NER for Medical Domain

Contents

What is NER?

Named entity recognition (NER) is a natural language processing (NLP) method that extracts information from text. NER involves detecting and categorizing important information in text known as named entities. Named entities refer to the key subjects of a piece of text, such as names, locations, companies, events and products, as well as themes, topics, times, monetary values and percentages.

NER is also referred to as entity extraction, chunking and identification. It's used in many fields in artificial intelligence (AI), including machine learning (ML), deep learning and neural networks. NER is a key component of NLP systems, such as chatbots, sentiment analysis tools and search engines. It's used in healthcare, finance, human resources (HR), customer support, higher education and social media analysis.

Applications of NER in Medical and Hospital Domain:

Named Entity Recognition (NER) plays a crucial role in the healthcare and medical domain by extracting and categorizing relevant information from unstructured text. Here are some applications of NER in hospitals and the medical field:

1. Clinical Information Extraction:

Patient Records: NER can be used to extract key information such as patient names, medical conditions, medications, and treatment plans from electronic health records (EHRs). This helps in organizing and summarizing patient data for healthcare professionals.

2. Disease Surveillance:

Epidemiological Studies: NER can assist in tracking and monitoring the spread of diseases by extracting relevant information from medical literature, news articles, and social media. This helps public health agencies in early detection and response to outbreaks.

3. Drug Discovery and Development:

Literature Mining: NER can aid in mining scientific literature to identify mentions of genes, proteins, and chemical compounds. This is valuable in drug discovery, as researchers can gather information about potential drug targets and their relationships.

4. Billing and Coding:

Medical Coding: NER can automate the process of assigning diagnostic and procedural codes to patient records. This improves accuracy in billing, reduces errors, and ensures proper reimbursement for healthcare services.

5. Clinical Trial Matching:

Patient Eligibility Criteria: NER can assist in matching eligible patients with clinical trial criteria by extracting relevant information from patient records. This streamlines the patient recruitment process for clinical trials.

6. Adverse Event Detection:

Pharmacovigilance: NER can help in identifying and extracting information related to adverse drug reactions or side effects from various sources such as medical literature, patient records, and social media. This is crucial for ensuring drug safety.

7. Health Information Management:

Entity Resolution: NER can aid in resolving references to the same entity across different documents or systems, ensuring that patient information is accurately linked and maintained.

8. Telemedicine and Chatbots:

Patient Communication: NER can be used in chatbots to understand and respond to patient queries, extract relevant information, and provide personalized healthcare advice.

9. Research and Literature Analysis:

NER facilitates the extraction of valuable information from a vast amount of medical literature, aiding researchers in literature reviews, evidence synthesis, and staying updated on the latest advancements in healthcare.

Given Data Set:

Mrs. May visited Leeds General Infirmary hospital which is located near Burley Road. She was not feeling well and had an appointment with Dr. Ray Johnson. Mrs. May was last seen by Dr. Parker in May 2019, who advised her Chest X-ray. Dr. Ray examined her and by looking at her xray report, he advised her to quit smoking. She told him that she already worked very hard to quit it and used to have coffee instead every day from Burley Café. Dr. Johnson requested her to go to Radiography lab to get a new X-ray so he could analyze her current situation. He also prescribed her medicines and asked her to collect them from Lloyds Pharmacy in front of Parkinson's building and visit again on 13th June.

Extended Dataset:

We requested GPT to expand the provided dataset to include 100 additional lines and subsequently conducted annotations. Attached are screenshots displaying the initial and concluding segments of the extended data for reference.

Initial Extended Dataset:

DATASET:

1. "Mrs. May visited Leeds General Infirmary hospital located near Burley Road for her health checkup."

2. "During her appointment with Dr. Ray Johnson, she discussed her ongoing health issues."

3. "Dr. Parker, who saw Mrs. May in May 2019, recommended a Chest x-Ray at that time."

4. "The Chest x-Ray report was later examined by Dr. Ray, who advised Mrs. May to quit smoking."

5. "Mrs. May informed Dr. Ray that she had already successfully quit smoking and replaced it with daily coffee from Burley Cafe."

6. "Dr. Johnson suggested visiting the Radiography lab for a new x-Ray to assess her current health status."

7. "In the Radiography lab, Mrs. May underwent the recommended x-Ray for further analysis."

8. "Dr. Ray Johnson, after reviewing the new x-Ray report, prescribed specific medications for Mrs. May."

