# How NLP Extracts Key Health Information from Unstructured Medical Reports: Challenges in Language Understanding and Accuracy

WenChing Li,
wl33@illinois.edu

December 8, 2024

**Abstract**

Natural Language Processing (NLP) is becoming an essential tool in the medical domain, especially in extracting key information from unstructured medical reports. This paper integrates insights from six selected papers that attempt to answer a core question: **How NLP Extracts Key Health Information from Unstructured Medical Reports: Challenges in Language Understanding and Accuracy**. Each study presents innovative methodologies for extracting critical health information, whether in structured or unstructured formats, with the strong desired goal of enhancing healthcare quality. These advancements demonstrate practical applications, such as supporting physicians in clinical decision-making and monitoring the progression of schizophrenia in patients. Despite these achievements, the studies highlight persistent challenges, including handling domain-specific language, ambiguity in medical terminologies, and ensuring data accuracy for sensitive use cases. Some of papers didn't provide clear and solid result, however, authors did show a series of great process and achievement.

## 1 Introduction

This paper examines the critical and expanding role of Natural Language Processing (NLP) in the medical field, highlighting its transformative potential despite the general public's limited awareness of its profound implications for the future. While some may view machine learning as a complex technology removed from everyday life, my literature review draws on six selected published papers to provide evidence of how NLP's capability to efficiently process critical data can significantly improve people's lives—both mentally and physically. The review explores how researchers investigate clinical conversations, electronic medical records (EMRs), and medical annotations, seeking connections between meditation practices and diseases.

## 2 Background

Reflecting on my personal experiences with multiple surgeries since I was born, I am deeply intrigued by how the earlier adoption of Natural Language Processing (NLP) could have reshaped my medical journey and positively impacted others facing similar challenges. For instance, if NLP technologies integrated with deep learning had been available decades ago, they might have enabled researchers and doctors to uncover critical findings, such as how the intake of folic acid and Vitamin A reduces the likelihood of congenital conditions like cleft lip and palate. Such insights could have saved my family and me significant amounts of life time and money spent on surgeries. This personal connection has fueled my profound interest in exploring the potential of NLP to transform healthcare and improve the lives of countless individuals from a scientific perspective.

To deepen my understanding, I have selected six insightful papers from the Association for Computational Linguistics(ACL). For example, in ”*Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge- Starting From Critical Lung Diseases*”, the group of authors categorized sentences into various medical-Annotation guideline, providing a foundation for understanding nuanced medical documents without deep medical knowledge. Similarly, ”*DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records*” focused on leveraging clinic annotations to uncover supportive documents that healthcare professionals might otherwise overlook due to their overwhelming volume or tedious nature. From ”*Summarizing Medical Conversations via Identifying Important Utterances*”, researchers analyzed conversations during doctor-patient appointments, uncovering actionable insights to enhance communication and decision-making. Meanwhile, ”*Prediction of Key Patient Outcomes from Sentence and Word of Medical Text Records*” investigated necessary but often overlooked adjustments in medication regimens, aiming to bridge the gap in doctors' understanding of patient needs. Additionally, ”*Context-aware Medication Event Extraction from Unstructured Text*” creatively explored the connection between prescriptions and meditation practices, highlighting how NLP could link seemingly unrelated factors for improved health outcomes. Lastly, ”*Time Expressions in Mental Health Records for Symptom Onset Extraction*”, though focused on mental health disorders, demonstrated an innovative application of NLP to trace optimal medication schedules, revealing its versatility across medical domains. Each assists me to gain knowledge about how to preprocess overwhelming data and improve accuracy.

# 3 Analysis

## 3.1 Extracting and Annotating Medical Text

”*Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge- Starting From Critical Lung Diseases.*”[1]

Its key concept is to focus on parsing and understanding unstructured text for actionable insights and apply the nature of ontology linking that connects entities across different ontology. They developed the tags below while parsing the medication annotation.

Now consider the following medication annotation: ”The patient was admitted on $< TIMEX3$ type=”DATE”>November 1, 2023$< /TIMEX3 >$, with $< D$ **certainty=”positive”>fever and cough**$< /D >$. Diagnosed with **pneumonia** and treated with **Amoxicillin 500 mg twice daily**. **Mild diarrhea**, likely a side effect, was observed during treatment. Discharged on **November 7, 2023**.”

Note that they successfully link ”mild diarrhea” as a potential side effect of ”Amoxicillin.” They further integrates annotation techniques with advanced models like BERT and BiLSTM-CRF to perform Named Entity Recognition (NER) on medical records. The annotation process tags key entities such as diseases ($< D >$), temporal expressions ($< TIMEX3 >$), and treatment states

| feature | tag | example |
|---|---|---|
| Diseases and symptoms | $< D >$ | I consider it $< D$ certainty=”positive”$>$primary lung cancer $< /D >$ |
| Anatomical entities | $< A >$ | in $< A >$the right lung$< /A >$ |
| Features and measurements | $< F >$ | $< F >$are diffused$< /F >$ in $< A >$the right lung$< /A >$ |
| Change | $< C >$ | $< C >$has disappeared$< /C >$ |
| TIMEX3 | $< TIMEX3 >$ | from $< TIMEX3$ type=”DATE”$>$last time$< /TIMEX3 >$ |
| Test | $< T >$ | $< T - test$ state=”executed”$>$Chest CT $< /T - test >$ |
| Medicine | $< M >$ | $< M - val >$100mL/1hr$< /M - val >$ |
| Remedy | $< R >$ | $< R$ state=”executed”$>$resection of superior lobe of left lung$< /R >$ |
| Clinical Context | $< CC >$ | He $< CC$ state=”executed”$>$was going$< /CC >$ |
| Pending | $< P >$ | was $< P >$spared$< /P >$ |

Table 1: Tag Entity

|  | Medical records | Radiography reports |
|---|---|---|
| Total documents annotated | 156 | 1000 |
| Average sentence count per document | 30.89 | 13.36 |
| Average word count per document | 268.73 | 142.02 |
| Total tag count |  |  |
| Disease | 2008 | 13897 |
| Anatomical Feature | 742 | 7123 |
| Feature | 325 | 5345 |
| Change | 678 | 1100 |
| TIMEX3 | 1820 | 1550 |
| T-Test | 716 | 852 |
| T-Key | 1957 | 40 |
| T-Val | 2116 | 3 |
| M-Key | 399 | 0 |
| M-Val | 170 | 0 |
| Remedy | 439 | 137 |
| Clinical Context | 331 | 28 |

