Database Credentialed Access

Medication Extraction Labels for MIMIC-IV-Note Clinical Database

Akshay Goel , Almog Gueta , Omry Gilon , Sofia Erell , Amir Feder

Published: Dec. 12, 2023. Version: 1.0.0

When using this resource, please cite: (show more options)
Goel, A., Gueta, A., Gilon, O., Erell, S., & Feder, A. (2023). Medication Extraction Labels for MIMIC-IV-Note Clinical Database (version 1.0.0). *PhysioNet*. https://doi.org/10.13026/ps1s-ab29.

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Abstract

This dataset release provides medication extraction labels for a subset of 600 discharge summaries from the 2023 MIMIC-IV-Note dataset. These labels are consistent with the schema from the 2009 i2b2 Workshop on NLP Challenges dataset. We utilized a Large Language Model (LLM) pipeline to generate these labels, achieving performance on par with the average human annotation specialist.

Background

The MIMIC-IV-Note dataset is an invaluable asset comprising a comprehensive collection of 331,794 de-identified discharge summaries from 145,915 patients at Beth Israel Deaconess Medical Center in Boston, MA, USA [1,2]. However, the MIMIC-IV-Note dataset does not contain detailed Natural Language Processing (NLP) labels for the medication extraction task, similar to the 2009 i2b2 challenge dataset [3]. In this dataset release, we apply a novel approach to data labeling using Large Language Models (LLMs).

Traditionally, clinical text analysis relies heavily on human experts for labeling, a process both time-consuming and costly. Our methodology employs LLMs for initial label generation on a subset of the MIMIC-IV-Note dataset, aiming to streamline the labeling process and address the practical limitations of manual annotation. The LLM-generated labels have demonstrated quality comparable to that of an average human annotation-specialist in a rigorous evaluation using the 2009 i2b2 challenge dataset [4].

Methods

The data in this release was processed using a Large Language Model (LLM) labeling pipeline applied to a subset of the MIMIC-IV-Note dataset. In our LLM pipeline, we utilized Google's PaLM 2 model for its exceptional performance across various tasks [5]. The LLM was conditioned with a task-specific, few-shot prompt tailored for medication extraction [6]. This prompt included context, a comprehensive task description, and example input-output pairs. The prompt also specifies the output format, for example, using YAML syntax in the provided examples (example provided below). Segments of text from the dataset were input into the LLM using a structured prompt. The LLM's output was then converted into structured data objects by a post-processing module, which transformed the generative text into NER-RE (Named Entity Recognition-Relation Extraction) structured objects. These encapsulated pertinent medication details and associated RE information.

To ensure high-quality, relevant LLM outputs, a series of iterative experiments were conducted. These refined the prompts based on error patterns noted in the development set, with evaluations aimed at optimizing NER-RE task performance. The effectiveness of LLM-generated annotations was also rigorously compared to those made by human annotation specialists, demonstrating the significant value of integrating LLMs into the annotation process.

For detailed examples of prompt implementations, resolver logic, and evaluation process, please refer to the accompanying manuscript [4].

Example Abbreviated Prompt

~~~
description:
You are a medical assistant with expertise in document processing.

Your task is to identify and tag medication-related entity groups.

examples:
  question: "'Ibuprofen as needed, and diclofenac for one month as needed, for abdominal discomfort.'"
  answer:
    ```yaml
    entities:
      - group: 1
        MEDICATION:
          text: Ibuprofen
        DOSE:
          text: ''
        FREQUENCY:
          text: as needed
        DURATION:
          text: ''
        REASON:
          text: abdominal discomfort
        MODE:
          text: ''
      - group: 2
        MEDICATION:
          text: diclofenac
        DOSE:
          text: ''
        FREQUENCY:
          text: as needed
        DURATION:
          text: for one month
        REASON:
          text: abdominal discomfort
        MODE:
          text: ''
    ```
~~~

Data Description

The dataset includes labels for a random subset of MIMIC-IV-Note discharge summaries, stored in CSV files. The ‘note id’ in each filename corresponds to a document in the MIMIC-IV-Note dataset. The label files contain detailed information about medications, including their name, dosage, mode of administration, frequency, duration, and reason for administration, mirroring the fields in the 2009 i2b2 challenge dataset. These details are organized into medication group entries, each assigned a unique group code. Future updates aim to include an expanded set of labels covering more MIMIC-IV-Note discharge summaries.