9. "He advised her to collect the prescribed medicines from Lloyds Pharmacy, situated in front of the Parkinson building."

10. "Mrs. May was scheduled to visit Dr. Johnson again on 13th June for a follow-up."

11. "The Parkinson building, near Lloyds Pharmacy, is a prominent landmark in the hospital vicinity."

11. "The Parkinson building, near Lloyds Pharmacy, is a prominent landmark in the hospital vicinity."

12. "Burley Cafe, where Mrs. May enjoys her daily coffee, is known for its cozy ambiance."

13. "Patients at Leeds General Infirmary often have their appointments scheduled on Burley Road."

14. "A common concern among patients is the waiting time at Leeds General Infirmary hospital."

15. "Dr. Ray Johnson specializes in respiratory health and is known for his expertise in interpreting x-Ray reports."

16. "Radiography lab technicians at Leeds General Infirmary are trained to operate advanced imaging equipment."

17. "Patients are encouraged to share their complete medical history during appointments at Leeds General Infirmary."

18. "Mrs. May's commitment to quitting smoking impressed both Dr. Parker and Dr. Ray Johnson."

19. "Patients often gather health-related information from the notice board near the hospital's main entrance."

Ending Extended Dataset:

90. "The hospital's research department collaborates with pharmaceutical companies to bring cutting-edge treatments to patients."

91. "Mrs. May's successful journey towards better health serves as a motivating story in the hospital's newsletter."

92. "Patients attending the radiology department often comment on the efficiency of the imaging equipment."

93. "Leeds General Infirmary's commitment to continuous improvement is reflected in its ongoing staff training programs."

94. "The hospital's dedication to patient-centered care is embedded in its mission statement and core values."

95. "Dr. Ray Johnson's office is equipped with telehealth capabilities for virtual consultations when necessary."

96. "Patients attending the hospital's rehabilitation services experience personalized plans for their recovery."

97. "Leeds General Infirmary's outreach program extends to local schools, promoting health education for children."

98. "The hospital's emergency response team is trained to handle a variety of medical situations with promptness and efficiency."

99. "Patients are encouraged to utilize the hospital's online resources for health-related information and self-care tips."

100. "Mrs. May's consistent follow-up appointments with Dr. Ray Johnson showcase the importance of ongoing patient-doctor collaboration."

Solution a/c to given guidelines:

1. Baseline Solution:

1. Data Preprocessing:

Description: Annotate the dataset to mark entities such as names of people, locations, dates, and medical terms. Split the dataset into training and testing sets.

Example: In the sentence "Dr. Parker, who saw Mrs. May in May 2019, recommended a Chest x-Ray at that time," entities like "Dr. Parker" and "May 2019" would be annotated during data preprocessing.

