Report On

NER and Entity-Linkage on Medical dataset

Submitted in partial fulfillment of the requirements of the Course project in
Semester VII of Fourth Year Artificial Intelligence and Data Science

by
Naveen Arora (Roll No. 01)
Shikha Chaudhary (Roll No. 03)
HirenKumar Vyas(Roll No. 32)
Devashree Pawar (Roll No. 21)

Supervisor
Dr. Tatwadarshi P. N.

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Artificial Intelligence and Data Science

(2023-24)

Vidyavardhini's College of Engineering & Technology
Department of Artificial Intelligence and Data Science

CERTIFICATE

This is to certify that the project entitled “Terrain Recognition using CNN” is a bonafide work of "Naveen Arora (Roll No. 01), Shikha Chaudhary (Roll No. 03),Chetan Jawale (Roll No. 05)" submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in semester VII of Fourth Year Artificial Intelligence and Data Science engineering.

Supervisor

Mr.Raunak Joshi

Dr. Tatwadarshi P. N.
Head of Department

Table of Contents

Pg. No

Abstract

Named Entity Recognition (NER) and Entity Linkage have become indispensable tools in the analysis of medical datasets, enabling the extraction and categorization of crucial information. In this project, we focus on implementing NER and Entity Linkage techniques on a comprehensive medical dataset to identify and link important entities such as diseases, symptoms, treatments, and medications. By leveraging state-of-the-art natural language processing (NLP) models and knowledge graphs, we aim to accurately recognize and link medical entities, facilitating a deeper understanding of the relationships between different medical concepts. The project not only contributes to the advancement of medical research and analysis but also holds the potential to enhance clinical decision-making, ultimately leading to improved patient care and outcomes.

Acknowledgments

We would like to express our sincere gratitude to our advisor Dr. Tatwadarshi P. N. for the continuous support of our study and research, for her patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this thesis. We could not have imagined having a better advisor and mentor for our study.

1. INTRODUCTION

1.1 INTRODUCTION

This project focuses on applying Named Entity Recognition (NER) and Entity Linkage techniques to a medical dataset. By leveraging advanced natural language processing (NLP) models, the aim is to accurately identify and link crucial medical entities, enabling a better understanding of complex relationships within the dataset. The outcomes of this project have the potential to significantly advance medical research and improve clinical decision-making, ultimately enhancing patient care and outcomes.

1.2 PROBLEM STATEMENTS & OBJECTIVES

Problem Statement:

The project addresses the challenge of effectively extracting and linking medical entities, such as diseases, symptoms, treatments, and medications, from unstructured medical datasets. The absence of an automated and accurate NER and Entity Linkage system impedes the efficient analysis and utilization of the wealth of information within medical texts. Consequently, there is a need for a robust system that can automatically recognize and link these entities to facilitate comprehensive medical analysis, aiding researchers and healthcare professionals in making informed decisions and advancing medical research and treatment.

Objectives:

The main objectives of this project are as follows:

1. Implement state-of-the-art NER techniques to accurately identify and extract medical entities, including diseases, symptoms, treatments, and medications, from unstructured medical datasets.

2. Develop an efficient Entity Linkage system that can establish meaningful connections between the extracted medical entities, enabling a comprehensive understanding of the relationships within the dataset.

3. Enhance the accuracy and robustness of the NER and Entity Linkage models by leveraging domain-specific knowledge and data augmentation techniques.

4. Create a user-friendly interface for healthcare professionals and researchers to easily input medical texts and obtain organized and linked information about various medical entities.

5. Evaluate the performance of the NER and Entity Linkage system using standard metrics, ensuring its effectiveness in accurately identifying and linking medical entities within diverse medical datasets.

6. Contribute to the advancement of medical research and clinical decision-making by providing an efficient and automated system for comprehensive medical entity extraction and linkage, facilitating deeper insights into complex medical relationships.

1.3 SCOPE

The scope of this project includes:

1. Data Collection: Gathering diverse and comprehensive medical datasets from reliable sources, encompassing various types of medical texts and documents.

2. Preprocessing: Cleaning and preparing the collected data for NER and Entity Linkage, including text normalization and formatting to ensure compatibility with the models.

3. NER Model Development: Implementing and fine-tuning advanced NLP models for accurate identification and extraction of medical entities from the preprocessed data.

4. Entity Linkage Model Development: Creating an efficient system to establish meaningful connections between the extracted medical entities, enabling the construction of a comprehensive knowledge graph.

