Symptom Extraction and Linking from Vaccine Adverse Event Reports

Thanapoom Phatthanaphan
Stevens Institute of Technology
tphattha@stevens.edu

Abstract

The Vaccine Adverse Events Reporting System (VAERS) serves as a critical resource for monitoring the safety of vaccines. However, the vast amount of unstructured narrative text within VAERS reports poses a challenge for efficient analysis and symptom identification. In this paper, we present a novel approach to automatically extract and link vaccine-related symptoms from VAERS reports using sequence labeling techniques. Our method combines the power of named entity recognition (NER) and named entity linking (NEL) to achieve comprehensive and standardized symptom identification. We utilize NER packages and develop custom NEL methods to map identified terms to standard terminology from a symptom dictionary. By doing so, we aim to enhance the efficiency and accuracy of vaccine adverse event monitoring, ultimately contributing to the safety assessment of vaccines.

Introduction

Vaccines are crucial for public health, and to ensure their safety, the Vaccine Adverse Events Reporting System (VAERS) was jointly established by the CDC and FDA. VAERS collects reports of adverse events following vaccination, allowing continuous safety monitoring.
However, VAERS faces a challenge due to the high volume of unstructured narrative reports it receives each year. These reports contain valuable information about adverse events, including symptoms. Identifying these symptoms promptly is essential for taking action and maintaining public trust in vaccination programs.
In recent years, sequence labeling techniques, used in linguistics and computational linguistics, have proven valuable for tasks like identifying named entities. Our project introduces an innovative method to automatically extract and connect vaccine-related symptoms from VAERS reports. Our project focuses on two main goals: identifying symptom-related terms within the narrative text reports using NER tools and linking these terms to standardized symptom terminology. This approach addresses the challenge of unstructured VAERS reports while ensuring alignment with recognized medical terms.

This paper explains our methodology, including NER tool selection, symptom dictionary creation, and linking techniques. We also present experimental results to demonstrate the effectiveness of our approach. Ultimately, our project enhances vaccine adverse event monitoring, strengthens public health, and ensures vaccine safety.

Problem formulation

The core machine learning task in this project is formally defined as a **multi-class classification** problem. This task involves classifying text descriptions of vaccine adverse events, specifically the SYMPTOM TEXT in the VAERS DATA table, into predefined categories representing various symptoms or symptom-related entities. The formulation can be represented as follows:

- **Input**: Text descriptions of vaccine adverse events (SYMPTOM TEXT).
- **Output**: Class labels representing symptom-related entities or symptoms.
- **Classes:** Multiple predefined classes representing different symptoms or symptom-related entities.

Methods

Several methodologies, techniques, and tools will be employed. These approaches aim to automatically identify symptoms from VAERS reports and link them to standard terms in a dictionary. Here's an outline of the key methods and tools required for this project:

Named Entity Recognition (NER) Packages

Utilization of specialized NER packages tailored for biomedical and clinical text analysis, including Stanza, i2b2, and similar libraries, to identify symptom-related entities.

Named Entity Linking Methods

Development of Named Entity Linking (NEL) methods to establish connections between identified symptom entities and standard terms within a dedicated symptom dictionary.

Rule-Based Matching

Development of rule-based matching algorithms, including exact matching and fuzzy matching, to establish direct mappings of symptoms to standard terms through predefined rules.

Similarity-Based Matching

Investigation of similarity-based matching techniques employing word embeddings (e.g., GloVe or clinical word embeddings) to optimize entity linking accuracy.

Datasets and Experiments

The Vaccine Adverse Event Reporting System (VAERS) is a national early warning system to detect possible safety problems in U.S.-licensed vaccines. VAERS is co-managed by the Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA). VAERS accepts and analyzes reports of adverse events (possible side effects) after a person has received a vaccination.
There are three tables in the VAERS dataset that will be used in this project, including VAERS Data, VAERS Symptoms, and VAERS Vaccine.

- **VAERS Data**: The dataset contains detailed information about individual VAER submitted to the VAERS system.
- **VAERS Symptoms**: The dataset contains a list of symptoms and their corresponding codes that are reported in adverse event narratives.
- **VAERS Vaccine:** The dataset provides information about the vaccines including their characteristics and administration.

The datasets spanning from 2010 to the present day will serve as the foundation for creating new named entity recognition packages to identify terms related to symptoms and developing named entity linking methods to connect these identified terms to standardized terms present in a dictionary.
The evaluation of the machine learning model's performance is a critical aspect of this project. However, it's essential to note that there is no ground truth annotation available for the data, making evaluation challenging. To overcome this limitation, a combination of automatic and manual evaluations will be employed.

- **Automatic Evaluation**: The model's accuracy in correctly classifying symptoms will be assessed through automatic evaluation metrics.
- **Manual Evaluation**: To ensure the accuracy of the model's predictions, a sample of clinical notes (typically 20~50) will be selected for manual evaluation. Experts manually review the results, verifying the correctness of the model's predictions.

The three key metrics, including precision, recall, and F1-score, will serve as essential tools for assessing the model's performance in identifying and linking symptoms.

- **Precision**: Precision will measure the accuracy of identified symptoms, ensuring their trustworthiness and relevance
- **Recall**: Recall will determine the model's ability to capture all relevant symptoms, minimizing the risk of missing significant health indicators
- **F1-score**: F1-score will provide an overall evaluation of the model's effectiveness in both accuracy and completeness.

Project management

The timeline of the project is set as below:

- **Oct 10 – Oct 15**: Data Preprocessing.
- **Oct 16 – Oct 20**: Midterm Exams (Other courses).
- **Oct 21 – Oct 27**: Extracting Symptom-related Entities (Step 1).
- **Oct 28 – Nov 2**: Link Entities to Standard Symptoms (Step 2).
- **Nov 3 – Nov 6**: Reporting Midterm progress
- **Nov 7** – **Nov 12**: Continue to finish Step 2.
- **Nov 13** – **Nov 17**: Evaluating the system.
- **Nov 18** – **Nov 28**: Preparing presentation material.
- **Nov 29** – **Dec 3**: Preparing Final report
- **Dec 4 – Dec 11**: Presentation period
- **Dec 11**: Submitting Final Report

References

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, D.C. 2020. A Python Natural Language Processing Toolkit for Many Human Languages. The 58th Annual Meeting of the Association for Computational Linguistics. (pp. 101-108)

Zhang, Y., Zhang, Y., Qi, P., Manning, D.C., & Langlotz, P.C. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. Journal of the American Medical Informatics Association (Vol. 28, Issue 9). (pp. 1892-1899). https://doi.org/10.1093/jamia/ocab090