extended.json

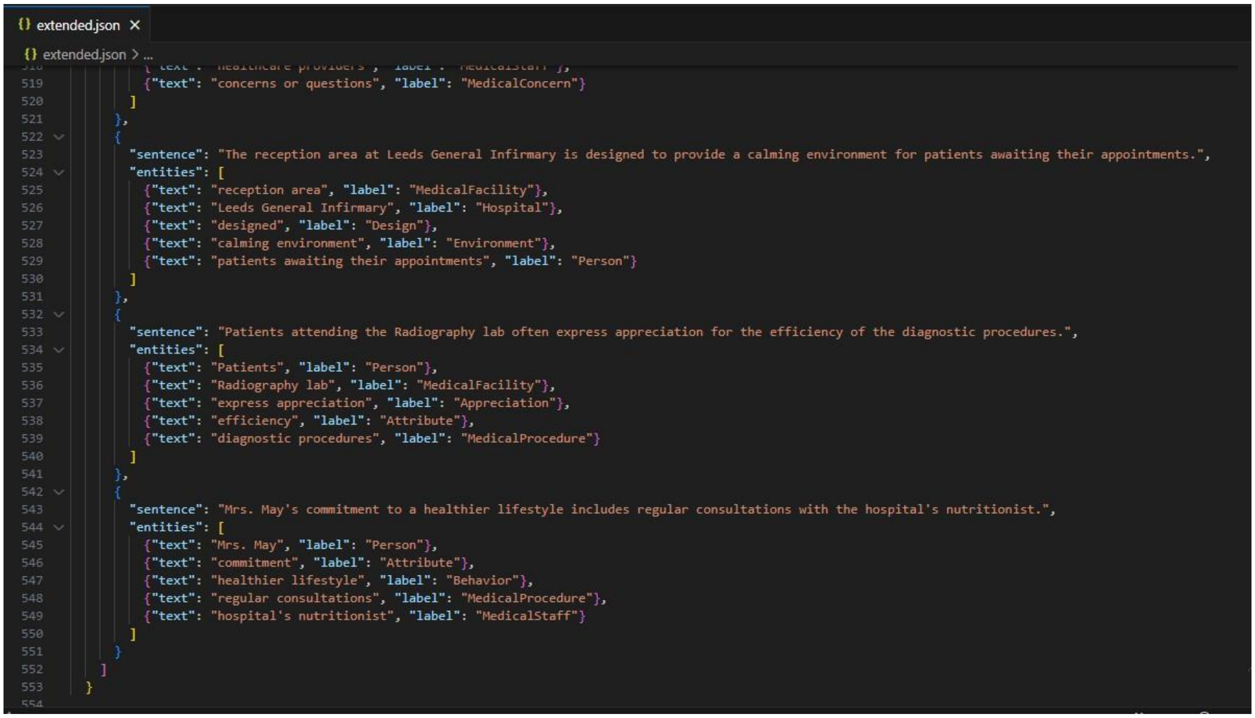
extended.json > ...


```json
{
  "data": [
    {
      "sentence": "Mrs. May visited Leeds General Infirmary hospital located near Burley Road for her health checkup.",
      "entities": [
        {"text": "Mrs. May", "label": "Person"},
        {"text": "Leeds General Infirmary", "label": "Hospital"},
        {"text": "Burley Road", "label": "Location"},
        {"text": "health checkup", "label": "MedicalProcedure"}
      ]
    },
    {
      "sentence": "During her appointment with Dr. Ray Johnson, she discussed her ongoing health issues.",
      "entities": [
        {"text": "Dr. Ray Johnson", "label": "Doctor"},
        {"text": "health issues", "label": "MedicalCondition"}
      ]
    },
    {
      "sentence": "Dr. Parker, who saw Mrs. May in May 2019, recommended a Chest x-Ray at that time.",
      "entities": [
        {"text": "Dr. Parker", "label": "Doctor"},
        {"text": "Mrs. May", "label": "Person"},
        {"text": "May 2019", "label": "Date"},
        {"text": "Chest x-Ray", "label": "MedicalProcedure"}
      ]
    },
    {
      "sentence": "The Chest x-Ray report was later examined by Dr. Ray, who advised Mrs. May to quit smoking.",
      "entities": [
        {"text": "Chest x-Ray", "label": "MedicalProcedure"},
        {"text": "Dr. Ray", "label": "Doctor"},
        {"text": "Mrs. May", "label": "Person"},
        {"text": "quit smoking", "label": "Behavior"}
      ]
    },
```

```json
{
  "sentence": "Mrs. May informed Dr. Ray that she had already successfully quit smoking and replaced it with daily coffee from Burley Cafe.",
  "entities": [
    {"text": "Mrs. May", "label": "Person"},
    {"text": "Dr. Ray", "label": "Doctor"},
    {"text": "quit smoking", "label": "Behavior"},
    {"text": "daily coffee", "label": "Behavior"},
    {"text": "Burley Cafe", "label": "Location"}
  ]
},
{
  "sentence": "Dr. Johnson suggested visiting the Radiography lab for a new x-Ray to assess her current health status.",
  "entities": [
    {"text": "Dr. Johnson", "label": "Doctor"},
    {"text": "Radiography lab", "label": "MedicalFacility"},
    {"text": "new x-Ray", "label": "MedicalProcedure"},
    {"text": "current health status", "label": "MedicalCondition"}
  ]
},
{
  "sentence": "In the Radiography lab, Mrs. May underwent the recommended x-Ray for further analysis.",
  "entities": [
    {"text": "Radiography lab", "label": "MedicalFacility"},
    {"text": "Mrs. May", "label": "Person"},
    {"text": "recommended x-Ray", "label": "MedicalProcedure"},
    {"text": "further analysis", "label": "MedicalCondition"}
  ]
},
{
  "sentence": "Dr. Ray Johnson, after reviewing the new x-Ray report, prescribed specific medications for Mrs. May.",
  "entities": [
    {"text": "Dr. Ray Johnson", "label": "Doctor"},
    {"text": "new x-Ray report", "label": "MedicalReport"},
    {"text": "specific medications", "label": "Medication"},
    {"text": "Mrs. May", "label": "Person"}
  ]
```


