Enhancing Medical Dialogue Summarization with Key Word and Medical Entity Extraction

Tingting, Li; Michelle, Sinani; Agnese, Minazzo
DATASCI 266 Natural Language Processing with Deep Learning - Final Project (Summer 2024)

Abstract
Accurate documentation of patient-clinician interactions is essential for high-quality healthcare. The increasing volume of medical dialogues necessitates efficient systems to generate precise and coherent summaries, relieving clinicians from the burden of extensive note taking. This project explores enhancing medical dialogue summarization through keyword and medical entity extraction. We utilize the MTS-Dialog dataset for short dialogues and the ACI Bench and Dialogue_G for long dialogues, comparing Bio-BERT, Medical NER and KeyBERT for keyword and entity extraction. Our summarization approach involves fine-tuning LED_LARGE and Flan-T5. The evaluation metrics include ROUGE Scores and Precision, Recall, and F1 Score. By integrating these techniques, we aim to enhance the presence of critical medical terminology in summaries, thereby improving clinical documentation practices and patient care workflows.

1 Introduction

Accurate summarization of medical dialogues between patients and clinicians is crucial for effective healthcare delivery. These summaries aid in clinical decision-making, facilitate provider hand-offs, and serve as patient references. Manual documentation is time-consuming and contributes to clinician burnout, underscoring the need for automated solutions (Eschenroeder et al., 2021). Despite advancements in Natural Language Processing (NLP), current methods often fail to capture key medical terms, resulting in summaries lacking critical information or context (Nair et al., 2023; Wang et al., 2024).

Our project addresses these challenges by enhancing medical dialogue summarization through the extraction of keywords and medical entities, ensuring crucial information is included for improved accuracy and coherence. We use the ACI Bench and Dialogue_G datasets for long dialogues and the MTS-Dialog dataset for short dialogues. We compare models like Bio-BERT, Medical NER, and KeyBERT for keyword and entity extraction, and fine-tune models such as LED and Flan-T5 for summarization.

Our approach is necessary due to the complex language of medical dialogues, inefficiencies of manual documentation, and scalability challenges in current NLP methods. By integrating advanced models and domain-specific adaptations, we enhance the precision and adaptability of summaries across diverse medical dialogues. This not only streamlines clinical workflows but also ensures compliance with regulatory requirements, improving both care quality and privacy adherence.

Our contributions include developing a system that integrates keyword and medical entity extraction to ensure the inclusion of critical medical terminology. By using a combination of short and long dialogue datasets, we ensure comprehensive coverage of diverse medical dialogues. Implementing and comparing state-of-the-art NLP models, we aim to significantly improve summarization quality. Our evaluation strategy uses ROUGE scores, Precision, Recall, and F1 Score to rigorously assess performance and accuracy.

2. Background Work in Medical Dialogue Summarization and BERT

2.1 Medical Dialogue Summarization

Initially, medical dialogue summarization relied on rule-based systems, limited by their inability to adapt to diverse dialogues and terminologies (Liu et al., 2017). The adoption of RNNs and LSTMs introduced capabilities for handling sequential data more dynamically (Chung et al., 2014). Recently, transformer-based models like BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2019) have significantly improved clinical text comprehension and generation. The Longformer Encoder-Decoder (LED) model (Beltagy et al., 2020) is particularly notable for its efficient management of long texts, enhancing summarization quality.

2.2 Named Entity Recognition in Biomedical Text

The field advanced with BioBERT (Lee et al., 2019), which set new benchmarks in biomedical NER, followed by Medical NER and SciBERT, which were optimized through training on specialized corpora (Alsentzer et al., 2019; Beltagy et al., 2019). Additionally, KeyBERT's keyword extraction represents a major step forward by effectively identifying crucial terms with minimal data requirements (Grootendorst, 2020).

2.3 Integration of NER and Summarization

Recent research demonstrates that integrating NER with summarization significantly enriches summary content by incorporating essential medical terms, thus enhancing both the coherence and the clinical utility of summaries (Zhang et al., 2020). This methodological synergy is pivotal to our project's aim to refine medical dialogue summarization.

3. Methods

This section introduces our dataset, baseline model, and the custom implementation of medical NER tokens used in conjunction with the finalized summaries. Our intuition is that incorporating entities will help the model filter out noise and focus more on the medically relevant content of the dialogue, thereby generating more accurate and comprehensive medical notes.

3.1 Dataset

To train and evaluate our model on diverse data, we used both short and long dialogue-summarization datasets. For long dialogues, we used ACI-BENCH (Yim et al., 2023) from real clinician-patient interactions and Dialogue-G (Wang et al., 2023), consisting of

artificially generated clinician-patient conversations, totaling 1,378 dialogues. For short dialogues, we utilized the MTS-Dialogue dataset (Abacha et al., 2023), which contains 1,501 dialogue pairs. All datasets include dialogues and reference notes. Appendix 1 shows the dialogue and summary lengths for the long and short dialogues. After performing data cleaning, the final data splits followed the MTS-Dialogue ratios: 80:7:13 for training, validation, and testing, respectively.