Table 2: Tag Count

| Tag | Precision | Recall | F-score | Baseline |
|---|---|---|---|---|
| Diseases | 95.90% | 96.83% | 96.36 | 95.44 |
| Anatomical | 95.08% | 95.93% | 95.50 | 94.21 |
| Features | 92.48% | 94.77% | 93.61 | 93.09 |
| Change | 88.56% | 91.67% | 90.09 | 89.73 |
| TIMEX3 | 95.22% | 97.39% | 96.30 | 95.27 |
| T-Test | 94.80% | 95.35% | 95.07% | 92.13% |
| T-Key | 66.67% | 100.00% | 80.00 | 66.67 |
| T-Val | - | - | - | - |
| Remedy | 81.48% | 81.48% | 81.48 | 63.64 |
| Clinical Con | 83.33% | 71.43% | 76.92% | 44.44 |
| Overall | 94.65% | 95.95% | 95.30 | 94.26 |

Table 3: NER results of our BERT-based classifier. 'Baseline' denotes the F-score of the baseline system

$(< State >)$, resulting in a high-quality corpus optimized for training NLP models.

BERT, a Transformer-based pre-trained model, is fine-tuned for sequence labeling tasks. Using byte-pair encoding for tokenization, its final layer representations are classified into BIO tags. With its ability to capture rich contextual information, BERT achieves an F1 score of 95.30, excelling particularly in recognizing frequent labels like $< D >$ and $< TIMEX3 >$.

The baseline model, BiLSTM-CRF, leverages Word2Vec embeddings processed through a bidirectional LSTM, followed by a Conditional Random Field (CRF) layer to ensure valid tag sequences. While BiLSTM-CRF performs competitively on smaller datasets, its F1 scores are lower than BERT's, especially for categories with abundant annotations. For example, in a **Named Entity Recognition (NER)** task, consider the sentence:

*"Microsoft announced new features for Windows in Seattle on January 1, 2024."*

**BiLSTM-CRF** processes this through a bidirectional LSTM, capturing contextual dependencies, and uses a CRF layer to ensure valid label sequences, identifying "Microsoft" as an organization and "Seattle" as a location. However, **BERT** captures deeper contextual meanings and long-range dependencies more effectively, understanding the relationships between words and entities. BERT performs better, especially on larger datasets with abundant annotations, resulting in higher F1 scores for complex or varied entity categories.

Both models are trained on 1,000 annotated radiology reports, split into 80% for training and 20% for testing. BERT consistently outperforms BiLSTM-CRF across most categories. However, rare labels like$< T - key >$ show limitations due to data sparsity, highlighting areas for further research.

The study underscores BERT's strength in generalizing without overfitting, even on sparse annotations, while demonstrating the synergy between high-quality annotation and state-of-the-art NLP models to achieve robust performance in clinical NER tasks. But my focus is to look at how authors use invented tag entity to capture each weighted medical term and how they link each term's importance.

## 3.2 Extracting clinical notes and identify rug-disease relationships

*"DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records"*[3]

The goal of this paper is to create a benchmark dataset for multi-modal QA systems, addressing gaps in linking structured prescriptions with unstructured clinical notes to identify drug-disease relationships. The motivation is because EHRs contain both structured data (e.g., diagnoses, medications, lab results) and unstructured clinical notes (e.g., detailed descriptions of patient history and conditions), both of which are critical for answering medical-related questions. However, combining structured and unstructured data requires overcoming challenges such as limited relationships in structured data and the complexity of processing unstructured text.

"What medication is the patient with an admission ID of 105104 taking for Hypoxemia?" The patient of interest being diagnosed with "Hypoxemia" They also contain the list of medicines prescribed to the patient of interest. However, the tables may contain records (medicines) prescribed to the patient for non-Hypoxemia related conditions. The **PRESCRIPTIONS** table only displays the medications being taken, if they are determined to find a clear relation between medications and diseases, that relies on clinical notes to supplement background information on prescriptions and establish links. In this scenario, the answer from unstructured data for such missing relations is more reliable since the answer is directly available in the clinical notes. They expect a modality selection network to determine whether a query should be answered using structured or unstructured data.

The authors invented DrugEHRQA, the first multimodal QA dataset that integrates structured tables and unstructured clinical text. The dataset includes, "**Natural language questions**", "**Corresponding SQL queries**", "**Answers from both structured tables and unstructured text**" and "**Automatically generated and human-verified multimodal answers**." Each template was designed to address medicine-related topics such as drug dosage, strength, route, and form. To ensure a range of complexity, SQL query templates were categorized into four levels—easy, medium, hard, and very hard—based on attributes like the number of conditions in the WHERE clause, aggregation operators, and the complexity of joins or nesting. For instance, a simple query with one WHERE condition is classified as easy, while a nested query involving multiple joins is categorized as very hard. NLP methods were key to ensuring that the natural language questions matched their SQL counterparts both semantically and syntactically.

Unstructured clinical notes from the MIMIC-III database were leveraged for extracting drug-related attributes, using annotations from the 2018 Adverse Drug Event (ADE) dataset and the Medical Extraction Challenge dataset in the n2c2 (National NLP Clinical Challenges) repository. Six attributes, including drug strength, form, dosage, and route, were extracted for 505 patient discharge summaries. Using these attributes, NL question-answer pairs were generated by filling in placeholders with relevant patient admission IDs and drug-related data. This allowed for the creation of diverse, context-specific QA pairs, despite the challenges posed by unstructured text.

Answers were also retrieved from structured MIMIC-III tables, including PRESCRIPTIONS, DIAGNOSES_ICD, and D_ICD_DIAGNOSES. Slot-filling techniques were used to populate place-