Ln 1, Col 1 Spaces: 2 UTF-8 CRLF JSON Go Live

This annotation continues …




```json
{} extended.json
{"text": "concerns or questions", "label": "MedicalConcern"}
]
},
{
"sentence": "The reception area at Leeds General Infirmary is designed to provide a calming environment for patients awaiting their appointments.",
"entities": [
{"text": "reception area", "label": "MedicalFacility"},
{"text": "Leeds General Infirmary", "label": "Hospital"},
{"text": "designed", "label": "Design"},
{"text": "calming environment", "label": "Environment"},
{"text": "patients awaiting their appointments", "label": "Person"}
]
},
{
"sentence": "Patients attending the Radiography lab often express appreciation for the efficiency of the diagnostic procedures.",
"entities": [
{"text": "Patients", "label": "Person"},
{"text": "Radiography lab", "label": "MedicalFacility"},
{"text": "express appreciation", "label": "Appreciation"},
{"text": "efficiency", "label": "Attribute"},
{"text": "diagnostic procedures", "label": "MedicalProcedure"}
]
},
{
"sentence": "Mrs. May's commitment to a healthier lifestyle includes regular consultations with the hospital's nutritionist.",
"entities": [
{"text": "Mrs. May", "label": "Person"},
{"text": "commitment", "label": "Attribute"},
{"text": "healthier lifestyle", "label": "Behavior"},
{"text": "regular consultations", "label": "MedicalProcedure"},
{"text": "hospital's nutritionist", "label": "MedicalStaff"}
]
}
]
}
```


2. Choose a NER Framework:

Description: Select a suitable NER framework for the model. The baseline solution uses spaCy's pre-trained English model (en_core_web_sm).

Example: When applying spaCy to the sentence "The Chest x-Ray report was later examined by Dr. Ray, who advised Mrs. May to quit smoking," the pre-trained model recognizes entities like "Dr. Ray" as a person.

3. Define Entity Types:

Description: Specify the types of entities to recognize, such as person names, locations, dates, and medical terms.

Example: In the sentence "The Parkinson building, near Lloyds Pharmacy, is a prominent landmark in the hospital vicinity," the entity "The Parkinson building" is defined as a location during the annotation process.

4. Feature Extraction:

Description: Convert the text data into a format suitable for training, involving tokenization and feature extraction.

Example: Tokenization and feature extraction would be applied to sentences like "Patients are encouraged to share their complete medical history during appointments at Leeds General Infirmary" during the model training process.

5. Model Selection:

Description: Choose a model architecture based on the selected framework.

Example: The chosen model would learn to identify entities like "Radiography lab technicians" in sentences such as "Radiography lab technicians at Leeds General Infirmary are trained to operate advanced imaging equipment."

6. Evaluation Metrics:

Description: Evaluate the model on a separate test set, measuring metrics like precision, recall, and F1-score.

Example: During evaluation, the model's effectiveness in identifying entities like "Burley Cafe" in sentences such as "Burley Cafe, where Mrs. May enjoys her daily coffee, is known for its cozy ambiance" is assessed.

7. Deployment:

Description: Deploy the NER model for extracting entities from new text data.

