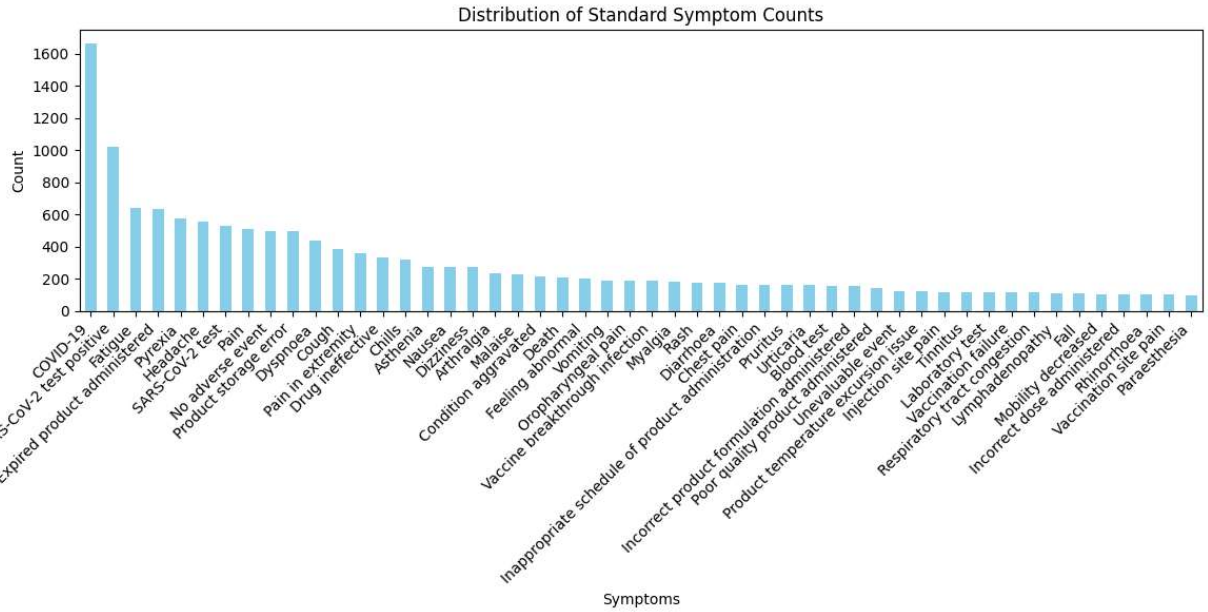

Figure 1

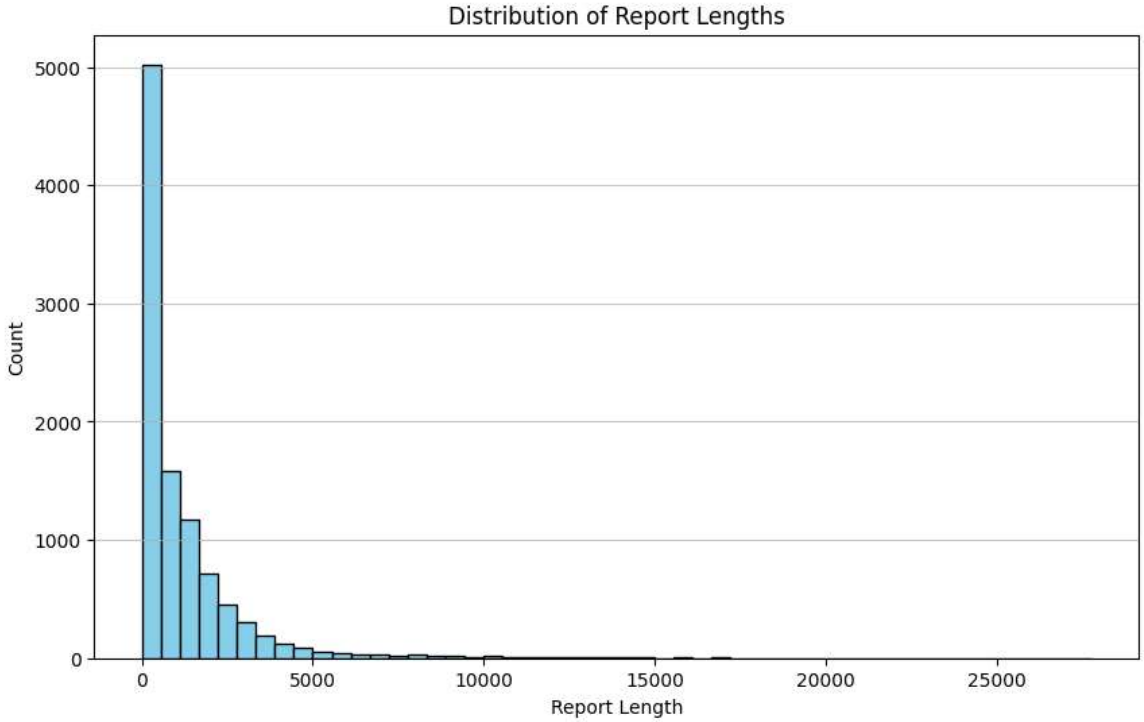

Figure 2

As shown in *Figure 1*, the top 5 standard symptoms from the reports were COVID-19, a positive SARS-CoV-2 test, fatigue, and an expired product administered.

To help justify the distribution given in *Figure 2*, I also performed some numerical statistics of the report lengths, including mean, standard deviation, minimum, maximum, and the number of missing reports.

There are 10 missing reports, meaning the count has been reduced to 9990. The mean of the report lengths is 1175.03 with a standard deviation of 1768.18. The minimum report length is 3 while the maximum report length is 27774. However, do note that these lengths are in terms of number of characters, not words. I may add additional statistics in the following weeks for the number of words in the reports.

I have also accomplished Named-Entity Recognition using a pipeline consisting of the MIMIC III syntactic analysis model using the i2b2 clinical processor for NER.

An example input and output of a single sample's Named-Entity Recognition is shown below:

Input: Pt was admitted to hospital with a positive COVID test on 1/21 and transferred to another hospital on 1/25. Prior to arrival the pt had 6 days of worsening cough, fever, extreme fatigue, and worsening shortness of breath.

Output Symptoms (PROBLEM): ['a positive COVID test', 'worsening cough', 'fever', 'extreme fatigue', 'worsening shortness of breath']

I have manually checked over 20 samples, but as seen in this single example, our pipeline does a very good job at extracting symptom-related entities. This task has been successfully accomplished.