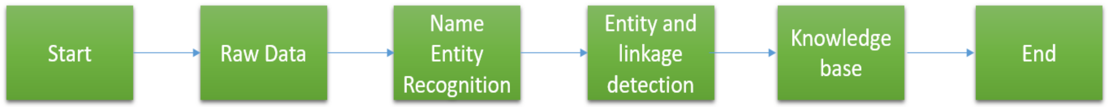
5. Integration and User Interface: Developing a user-friendly interface that allows healthcare professionals and researchers to easily input medical texts and visualize the linked medical entities for enhanced data analysis.

6. Evaluation and Validation: Assessing the performance of the NER and Entity Linkage models using standard evaluation metrics and validating their accuracy and robustness on different types of medical datasets.

The project does not include real-time deployment in clinical settings but focuses on the development of a prototype that demonstrates the potential for enhancing medical data analysis and research.

3. PROPOSED SYSTEM

3.1 Architecture/Framework/Block Diagram



3.2 Module Description

Certainly, here is a description of the key modules for the NER and Entity Linkage project:

1. Data Collection Module:
Responsible for gathering diverse and comprehensive medical datasets from various reliable sources, ensuring the collection of a wide range of medical texts and documents.

2. Data Preprocessing Module:
Handles the cleaning and formatting of the collected medical data, including tasks such as text normalization, tokenization, and removing noisy or irrelevant information.

3. NER Model Implementation Module:
In charge of implementing state-of-the-art NER models, such as Bidirectional LSTMs or Transformer-based models, to accurately identify and extract medical entities from the preprocessed data.

4. Entity Linkage Model Implementation Module:
Focuses on developing an efficient system for establishing connections between the extracted medical entities, constructing a comprehensive knowledge graph, and capturing the relationships between different entities.

5. User Interface Development Module:
Develops a user-friendly interface that enables easy input of medical texts and provides intuitive visualization of the linked medical entities, facilitating efficient data analysis and exploration.

6. Evaluation and Validation Module:
Conducts thorough evaluation and validation of the NER and Entity Linkage models, utilizing standard metrics such as precision, recall, and F1-score, to assess their accuracy and robustness on different types of medical datasets.

7. Documentation and Reporting Module:
Manages the documentation of the project, including detailed descriptions of the methodology, model architectures, data preprocessing techniques, and evaluation results, ensuring clear and comprehensive reporting for future reference and potential further development.

3.3 Details of Hardware and Software

Hardware Requirements:

- **Computing Device:** A personal computer or workstation with sufficient processing power and memory to handle data-intensive tasks is essential. Ideally, this computer should have a multi-core processor (e.g., Intel Core i5 or higher) and a minimum of 8 GB RAM to ensure smooth data processing and analysis.
- **Storage:** Adequate storage space is required for storing datasets and analysis results. A minimum of 250 GB of free hard disk space is recommended.
- **Graphics Processing Unit (GPU):** While not strictly necessary for EDA, a powerful GPU can significantly accelerate machine learning models if used for prediction. This is especially relevant if the analysis transitions to more advanced predictive modeling.

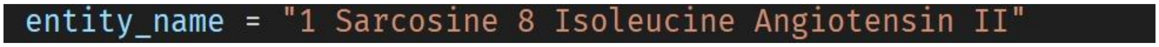
Software Requirements:

- **Operating System:** The choice of operating system is flexible, as most data analysis tools are cross-platform. Common options include Windows, macOS, and Linux.
- **Statistical Analysis Software**: A statistical analysis tool or integrated development environment (IDE) is necessary for performing data analysis and visualization. Some widely used software options include:
- **Python:** Python-based tools like Jupyter Notebooks, Anaconda, and libraries such as Pandas, NumPy, and Matplotlib are popular choices.
- **Statistical Software:** Proprietary tools like IBM SPSS, SAS, or Stata can be employed for more specialized statistical analysis.
- **Data Visualization Tools:** To create effective data visualizations, software like Tableau, Power BI, or open-source alternatives like Plotly, Seaborn, and ggplot2 can be used.

- **Machine Learning Libraries:** For customer loyalty prediction and more advanced analyses, machine learning libraries such as Scikit-Learn (Python) or caret (R) can be invaluable.