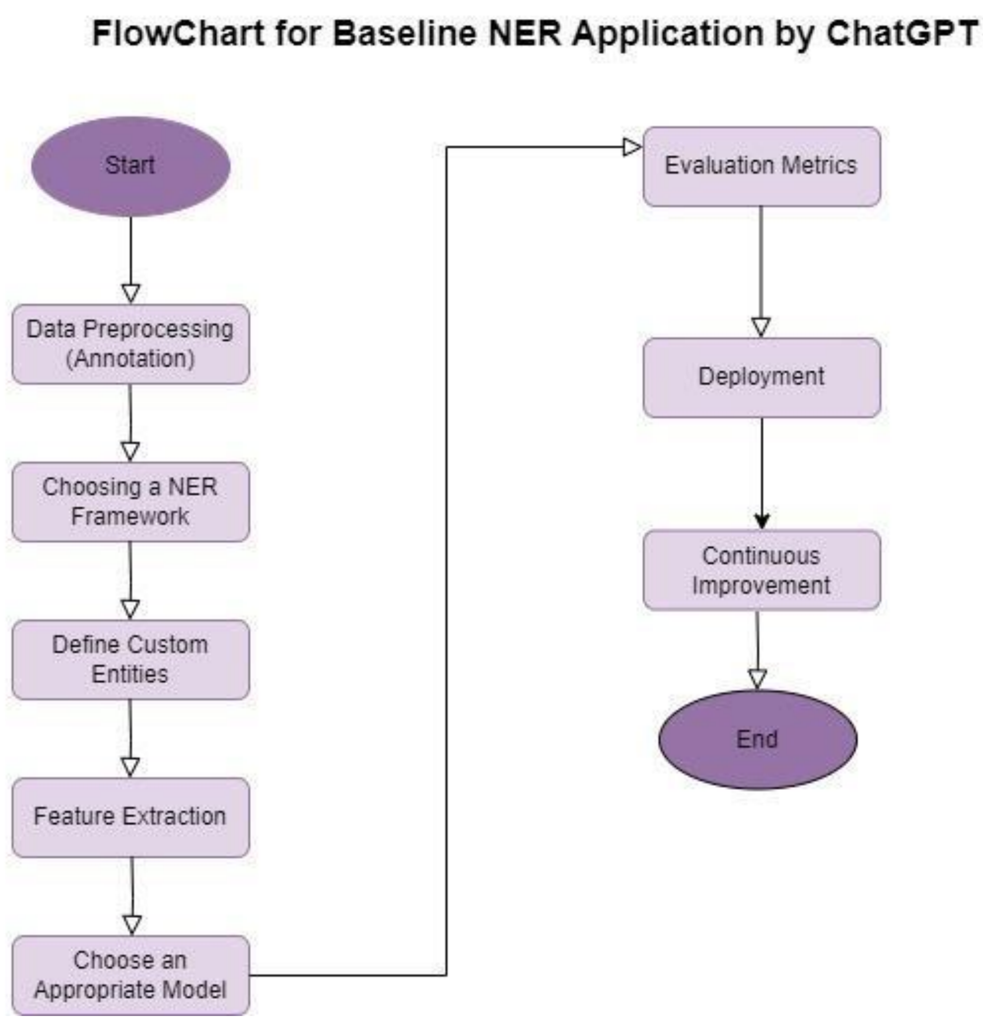
Example: Upon deployment, the model can extract entities like "Dr. Ray Johnson" from sentences like "Dr. Ray Johnson's seminars on respiratory health attract healthcare professionals from neighboring institutions."

8. Continuous Improvement:

Description: Continually monitor the model's performance and update it as needed, adapting to new entities.

Example: Continual improvement would involve adapting the model to handle evolving entities, ensuring it recognizes terms introduced in sentences such as "The hospital's commitment to sustainability is evident in its eco-friendly initiatives.

Baseline Design of System by Chatgpt:



Baseline Implementation by Chatgpt:

```python
import spacy
import json
```

```python
nlp = spacy.load("en_core_web_sm")
```

```python
with open("extended.json", "r") as file:
  jsondata = json.load(file)

data = jsondata["data"]
```

```python
for entry in data:
    sentence = entry["sentence"]
    doc = nlp(sentence)
     # Extract entities and labels
    entities = [{"text": ent.text, "label": ent.label_} for ent in doc.ents]
    ground_truth = entry["entities"]

    print("Sentence:", sentence)
    print("Predicted Entities:", entities)
    print("Ground Truth Entities:", ground_truth)
    print("\n" + "="*50 + "\n")
```

OUTPUT:

```
Sentence: Mrs. May visited Leeds General Infirmary hospital located near Burley Road for her health checkup.
Predicted Entities: [{'text': 'May', 'label': 'PERSON'}, {'text': 'Leeds General', 'label': 'ORG'}, {'text': 'Burley Road', 'label': 'FAC'}]
Ground Truth Entities: [{'text': 'Mrs. May', 'label': 'Person'}, {'text': 'Leeds General Infirmary', 'label': 'Hospital'}, {'text': 'Burley Road', 'label': 'Location'}, {'text': 'health checkup', 'label': 'MedicalProcedure'}]

==================================================

Sentence: During her appointment with Dr. Ray Johnson, she discussed her ongoing health issues.
Predicted Entities: [{'text': 'Ray Johnson', 'label': 'PERSON'}]
Ground Truth Entities: [{'text': 'Dr. Ray Johnson', 'label': 'Doctor'}, {'text': 'health issues', 'label': 'MedicalCondition'}]

==================================================

Sentence: Dr. Parker, who saw Mrs. May in May 2019, recommended a Chest x-Ray at that time.
Predicted Entities: [{'text': 'Parker', 'label': 'PERSON'}, {'text': 'May', 'label': 'PERSON'}, {'text': 'May 2019', 'label': 'DATE'}, {'text': 'Chest x-Ray', 'label': 'PRODUCT'}]
Ground Truth Entities: [{'text': 'Dr. Parker', 'label': 'Doctor'}, {'text': 'Mrs. May', 'label': 'Person'}, {'text': 'May 2019', 'label': 'Date'}, {'text': 'Chest x-Ray', 'label': 'MedicalProcedure'}]

==================================================
```