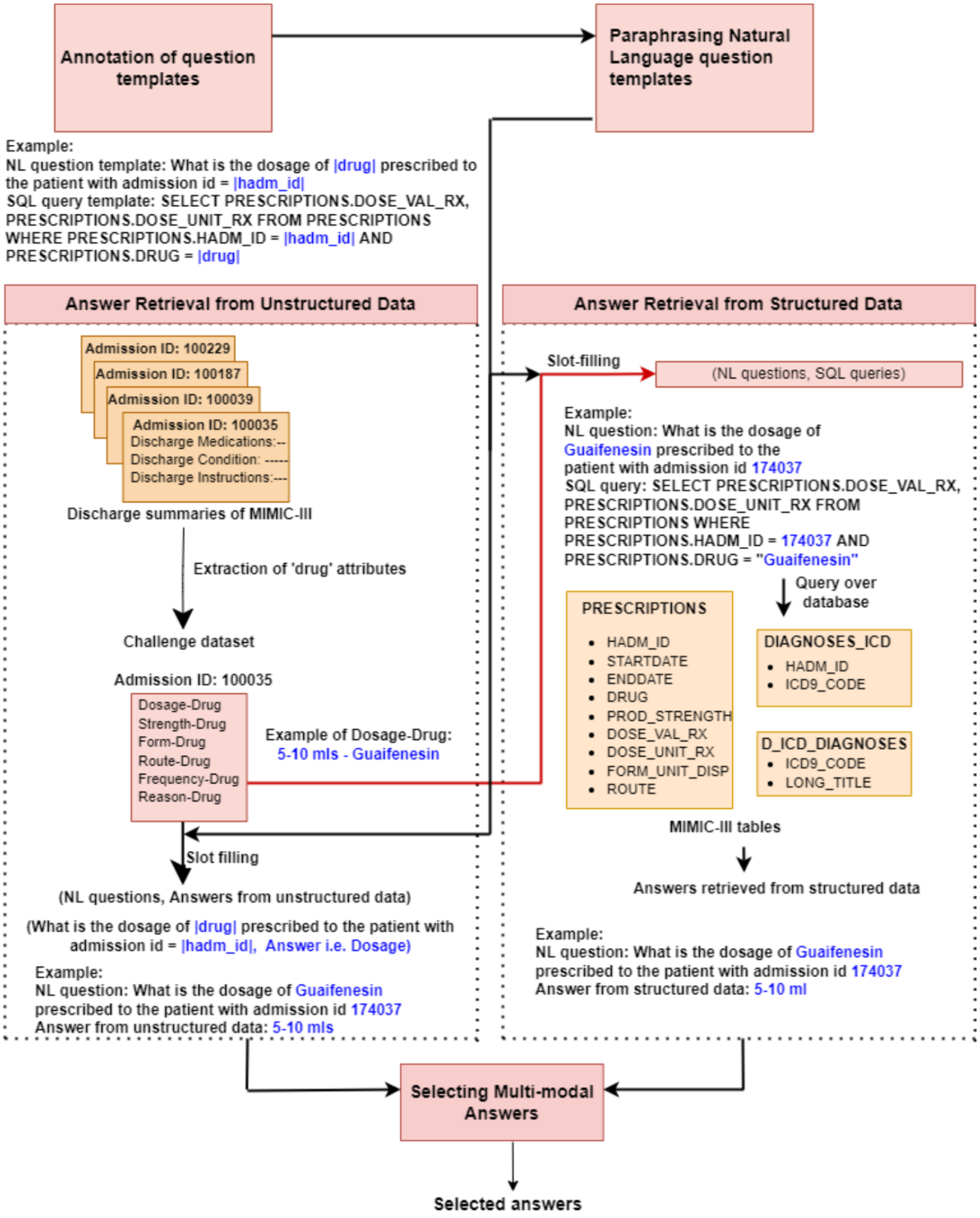
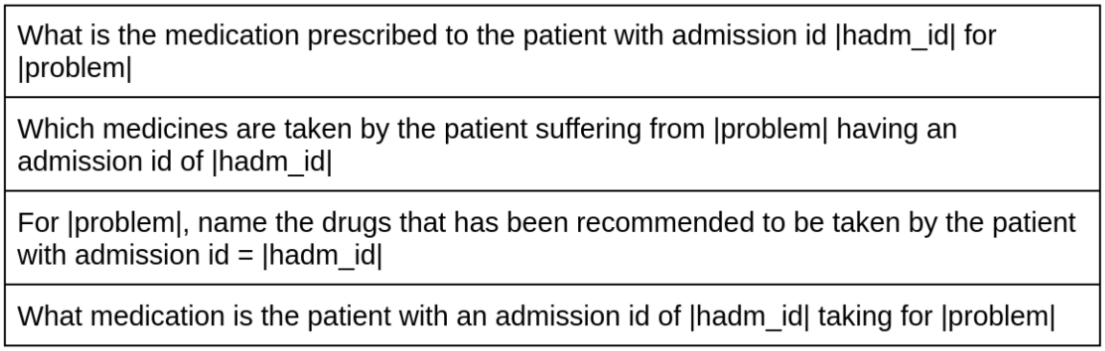

Annotation of question templates
Paraphrasing Natural Language question templates
Example:
NL question template: What is the dosage of |drug| prescribed to the patient with admission id = |hadm_id|
SQL query template: SELECT PRESCRIPTIONS.DOSE_VAL_RX, PRESCRIPTIONS.DOSE_UNIT_RX FROM PRESCRIPTIONS WHERE PRESCRIPTIONS.HADM_ID = |hadm_id| AND PRESCRIPTIONS.DRUG = |drug|
Answer Retrieval from Unstructured Data
Answer Retrieval from Structured Data
Admission ID: 100229
Admission ID: 100187
Admission ID: 100039
Admission ID: 100035
Discharge Medications:--
Discharge Condition: -----
Discharge Instructions:---
Discharge summaries of MIMIC-III
Extraction of 'drug' attributes
Challenge dataset
Admission ID: 100035
Dosage-Drug
Strength-Drug
Form-Drug
Route-Drug
Frequency-Drug
Reason-Drug
Example of Dosage-Drug:
5-10 mls - Guaifenesin
Slot filling
(NL questions, Answers from unstructured data)
(What is the dosage of |drug| prescribed to the patient with admission id = |hadm_id|, Answer i.e. Dosage)
Example:
NL question: What is the dosage of Guaifenesin prescribed to the patient with admission id 174037
Answer from unstructured data: 5-10 mls
Slot-filling
(NL questions, SQL queries)
Example:
NL question: What is the dosage of Guaifenesin prescribed to the patient with admission id 174037
SQL query: SELECT PRESCRIPTIONS.DOSE_VAL_RX, PRESCRIPTIONS.DOSE_UNIT_RX FROM PRESCRIPTIONS WHERE PRESCRIPTIONS.HADM_ID = 174037 AND PRESCRIPTIONS.DRUG = "Guaifenesin"
Query over database
PRESCRIPTIONS
HADM_ID
STARTDATE
ENDDATE
DRUG
PROD_STRENGTH
DOSE_VAL_RX
DOSE_UNIT_RX
FORM_UNIT_DISP
ROUTE
DIAGNOSES_ICD
HADM_ID
ICD9_CODE
D_ICD_DIAGNOSES
ICD9_CODE
LONG_TITLE
MIMIC-III tables
Answers retrieved from structured data
Example:
NL question: What is the dosage of Guaifenesin prescribed to the patient with admission id 174037
Answer from structured data: 5-10 ml
Selecting Multi-modal Answers
Selected answers