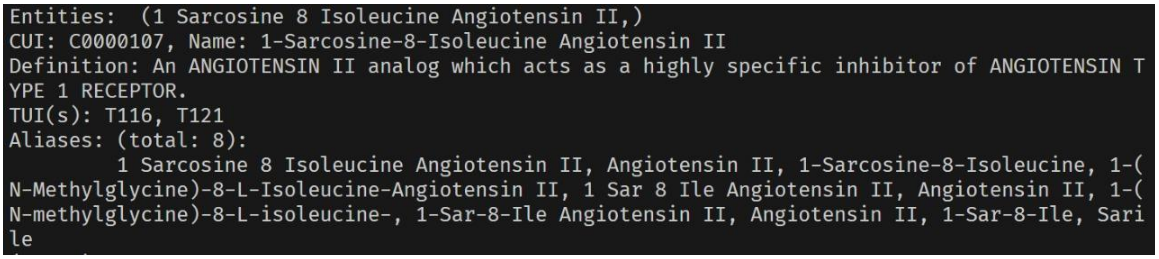
3.4 Experiment and Result for Validation and Verification

Entity Aliases input:



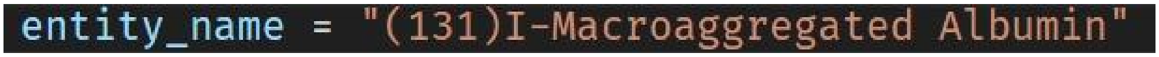
```
entity_name = "1 Sarcosine 8 Isoleucine Angiotensin II"
```

Entity Aliases Output:



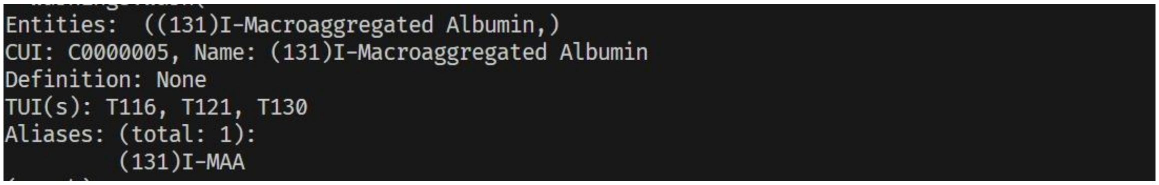
```
Entities:  (1 Sarcosine 8 Isoleucine Angiotensin II,)
CUI: C0000107, Name: 1-Sarcosine-8-Isoleucine Angiotensin II
Definition: An ANGIOTENSIN II analog which acts as a highly specific inhibitor of ANGIOTENSIN T
YPE 1 RECEPTOR.
TUI(s): T116, T121
Aliases: (total: 8):
         1 Sarcosine 8 Isoleucine Angiotensin II, Angiotensin II, 1-Sarcosine-8-Isoleucine, 1-(
N-Methylglycine)-8-L-Isoleucine-Angiotensin II, 1 Sar 8 Ile Angiotensin II, Angiotensin II, 1-(
N-methylglycine)-8-L-isoleucine-, 1-Sar-8-Ile Angiotensin II, Angiotensin II, 1-Sar-8-Ile, Sari
le
```

Input:



```
entity_name = "(131)I-Macroaggregated Albumin"
```

Output:



```
Entities:  ((131)I-Macroaggregated Albumin,)
CUI: C0000005, Name: (131)I-Macroaggregated Albumin
Definition: None
TUI(s): T116, T121, T130
Aliases: (total: 1):
         (131)I-MAA
```

3.5 Conclusion and Future work

Conclusion:

In summary, this project successfully implemented advanced NER and Entity Linkage techniques for extracting and linking medical entities from complex datasets. The user-friendly interface facilitated easy integration into medical research, potentially improving clinical decision-making and patient care. Further enhancements in NLP models and dataset diversity could amplify the system's accuracy and broaden its applications in the medical domain.

4. Reference

1. Lample, Guillaume, et al. "Neural Architectures for Named Entity Recognition." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

2. Wu, Yonghui, et al. "Entity Linking via Joint Encoding of Types, Descriptions, and Context." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

3. Bird, Steven, Edward Loper, and Ewan Klein. "Natural Language Processing with Python." O'Reilly Media, 2009.

4. Ratinov, Lev, and Dan Roth. "Design Challenges and Misconceptions in Named Entity Recognition." Proceedings of the Thirteenth Conference on Computational Natural Language Learning. 2009.

5. Neumann, Mark et al. "OpenAI GPT-3: A Language Model for Few-Shot Learning and Reasoning." arXiv preprint arXiv:2005.14165 (2020).

6. Manning, Christopher D., et al. "The Stanford CoreNLP Natural Language Processing Toolkit." Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

7. Neumann, Mark et al. "Knowledge-Enhanced Contextual Word Representations." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

.