Python

The provided code lacks adequacy for addressing the issue, as it fails to offer annotations for any of the medical terms. Consequently, Chat-GPT is deemed unsuitable for developing a system that can be relied upon for such purposes.

2. Our Identified Problems:

Identifying Incompleteness in solution provided:

Here are some aspects of the baseline solution that could be considered incomplete or areas for improvement:

1. Training on a Custom Dataset:

The baseline solution uses spaCy's pre-trained English model (en_core_web_sm). For better performance, especially in a specialized domain like medical NER, it's advisable to train the model on a custom dataset that includes labeled examples relevant to the medical domain.

2. Evaluation Metrics:

The solution doesn't include an evaluation metric to measure the performance of the NER model. In a real-world scenario, you would need to evaluate the model's precision, recall, and F1-score on a separate test set to ensure its effectiveness.

3. Handling Ambiguities:

The solution doesn't address potential ambiguities or overlapping entities in the text. For instance, a location like "Burley Road" might be tagged as both a "Location" and a "MedicalFacility," leading to potential conflicts.

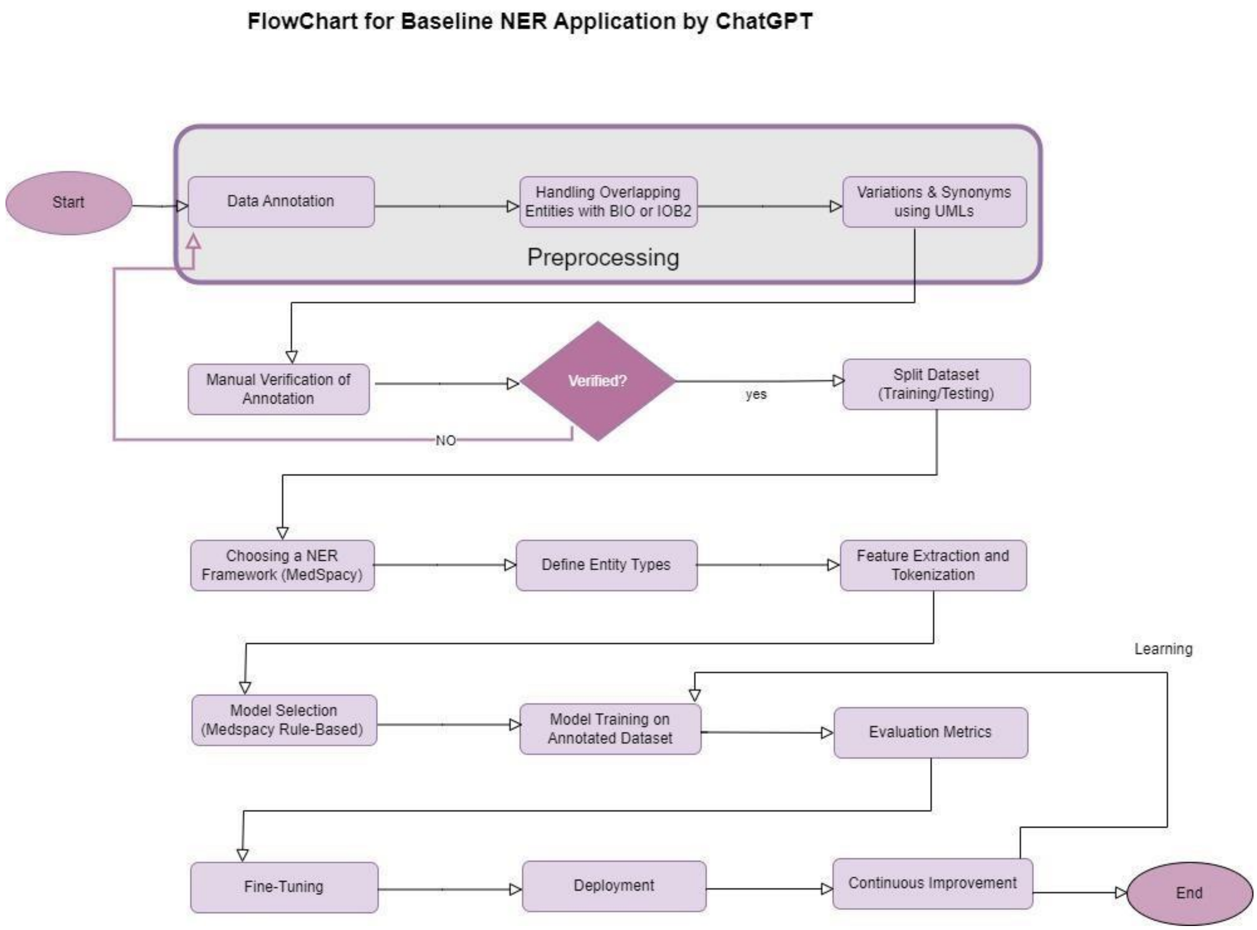
4. Optimizing for Domain-Specific Terms:

