Symptom Extraction and Linking from 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. 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.

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 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.

Data Preprocessing: Selection of a representative dataset, statistical analysis, and vocabulary creation.

Entity Extraction: Utilizing the Stanza package and biomedical/clinical NER models.

Entity Linking: Implementation of rule-based and similarity-based methods.

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 COVID-19 vaccine-related reports. Data preprocessing includes subset selection and standard symptom list creation. Experiments involve Named Entity Recognition (NER) and Entity Linking (EL) tasks, 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: Development of entity linking methods.

Week 5: Manual evaluation and model validation.

Weeks 6-7: 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. Domainspecific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23.