Figure 1: Dataset generation framework of DrugEHRQA. There are five steps in this process: (1) annotation of question templates, (2) answer retrieval from unstructured clinical notes, (3) answer retrieval from structured EHR Data, (4) paraphrasing natural language question templates, and (5) selecting multi-modal answers. Note that the challenge dataset mentioned in the figure refers to the ”The 2018 Adverse Drug Event (ADE) dataset and Medication Extraction Challenge dataset” present the n2c2 (National NLP Clinical Challenges) repository.



| |
|---|
| What is the medication prescribed to the patient with admission id \|hadm_id\| for \|problem\| |
| Which medicines are taken by the patient suffering from \|problem\| having an admission id of \|hadm_id\| |
| For \|problem\|, name the drugs that has been recommended to be taken by the patient with admission id = \|hadm_id\| |
| What medication is the patient with an admission id of \|hadm_id\| taking for \|problem\| |

Figure 2: Example of various paraphrases of a natural language question template in the DrugEHRQA dataset.

holders in the SQL query templates, enabling the retrieval of structured data answers. For example, a question template asking about a drug's dosage was matched with corresponding admission IDs and drug names to query the PRESCRIPTIONS table. However, not all questions yielded answers due to the limitations of structured records, resulting in empty responses for certain queries. This dual-mode approach ensured that the dataset included both structured and unstructured data answers.

To enhance the dataset's linguistic diversity and realism, NLP-based paraphrasing techniques were applied to the natural language questions. Each of the nine templates was expanded with three additional paraphrases, creating a total of four variations per question type. **This approach captured the variability in how patients and clinicians phrase questions, improving the dataset's robustness for real-world QA applications.** By randomly associating SQL queries with these paraphrased NL questions, the framework ensured semantic diversity while maintaining structural consistency.

In conclusion, the authors believe that the DrugEHRQA dataset represents a significant contribution to the field of multimodal QA for EHRs, filling gaps in existing datasets. In the future, they attempt to jointly train models for structured and unstructured data and explore how answers from one modality can complement the other. DrugEHRQA opens new research avenues for multimodal QA in EHR systems, addressing critical challenges and laying the groundwork for more advanced medical QA systems.

### 3.3 Extracting key information from Medical Conversations

*”Summarizing Medical Conversations via Identifying Important Utterances”*[5]

When a conversation is too long or the key information is scattered in it, one could hardly find the essential contents or misread them in many cases. As a result, the summarization of the conversation, especially for problem statements and treatment recommendations, is an important task to help new patients locate useful information to address their medical concerns.

| | Role | Utterance | Translation | Tag |
|---|---|---|---|---|
| Conv. | P | 胆碱能性荨麻疹怎么治疗 | How to treat cholinergic urticaria and measles | PD |
| | D | 这种情况多长时间了？用什么治疗过? | How long has this condition last? What treatment have you used? | OT |
| | P | 好长时间了，之前治疗过，中西药都吃过就没治好 | It has been a long time. I have taken both Chinese and Western medicine, but it is not working. | OT |
| | D | 主要是避免诱因。胆碱能性荨麻疹要保持身体凉爽、避免出汗、避免精神紧张、进食热饮或酒精饮料等。 | You need to avoid triggers of cholinergic urticaria. Keep your body cool and avoid sweating, mental stress, hot drink, alcoholic beverages, etc. | DT |
| | P | 那怎么样能根治呢 | How can it be cured? | OT |
| | D | 目前医疗上，没有明确的根治方法。 | At present, there is no clear cure for this disease. | OT |
| | D | 内服药物之外还可以中药外洗.这个方法也有一定的效果。蚕砂、苦参、芒硝、白矾、荆芥准备二十克，把这些药一起煎了进行外洗，一天二次。 | In addition to taking medicines, you can also wash the skin with Chinese medicine, which has some effect. Use the decoction of Silkworm litter, Sophora flavescens, Glauber's salt, alum, and Nepeta, 20 grams for each, to wash your skin twice a day. | DT |
| SUM1 | | 胆碱能性荨麻疹怎么治疗 | How to treat cholinergic urticaria and measles | |
| SUM2 | A | 胆碱能性荨麻疹要保持身体凉爽，避免出汗，避免精神紧张、进食热饮或酒精饮料等。内服药物之外还可以通过中药进行外洗。这个方法也有一定的效果。蚕砂、苦参、芒硝、白矾、荆芥准备二十克,把这些药一起煎了进行外洗，一天二次。 | You need to avoid triggers of cholinergic urticaria. keep your body cool and avoid sweating, mental stress, hot drink, alcoholic beverages, etc. In addition to taking medicines, you can also wash the skin with Chinese medicine, which has some effect. Use the decoction of Silkworm litter, Sophora flavescens, Glauber's salt, alum, and Nepeta, 20 grams for each, to wash your skin twice a day. | |
| | B | 口服脱敏药物，同时避免诱因。胆碱能性荨麻疹要保持身体凉爽，避免出汗，避免精神紧张、进食热饮或酒精饮料等。 | You need to take desensitization drugs and avoid triggers. You should keep the body cool, avoid sweating, mental stress, hot drink, alcoholic beverages, etc. | |

Figure 3: An example of a conversation and its different types of summaries. *P* and *D* stand for speaker roles, i.e., patient and doctor, and *PD*, *DT*, and *OT* in the last column refer to the utterance tags for problem description, diagnosis or treatment, and others, respectively. *SUM1* is a summary of the medical problem from the patient; *SUM2* is a summary of the diagnosis and treatment from the doctor. The English translation is not part of the corpus, which is added as a reference.

| Data | Total # | | Avg. # per Case | | | | | | Avg. Length | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Case | Utt. | Utt. | PD | DT | OT | P | D | Utt. | SUM1 | SUM2 |
| All | 44,983 | 855,403 | 19.0 | 1.3 | 4.5 | 13.3 | 9.7 | 9.3 | 16.4 | 22.8 | 113.0 |
| Train | 35,987 | 684,611 | 19.0 | 1.3 | 4.5 | 13.3 | 9.8 | 9.3 | 16.4 | 22.9 | 112.8 |
| Test | 8,996 | 170,792 | 19.0 | 1.3 | 4.5 | 13.2 | 9.7 | 9.3 | 16.5 | 22.8 | 114.0 |

| | A Only | BOTH |
|---|---|---|
| Train | 33,306<br>(92.5%) | 2,681<br>(7.5%) |
| Test | 8,299<br>(92.3%) | 697<br>(7.7%) |

Figure 4: *SUM1* describes the medical problem that the patient has; *SUM2* summarizes the doctor's diagnosis or treatment recommendations; *Avg. length* is the average number of Chinese characters in an utterance, SUM1, or SUM2.