Entity Linking (EL) tasks is the next step on the project timeline, which will be evaluated using precision, recall, and F1-score metrics.

5. Project Management

Matthew Hullstrung: Project Lead and Data Scientist

Responsibilities: Coordination, data collection, model implementation, and project oversight.

As the sole member of this project, I am responsible for all aspects of its execution. Here's an abbreviated (and tentative) project timeline:

Weeks 1-2: Data collection and preprocessing, including dataset selection.

Week 3: Implementation of entity extraction (NER).

Week 4: Midterm Report.

Weeks 5-6: Development of entity linking methods. Manual evaluation and model validation.

Week 7: Final evaluation, results analysis, and system refinement.

End of Semester: Project completion and submission.

6. Conclusion

This project aims to enhance vaccine safety monitoring by automating symptom extraction and linking from VAERS reports. By leveraging named entity recognition and entity linking techniques, I strive to contribute to public health efforts, ensuring timely responses to adverse events and data-driven insights into COVID-19 vaccine safety.

7. Key references

1. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations 2020 Jul (pp. 101-108).
2. Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, Curtis P Langlotz, Biomedical and clinical English model packages for the Stanza Python NLP library, Journal of the American Medical Informatics Association, Volume 28, Issue 9, September 2021, Pages 1892–1899, https://doi.org/10.1093/jamia/ocab090
3. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. The 7th IEEE International Conference on Healthcare Informatics. 2019.
4. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23.