CSV File Format

- Start Position: Indicates the starting character position of the annotated text in the source document.
- End Position: Indicates the ending character position.
- Annotation: Specifies the type of annotation (e.g., REASON, MEDICATION, MODE, DURATION).
- Group: Assigns a unique identifier that links the annotation to its respective medication group entry.

Usage Notes

The labels are directly mappable to their corresponding discharge summaries in the MIMIC-IV-Note dataset using the 'note_id'. For each document, entity labels are locatable within the text using the 'start position' and 'end position' fields. Below is an example Python code demonstrating how to map the text from MIMIC-IV-Note discharge summaries to the provided labels. The import_mimic_iv_text function integrates the textual data from MIMIC-IV-Note into the label CSV dataframe as a new column.

```
import os
import re
import pandas as pd

def import_mimic_iv_text(
    mimic_iv_discharge_df: pd.DataFrame,
    label_dir_path: str,
    label_filename: str,
    ) -> pd.DataFrame:
    """Imports MIMIC-IV-Note data into a “Text” column of the label dataframe.

    Parameters:
    mimic_iv_discharge_df (pd.DataFrame): DataFrame containing discharge
    summaries from MIMIC-IV-Note discharge.csv.gz.
    label_dir_path (str): The directory path of label CSV files.
    label_filename (str): The filename of a label CSV file.

    Returns:
    pd.DataFrame: The DataFrame with labels and associated MIMIC-IV text.
    """
    note_id = extract_note_id(label_filename)
    document_row = mimic_iv_discharge_df.loc[
        mimic_iv_discharge_df.note_id == note_id
    ]

    document_text = document_row.text.iloc[0]

    label_df = pd.read_csv(os.path.join(label_dir_path, label_filename))
    label_df["Text"] = None

    # Mapping each label to its corresponding text
    for idx, row in label_df.iterrows():
        text_slice = slice(row["Start Position"], row["End Position"])
        label_df.at[idx, "Text"] = document_text[text_slice]

    return label_df


def extract_note_id(s: str) -> str:
    """
    Extracts the note id from a string that contains it in a specific format.
    For example, it extracts '12345-DS-01' from the filename 'NoteID-12345-DS-01'.

    Parameters:
    s (str): The string to extract the note id from. The string must contain 'NoteID-' followed by the
actual note id.

    Returns:
    str: The extracted note id.

    Raises:
    ValueError: If the note id cannot be extracted.
    """
    match = re.search(r"NoteID-([0-9A-Za-z-]+)", s)
    if match:
        return match.group(1)
    else:
        raise ValueError(f"Could not extract note_id from {s}")
```

Release Notes

Version: 1.0.0: initial release.

Ethics

Data consists of extracted NLP labels from the existing MIMIC-IV-Note dataset. No additional data were collected from human subjects. No API services were utilized that log sensitive data for human review.

Conflicts of Interest

The authors of this data work for Google LLC.

References

1. Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet. https://doi.org/10.13026/1n74-ne17.
2. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.
3. Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 17(5), 514-518.
4. Goel, A., Gueta, A., Gilon, O., Erell, S., Liu, C., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Steiner, J., Laish, I., & Feder, A. (2023). LLMs Accelerate Annotation for Medical Information Extraction. In Machine Learning for Healthcare Conference. PMLR. (Accepted, In press).
5. Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Contents

Parent Projects

Medication Extraction Labels for MIMIC-IV-Note Clinical Database was derived from:

- MIMIC-IV-Note: Deidentified free-text clinical notes v2.2

Please cite them when using this project.

Share

Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
PhysioNet Credentialed Health Data License 1.5.0

Data Use Agreement:
PhysioNet Credentialed Health Data Use Agreement 1.5.0

Required training:
CITI Data or Specimens Only Research

Discovery

DOI (version 1.0.0):
https://doi.org/10.13026/ps1s-ab29

DOI (latest version):
https://doi.org/10.13026/6d3r-w470

Corresponding Author

You must be logged in to view the contact information.

Files

This is a restricted-access resource. To access the files, you must fulfill all of the following requirements:

- be a credentialed user
- complete required training:
    - CITI Data or Specimens Only Research
      You may submit your training here.
- sign the data use agreement for the project

PhysioNet is a repository of freely-available medical research data, managed by the MIT Laboratory for Computational Physiology.

Supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362.

For more accessibility options, see the MIT Accessibility Page.

Back to top