In this paper, to model the input, the proposed model follows the typical hierarchical structure in which the tokens and utterances are encoded by separate encoders and hierarchically stacked. Then a tagger is attached at the utterance-level to predict PD/DT/OT labels. (*P* and *D* stand for **speaker roles**, i.e., **patient and doctor**, and *PD*, *DT*, and *OT* refer to the utterance tags for **problem description, diagnosis or treatment, and others**, respectively.) Afterwards, we concatenate the utterances labeled by PD and DT to generate the summary of medical problems and doctor's diagnosis, respectively. To further enhance our model, authors adopt memory networks to incorporate the information from relevant utterances in the conversation. Therefore, this proposed model is a ***hierarchical encoder-tagger*** (HET) with the memory module applied between the token-level and utterance-level encoders, while NNs can capture contextual information, authors face challenges in handling long-term dependencies:

- **RNN/LSTM**: As the sequence length increases, the hidden states may forget earlier contextual information, leading to the ”memory decay” problem, for example, given the sentence:

  *”I love this movie!”*

  The RNN processes each word sequentially. The LSTM, a type of RNN, helps preserve long-term dependencies by updating its hidden state, ensuring that information like ”love” influences the final sentiment prediction. However, as the sequence length increases, LSTM may struggle with ”memory decay,” forgetting earlier context, which can degrade performance on longer texts.

- **Transformer**: Although it computes global attention for all tokens and captures long-range dependencies, its computational complexity grows quadratically with sequence length, which can hinder performance.

**Improvements with HET:**

To obtain a representation of each input utterance, BERT is used as the token-level encoder (TE), with the “[CLS]” token's hidden vector representing the utterance. The output is concatenated with a memory module's information to form a combined vector, which is then processed by an utterance-level encoder (UE).

(next encoded hidden vector = encoded hidden vector + a memory module)
$\mathrm{h}_i = h_i + \mathrm{a}_i$

For instance, LSTM is applied to encode the sequential utterance representations, where the hidden states at each step, oi, are updated by the previous state and the current utterance representation.

(step-wise state for utterances = LSTM(previous step-wise state, h is used as the input to the UE))
$\mathrm{o}_i = LSTM(o_{i-1}, h_i)$

(trainable matrix W and bias vector b is used to align o$_i$)
$o_i = W * o_i + b$

After encoding, a tagger layer maps the output to the label space, followed by a softmax or CRF to generate the final tags for tasks like summarizing patient problems and diagnoses.

In summary, a hierarchical encoder-tagger model (HET), enhanced with a memory module, tags each utterance as a problem statement or treatment recommendation. The labeled utterances are concatenated to form summaries. Experimental results validate the approach on a Chinese medical dataset. Future work aims to extract key information from conversation summaries, improving references for new patients with similar medical issues.

## 3.4 Take a further step: predict medication outcome and changes

*"Prediction of Key Patient Outcomes from Sentence and Word of Medical Text Records"*[6]

**Variance-Oriented Clinical Pathway Framework** This pathway consists of three levels, aiming to clearly document patients' treatment processes and deviations from standards, firstly, outcome layer, patient outcome goals, such as recovery or symptom relief, and **assessment layer**, measures to determine whether the patient meets the outcome standards, such as pain levels or blood pressure ranges, finally, **task layer**, specific medical or nursing actions, such as medication, examinations, or monitoring.Doctors or nurses document the specific tasks performed (Task) and their assessments of the patient's condition (Assessment).

The desired goal is to find when a patient's condition does not meet the standards set by the Assessment layer, this deviation is documented in the **outcome layer**, for example, goal is effective pain management. **Assessment** is conducted and pain index should be below a specific value. If the pain index exceeds the threshold, this anomaly is recorded. When time is sensitive and limited, medical staff can quickly understand patient anomalies through variance records. Variance information can guide changes in treatment interventions to improve the patient's condition. Constructing an electronic record system facilitates standardized use and analysis across different medical institutions.

But how to achieve the goal? The first step involves extracting meaningful words from patient admission records using the GETA2 system, which performs morphological analysis. This NLP technique breaks down medical texts into their base forms, such as root words, which are crucial for creating accurate feature vectors. The system uses a medical dictionary consisting of around 80,000 words, ensuring that the extracted terms are relevant and standardized. Once the words are extracted, they are vectorized, converting the text into numerical features that can be input into the SVM model for classification. This step is fundamental in NLP as it transforms unstructured text into a structured format suitable for machine learning.

After converting the text to vectors, each patient's record is classified based on whether it contains certain clinical outcomes, such as the presence of pain or neuropathy worsening. If a record contains information related to these outcomes, it is labeled as a **positive** example, while records that do not show such outcomes are labeled as **negative**. This classification process relies heavily on NLP, as it requires the system to understand and identify specific keywords or phrases within the

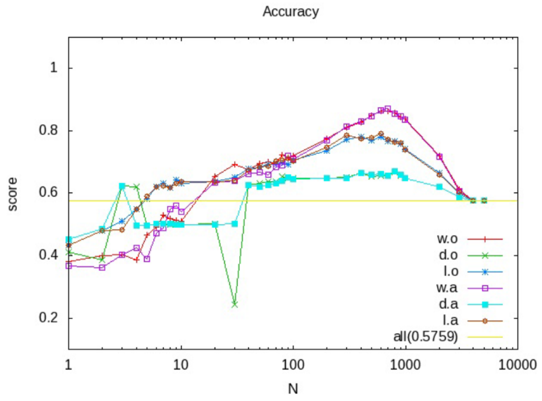


Figure 2: Accuracy(Pain)

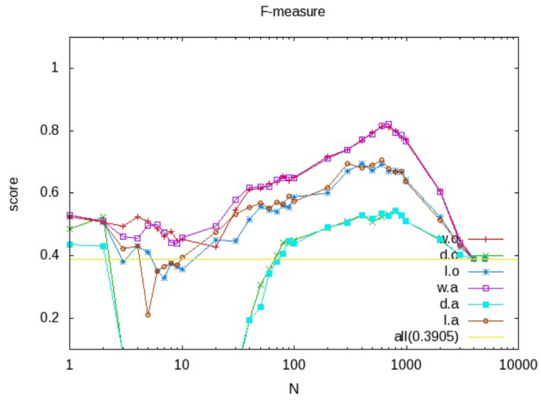


Figure 3: F-measure(Pain)