The model might not be optimized for medical domain-specific terms. Fine-tuning the model on a dataset containing medical entities would improve its ability to recognize and classify medical terms accurately.

5. Handling Large Datasets:

The solution processes a small set of 10 sentences. In a real-world scenario with a large dataset, you might need to handle memory and processing efficiency issues.

3. Improved NER Application Design:



Final Flow of Improved NER:

Data Preprocessing:

Description: Annotate the dataset to mark entities. For instance, label "Mrs. May" as a person, "Leeds General Infirmary" as a location, and "13th June" as a date entity.

Example: In the sentence "Mrs. May visited Leeds General Infirmary hospital located near Burley Road for her health checkup," annotate "Mrs. May" as a person, "Leeds General Infirmary" as a location, and "Burley Road" as another location.

- **Handling Overlapping Entities:**

  Description: Resolve any conflicts or overlapping entities during the annotation. Consider using tagging schemes like BIO or IOB2 to handle multiple entities in a sequence.

  Example: In the sentence "Dr. Parker, who saw Mrs. May in May 2019," annotate "Dr. Parker" as a medical professional, "Mrs. May" as a person, and "May 2019" as a date. Handle the overlap between "Dr. Parker" and "Mrs. May."

- **Accounting for Variations and Synonyms:**

  Description: During annotation, consider variations and synonyms. Ensure the model is exposed to diverse terms. Leverage external resources like UMLS for comprehensive term coverage.

  Example: In the sentence "Dr. Ray Johnson specializes in respiratory health," annotate "Dr. Ray Johnson" as a medical professional and "respiratory health" as a medical condition. Account for variations like "respiratory" and "breathing."

- **Manual Verification of Annotations:**

  Description: Have a domain expert, familiar with medical terminology, review and verify the initial annotations. Correct any mistakes or inconsistencies.

  Example: The manual annotator ensures that entities like "Mrs. May" are correctly labeled as a person, and that medical terms are appropriately identified.

- **Split Dataset:**

Description: Divide the annotated dataset into training and testing sets.

Example: If you have 100 sentences, you might use 80 for training and 20 for testing.

- **Choose a Framework:**

  Description: Select MedSpacy as the NER framework for medical entity recognition.

  Example: Utilize MedSpacy's capabilities to recognize entities like names, locations, dates, and medical terms in the dataset.

- **Define Entity Types:**

  Description: Identify and define various medical entities to be recognized, such as patient names, medical conditions, medications, medical professionals, and dates.

  Example: Define entities like "patient," "medical condition," "medication," "medical professional," and "date" based on the dataset.

- **Feature Extraction:**

  Description: Utilize MedSpacy for tokenization and feature extraction during the training process.

  Example: Tokenize sentences into individual words and extract features that capture the context of medical entities.

- **Model Selection:**

Description: Choose MedSpacy's rule-based and machine learning components as the model architecture.

Example: Leverage MedSpacy's rule-based system to capture specific patterns in medical entities, and use machine learning components for more nuanced recognition.