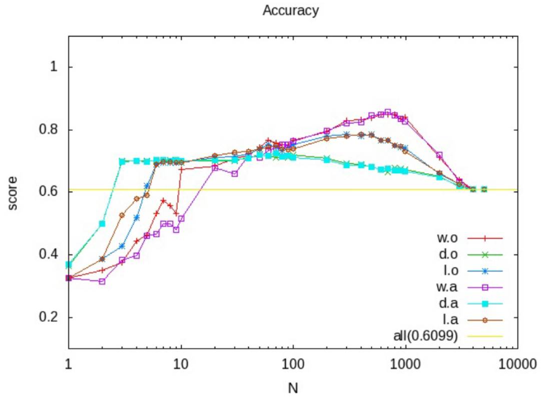


Figure 4: Accuracy(Neuropathy worsening)

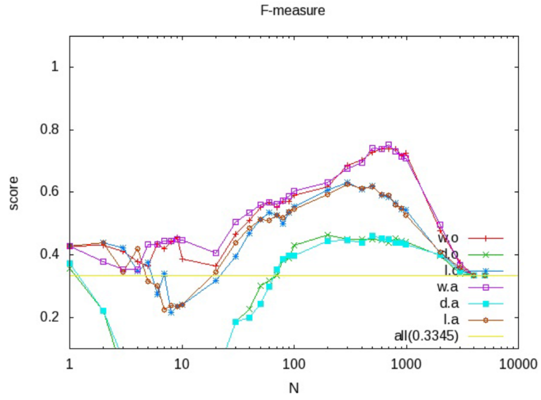
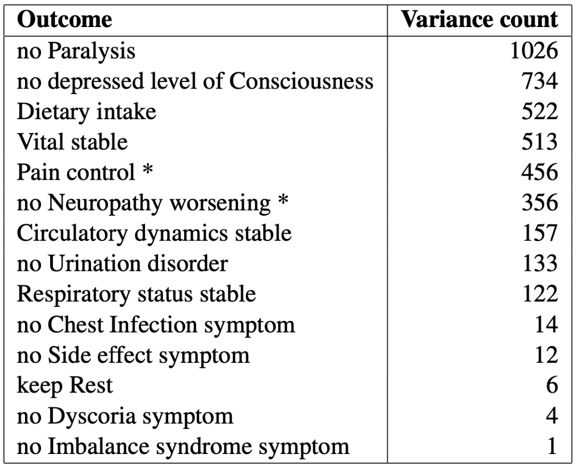


Figure 5: F-measure(Neuropathy worsening)

| **Outcome** | **Feature words** | **Feature setences** |
|---|---|---|
| Pain | dizzy(132), hypalgesia(14), aneurysm(181)*, headache(81)*, nifedipine(67), nausea(61), fibroid(27), right angular(53), calcification(68), pravastatin(23), hemianopsia(94) | dizzy when body move, severe nausea.<br>Feeling badness, a headache and dizziness appear suddenly.<br>found aneurysm in cavernous sinus.<br>anamnesis: fibroid, gallstone(postoperation), high blood pressure, irregular pulse. |
| Neuropathy worsening | paralysis(205), right face(106), renal failure(60), right knee(10), hypalgesia(14), sick sinus syndrome(15), Right facial paralysis(88), difficulty talking(165), flexion(65)* | When getting up, the paralysis of a right hand finger appeared and was also felt by the right face again.<br>With the paralysis senses in mandibular nerve area of right face.<br>The paralysis sense of the right face, the right forearm and the right thigh back side.<br>Difficulty talking appeared.<br>Forgetfulness and slow talking appeared. |

Figure 5: Feature words and sentences by SVM (* possibility or impossibility, presence or absence)



| Outcome | Variance count |
|---|---:|
| no Paralysis | 1026 |
| no depressed level of Consciousness | 734 |
| Dietary intake | 522 |
| Vital stable | 513 |
| Pain control * | 456 |
| no Neuropathy worsening * | 356 |
| Circulatory dynamics stable | 157 |
| no Urination disorder | 133 |
| Respiratory status stable | 122 |
| no Chest Infection symptom | 14 |
| no Side effect symptom | 12 |
| keep Rest | 6 |
| no Dyscoria symptom | 4 |
| no Imbalance syndrome symptom | 1 |

Figure 6: Outcome in clinical pathway of Cerebral Infarction (*: target in this study)

text that correspond to particular health outcomes. By identifying and labeling these records, the system can later learn to classify new records accurately.

Once the classification labels are applied, the next step is feature selection, which is a key task in NLP and machine learning. The SVM model is used to evaluate the **importance** of each word in predicting the health outcomes. Specifically, the model computes an SVM score for each word by treating it as if it appeared alone in a document. This score reflects how strongly each word contributes to classifying the record as positive or negative for a particular outcome. In addition to the raw SVM score, two additional metrics are calculated: one involves multiplying the SVM score by the word's document frequency (df), and the other takes the logarithm of the document frequency before multiplying it with the score. These measures are referred to as ”w.o,” ”d.o,” and ”l.o,” respectively, and are used to assess the relevance of each word more precisely. This step demonstrates how NLP can help determine which words are most indicative of specific outcomes by evaluating their statistical significance within the dataset.

With the scores calculated, the next step is to **identify the most significant feature words and feature sentences** associated with the clinical outcomes, such as pain or neuropathy worsening. Sentences containing high-scoring words are selected as typical examples of records associated with these outcomes. The feature words extracted from the sentences help interpret the meaning of the text in a way that makes the model's decision-making process more transparent. For example, words like “dizzy,” “headache,” and “nausea” appear frequently in sentences related to pain, while terms like “paralysis” and “difficulty talking” are commonly seen in sentences associated with worsening neuropathy. By highlighting these words in context, the authors provide insights into how certain terms correlate with specific health conditions, which can aid in clinical decision-making.

Finally, feature selection is refined using a performance evaluation based on various metrics, including accuracy and F-measure. Different sets of top words are selected based on their importance, and the model is retrained using these selected features. The evaluation shows that as the number of selected words increases, the accuracy and F-measure also improve, indicating that careful feature selection significantly enhances the model's predictive power.

**Algorithm 1** get_conjunction(head)

```
$acc \leftarrow []$, $list\_heads \leftarrow [head]$
while $list\_heads \neq []$ do
    $new\_heads \leftarrow []$
    for $h$ in $list\_heads$ do
        $children \leftarrow$ children of $h$ with dependency tags "conj" or "ccomp"
        if $children \neq []$ then
            append $children$ to $new\_heads$ and $acc$
        end if
    end for
    $list\_heads \leftarrow new\_heads$
end while
return $acc$
```

Figure 7: Algo 1 demonstrates how to extract phrases following conjunctions (such as ”and” or ”or”) based on syntactic dependency structures, which is crucial for capturing multiple medication events.

### 3.5 Take a further step: relation between medication and prescription

*”Context-aware Medication Event Extraction from Unstructured Text”*[2]

Accurately documenting a patient’s medication history is critical to providing effective healthcare. However, unstructured clinical narratives present challenges in extracting such information, especially when sentences simultaneously exhibit overlapping event labels like ”undefined” and ”disposition.” To address these challenges, the authors developed a preprocessing pipeline supported by advanced algorithms to enhance data consistency and the accuracy of information extraction and classification. For example, algo 1, ”get$_{c}onjunction(head)”andalgo2,”get_{c}hunks(sentence)$”.

The pipeline begins with text standardization to address inconsistencies within the clinical narratives. Data from the Contextualized Medication Event Dataset (CMED) is cleaned to ensure uniformity and eliminate noise, preserving patterns essential for identifying medication-related information. This preprocessing step ensures a robust foundation for downstream tasks, such as identifying and classifying medication mentions.

Medication mention identification is approached as a sequence tagging task, leveraging domain-specific pretrained models like BioBERT and Bio+Clinical BERT. These models, fine-tuned on CMED data, predict the beginning, inside, and outside (BIO) positions of medication-related phrases, enabling accurate localization of multi-token mentions. To further enhance recognition performance, additional training data from the DDI Extraction 2013 corpus is integrated, increasing the diversity of the training set and improving generalization. Comparisons reveal that using both CMED and DDI data boosts F1 scores, demonstrating the value of augmented datasets.

The pipeline also incorporates negation detection to interpret contextual nuances. NegspaCy, a specialized tool for recognizing negated terms in clinical text, is applied to determine whether a medication was prescribed, discontinued, or merely mentioned, for example, ”The patient was not prescribed aspirin.”. While negation occurs in only 2% of CMED samples, accurately detecting it is

**Algorithm 2** get_chunks(sentence)

```
$doc \leftarrow$ parse $sentence$ using spaCy, $chunks \leftarrow []$
for $sent$ in $doc$ do
    $conj\_phrases \leftarrow$ get coordinated conjunction phrases from $sent$'s root using get_conjunction(head)
    for $head$ in $conj\_phrases$ do
        append $head$'s subtree to $chunks$
    end for
end for
sort $chunks$ in ascending order of length
$seen \leftarrow$ empty set, $trimmed\_chunks \leftarrow []$
for $chunk$ in $chunks$ do
    $c2 \leftarrow$ list of unconsumed tokens in $chunk$
    update $seen$ set with indices of tokens in $c2$
    $c3 \leftarrow$ longest continuous sequence of tokens in $c2$
    append longest sequence in $c3$ to $trimmed\_chunks$
end for
$output \leftarrow []$
for $phrase$ in $trimmed\_chunks$ do
    remove any conjunctions at the beginning or end of $phrase$
    join the tokens in $phrase$ to form a string
    remove any leading or trailing commas from the string

    append the string to $output$
end for
sort $output$ in the original order of phrases in $sentence$
return $output$
```

Figure 8: Algo 2 breaks these sentences into smaller segments, each containing only one medication event. For example, the sentence ”Started lisinopril 10 mg p.o. daily, substituted for diltiazem” would be decomposed into two simpler clauses: ”started lisinopril 10 mg p.o. daily” and ”substituted for diltiazem,” making it easier to assign events and medications clearly.

(a) “The patient’s daily dose of **furosemide** was increased from 40mg to 80mg.
and then reduced to 60mg daily.”
LABELS: *increase*, *decrease*

(b) “The healthcare provider started the patient on a new regimen of **metformin** and discontinued the use of **pioglitazone**.”
LABELS: *start*, *stop*

(c) “The healthcare provider instructed the patient to take **acetaminophen**
if their fever rises above 100 degrees.”
LABELS: *conditional*

Figure 9: Examples from clinical notes where (a) one drug mention indicates two events with opposite action labels, and (b) two drug mentions, each with their own action labels. Also, (a) and (c) have grammatically valid sentences up until the line break, but the sentence con- tinues. Stopping at the line break will miss the language responsible for the decrease and conditional labels.

essential for distinguishing negated from affirmative mentions, such as identifying ”not prescribed aspirin” as negated. This process is further enhanced by integrating Med7, a medical Named Entity Recognition tool, to refine the detection of negated medication mentions.

Beyond identifying medication mentions, the pipeline classifies events and associated attributes like action, actor, negation, certainty, and temporality. This task involves the use of pretrained Transformer-based models, including Bio+Clinical BERT and RoBERTa, fine-tuned on CMED for multi-class classification. Algorithms are designed to process these attributes as distinct tasks, ensuring detailed and structured representations of medication-related events.

The system's evaluation uses both strict and lenient matching criteria for medication mention extraction. Strict matches require exact alignment between predicted and true mentions, whereas lenient matches allow for partial overlaps. BioBERT achieves the highest F1 score of 0.95 under both criteria, although Bio+Clinical BERT performs better in terms of precision, highlighting the trade-off between precision and recall. For event and attribute classification, Bio+Clinical BERT and RoBERTa demonstrate strong macro-average F1 scores, underscoring their suitability for clinical text analysis.

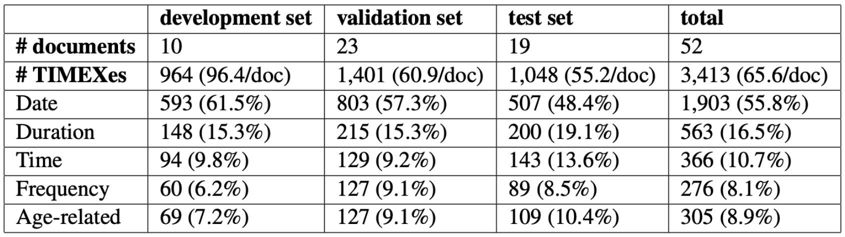
Overall, the combination of preprocessing techniques, integration of diverse datasets, and advanced algorithms provides a robust framework for extracting and classifying medication-related information. This pipeline effectively tackles the challenges posed by unstructured clinical text, contributing to a better understanding of patient treatments.