- **Train the Model:**

  Description: Train the MedSpacy NER model on the annotated training dataset. Adjust hyper parameters as needed.

  Example: Train the model using the 80% of the dataset, teaching it to recognize entities based on the annotated examples.

- **Evaluation Metrics:**

  Description: Evaluate the MedSpacy model on the testing dataset using metrics like precision, recall, and F1-score.

  Example: Assess how well the model performs on the 20% of the dataset it hasn't seen before, comparing its predictions to the annotated entities.

- **Fine-Tuning:**

  Description: Fine-tune the MedSpacy model based on the evaluation results. This may involve adjusting parameters or incorporating additional labeled data.

  Example: If the model struggles with specific types of entities, refine its understanding through additional training or parameter adjustments.

- **Deployment:**

  Description: Deploy the MedSpacy NER model for extracting entities from new medical text data.

  Example: Implement the model in a healthcare system to automatically extract relevant information from new patient records.

- **Continuous Improvement:**

  Description: Monitor the model's performance over time and update it as needed. Regularly update external resources to stay current with medical terminology.

  Example: Periodically retrain the model with new annotated data and update any external resources used for disambiguation.

Comparison & Analysis of Baseline and Improved Design:

1. Data Preprocessing:

Baseline: Annotation involves marking entities like names, locations, dates, and medical terms without specific entity type definitions.

Improved: Annotates entities with clear types (e.g., person, location, date) and includes handling of overlapping entities.

2. Handling Overlapping Entities:

Baseline: No explicit mention of handling overlapping entities during annotation.

Improved: Introduces the use of tagging schemes (BIO or IOB2) to handle overlapping entities, addressing conflicts in sequences.

3. Accounting for Variations and Synonyms:

Baseline: No explicit consideration for variations and synonyms during annotation.

Improved: Emphasizes accounting for variations and synonyms during annotation, ensuring exposure to diverse terms and leveraging external resources like UMLS.

4. Manual Verification of Annotations:

Baseline: No mention of manual verification by domain experts.

Improved: Recommends manual verification by domain experts to ensure correctness and consistency of annotations.

5. Split Dataset:

Baseline: Splits the dataset into training and testing sets without specific mention of the size or percentage.

Improved: Provides an example of an 80/20 split for training and testing, offering more concrete guidance.

6. Choose a Framework:

Baseline: Chooses spaCy's pre-trained English model.

Improved: Recommends MedSpacy as the NER framework, specifically designed for medical entity recognition.

7. Define Entity Types:

Baseline: Specifies entity types such as person names, locations, dates, and medical terms without focusing on medical subtypes.

Improved: Defines various medical entities such as patient names, medical conditions, medications, medical professionals, and dates.

8. Feature Extraction:

Baseline: Mentions tokenization and feature extraction without specifying the framework.

Improved: Utilizes MedSpacy for tokenization and feature extraction during the training process.

9. Model Selection:

Baseline: Recommends choosing a model architecture based on the selected framework, with no specific emphasis on medical entities.

Improved: Chooses MedSpacy's rule-based and machine learning components tailored for medical entity recognition.

10. Train the Model:

Baseline: Trains the NER model on the annotated training dataset without framework-specific details.

Improved: Trains the MedSpacy NER model on annotated examples with a focus on medical entities.

11. Evaluation:

Baseline: Evaluates the model on a separate test set, measuring metrics like precision, recall, and F1-score.

Improved: Evaluates the MedSpacy model on a testing dataset using similar metrics, assessing its performance on unseen medical text.

12. Fine-Tuning:

Baseline: Mentions fine-tuning based on evaluation results without specific details.

Improved: Recommends fine-tuning the MedSpacy model, potentially incorporating additional labeled data or adjusting parameters.

13. Deployment:

Baseline: Deploys the NER model for extracting entities from new text data.

Improved: Deploys the MedSpacy NER model specifically for extracting entities from new medical text data within a healthcare system.

14. Continuous Improvement:

Baseline: Mentions continual improvement but provides no specific details.

Improved: Emphasizes continuous monitoring, periodic retraining, and updating external resources for staying current with medical terminology.

In summary, the improved version provides more specific guidance tailored to medical entity recognition, addressing challenges such as overlapping entities, variations, and synonyms, and recommending the use of a specialized framework (MedSpacy). It also highlights the importance of manual verification by domain experts and continuous improvement through regular updates.