### 3.6 Take a further step: Time Onset Insights

*”Time Expressions in Mental Health Records for Symptom Onset Extraction”*[4]

For psychiatric disorders such as schizophrenia, prolonged periods of time without treatment are associated with worse intervention outcomes. Electronic Health Records (EHR) are valuable for retrospective clinical research. However, most data is stored as unstructured text and cannot be directly calculated. Natural Language Processing (NLP) methods can extract data from mental health records, identify symptoms and treatments, and determine their first occurrence times. When applying NLP techniques to the clinical domain, one crucial task involves the identification of *temporal information*. In general, for temporal information mod- eling, three different steps are typically outlined: (1) the identification of relevant concepts, such as symptoms (*hallucinations*) and treatments (*Clozapine*), (2) the identification of time expressions, and (3) the identification of temporal relations between entity pairs.

Unstructured data (e.g., clinical notes) are often written in natural language and may include doctors' descriptions, patient histories, and symptom records. These lack fixed fields or labels to regulate their organization. Unlike structured data, clinical notes are not constrained by predefined frameworks (e.g., rows or columns in data tables). Clinical notes may include various expressions, such as abbreviations, medical terms, misspellings, or unique recording styles of different physicians, complicating data interpretation. Natural language expressions cannot be stored in a standardized format like structured data (e.g., ICD diagnostic codes). Unstructured data may mix



| | development set | validation set | test set | total |
|---|---|---|---|---|
| **# documents** | 10 | 23 | 19 | 52 |
| **# TIMEXes** | 964 (96.4/doc) | 1,401 (60.9/doc) | 1,048 (55.2/doc) | 3,413 (65.6/doc) |
| Date | 593 (61.5%) | 803 (57.3%) | 507 (48.4%) | 1,903 (55.8%) |
| Duration | 148 (15.3%) | 215 (15.3%) | 200 (19.1%) | 563 (16.5%) |
| Time | 94 (9.8%) | 129 (9.2%) | 143 (13.6%) | 366 (10.7%) |
| Frequency | 60 (6.2%) | 127 (9.1%) | 89 (8.5%) | 276 (8.1%) |
| Age-related | 69 (7.2%) | 127 (9.1%) | 109 (10.4%) | 305 (8.9%) |

Figure 10: TIMEXannotationresults:prevalenceoftypesinourcorpus.

various information types, such as symptom descriptions and treatments, embedding essential details within narrative texts. Imagine analyzing a clinical note like this:

”The patient reported hearing voices starting about six months ago and increased stress since turning 30.”

The process begins with manual annotation, where a medical student uses the eHOST tool to tag time expressions. For example, ”six months ago” might be labeled as a Duration TIMEX, while ”since turning 30” could be tagged as an Age-related TIMEX. Guidelines are refined throughout the process to account for vague phrases commonly found in mental health records, such as ”recently” or ”a few weeks ago.” Automated tools like SUTime and HeidelTime are then applied to extract these time expressions. These tools are enhanced with customized rules to better handle the unique language in clinical notes. Finally, metrics such as precision, recall, and F1 scores are used to evaluate the tool’s performance, ensuring accurate detection of timelines related to psychosis symptoms.

To evaluate the corpus quality, inter-annotator agreement (IAA) was calculated using precision, recall, and F1 scores, following metrics from i2b2(Informatics for Integrating Biology  the Bedside) 2012 and Clinical TempEval 2016. Annotators compared their entities as gold references and system outputs. Another libraries, such as SUTime and HeidelTime, recognizing and normalizing time were assessed for precision, recall, and F1 scores by defining true positives (TP), false negatives (FN), and false positives (FP), ensuring accurate evaluation of their ability to extract TIMEXes from text. This comprehensive approach quantified annotation agreement and system performance in entity extraction tasks.

# 4 Conclusion

After analyzing these six papers, I believe the selected papers can be categorized into ”classic classification” and ”innovative application.” For example, classification refers to human intervention, where researchers use medical expertise to interpret and broadly define concepts. On the other hand, breakthroughs are scientifically supported, involving extensive reading to manage time more efficiently and identify new advancements. I have learned a lot of existing technologies and techniques to parse and extra data from these selected papers, for example, spaCy and negspaCy, and I also have seen how authors developed specific tag entity to remark unstructured data and provided unseen but precious medical histories of patients. In my own opinion, machine learning requires a complex combinations of steps to extract and analyze data, and it heavily relied on statistics. Back to my original research topic: **How NLP Extracts Key Health Information from Unstructured Medical Reports: Challenges in Language Understanding and Accuracy**, preprocess the raw data is a very important step, although it's very tedious and the amount is overwhelming. I appreciate those efforts and what have been done and I did learn a lot from these practical experiences in medical field.

## References

[1] Shuntaro Yada, Ayami Joh, Ribeka Tanaka, Fei Cheng, Eiji Aramaki, and Sadao Kurohashi(2020).*Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases.* In Proceedings of the Twelfth Language Resources and Evaluation Conference. pages 4565–4572, Marseille, France. European Language Resources Association.

[2] Noushin Salek Faramarzi, Meet Patel, Sai Harika Bandarupally, and Ritwik Banerjee(2023).*Context-aware Medication Event Extraction from Unstructured Text.* In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 86–95, Toronto, Canada. Association for Computational Linguistics.

[3] Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia(2020). *Summarizing Medical Conversations via Identifying Important Utterances* In Proceedings of the 28th International Conference on Computational Linguistics, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[4] Takanori Yamashita, Yoshifumi Wakata, Hidehisa Soejima, Naoki Nakashima, and Sachio Hirokawa(2016)*Prediction of Key Patient Outcome from Sentence and Word of Medical Text Records* In Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), pages 86–90, Osaka, Japan. The COLING 2016 Organizing Committee.

[5] Adachi, Yusuke Onimura, Naoya Yamashita, Takanori Hirokawa, Sachio(2016)*Standard measure and SVM measure for feature selection and their performance effect for text classification* 262-266. 10.1145/3011141.3011190.

[6] Natalia Viani, Lucia Yin, Joyce Kam, Ayunni Alawi, André Bittar, Rina Dutta, Rashmi Patel, Robert Stewart, and Sumithra Velupillai(2018) *Time Expressions in Mental Health Records for Symptom Onset Extraction* In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, pages 183–192, Brussels, Belgium. Association for Computational Linguistics.

[7] Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang(2022) *DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records For Medicine Related Queries* In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1083–1097, Marseille, France. European Language Resources Association.