3.2 Baseline Model

For our baseline model, we fine-tuned two different encoder-decoder models to handle the majority of samples in the dataset, based on the maximum encoder and decoder lengths.

For short dialogues (95% dialogue being 327 words count long, with summaries of 155 words), we fine-tuned the Flan-T5 model (Chung et al., 2022), experimenting with both the Flan-T5 Base and Flan-T5 Large models. We also evaluated the Longformer Encoder-Decoder (LED) base model (Beltagy et al., 2020). The Flan-T5 Large was chosen as our baseline due to its superior generalization on unseen data. Despite fine-tuning the Flan-T5 Base with various adjustments, it did not yield successful results. (See appendix 2 for ROUGE score)

For long dialogues, where 95% of the dialogue length is 966 and the summary length is 485, we fine-tuned and compared the Longformer Encoder-Decoder (LED) base (Beltagy et al., 2020) and LED Large models as LED models are more suitable for handling long text summarization. Finally, we chose the LED Large as our baseline model based on ROUGE score (See appendix 3 for ROUGE score)

3.3 Summarization with medical entity extraction

3.3.1 Entity extraction

In our exploration of named entity recognition mechanisms for medical dialogue, we evaluated three pre-trained NER and keyword models:

1. **KeyBert**: Demonstrated the ability to extract key terminology without requiring explicit labeling.
2. **BioBERT: biobert_chemical_ner and biobert_disease_ner** (hugging face: alvaroalon2): These models are fine-tuned versions of the BioBERT (Lee et al., 2019) model for NER tasks using the BC5CDR-chemicals and BC4CHEMD corpora. They extract entities such as diseases and chemicals.
3. **Medical-NER** (hugging face: Clinical-AI-Apollo): A fine-tuned version of DeBERTa (He et al., 2020) for NER tasks on the PubMed dataset. It extracts entities such as signs and symptoms, biological structures, diagnostic procedures, diseases, and disorders (see Appendix 4 for the full list).

Since BioBERT NER and Medical-NER could only extract tokens and labels, we merged tokens into words and combined them into phrases based on their BIO tagging. We also explored unique and frequency-based entity extraction to determine the best method for identifying key terminology (See Appendix 5 for sample NER extractions.)

3.3.2 Summarization with NER

After extraction, we concatenated the NER or keywords with the original dialogue and constructed a prompt to use as input for fine-tuning the model. Here is an example of the prompt, for long dialogue:

Summarize the following patient-doctor dialogue into 2 sections: "SUBJECTIVE CHIEF COMPLAINT" and "ASSESSMENT AND PLAN". Include all medically relevant information, including family history, diagnosis, past medical and surgical history, immunizations, lab results, and known allergies. Use the following medical and chemical entities extracted from the dialogue to help summarization, but do not overly use them. Entities:(extracted entities). Dialogue:(original dialogue).

In section 3.3.1, we discussed how unique and frequency-based entities from NER and KeyBERT were incorporated. Experiments on long and short dialogues compared entity inclusion approaches. Preliminary ROUGE scores (appendix 6) indicate frequency-based entities enhance medical representation in summaries, especially for long dialogues.

We then fine-tuned encoder-decoder models (Flan-T5 large for short, LED_large for long) with Low-Rank Adaptation to optimize computing power and training time. Models were fine-tuned with NER outputs from KeyBert, BioBERT, and Medical-NER to identify the most accurate representation of medical entities.

For predictions, we used 4-bit quantization to optimize performance and reduce computational load, ensuring efficient inference. Beam search and specific output lengths improved summary quality and coherence.

3.4 Metric for model evaluation

To evaluate the model, we used the ROUGE (Lin, 2004) score (f1, recall and precision). The ROUGE score measures the overlap of n-grams, word sequences, and word pairs between the generated summaries and the reference summaries, providing a robust indication of summary quality.

4 Results

Table 1: Model results

ROUGE Score for Short Dialogue

	ROUGE 1			ROUGE 2			ROUGE L		
	f1	recall	precision	f1	recall	precision	f1	recall	precision
Baseline: Flan-T5 Base	0.2896	0.5049	0.2601	0.1218	0.2385	0.1016	**0.2529**	0.4325	0.2302
Flan-T5 Base + Keybert NER	0.2826	0.5167	0.2499	0.1240	0.2497	0.1031	0.2464	0.4426	0.2203
Flan-T5 Base + Biobert NER	**0.2918**	0.5159	0.2605	**0.1251**	0.2424	0.1060	0.2527	0.4371	0.2290
Flan-T5 Base + Medical NER	0.2818	0.5085	0.2467	0.1199	0.2341	0.1019	0.2423	0.4268	0.2161
Baseline: Flan-T5 Large	**0.3269**	0.5601	0.3269	**0.1553**	0.2885	0.1325	**0.2861**	0.4833	0.2639
Flan-T5 Large + Keybert NER	0.3160	0.5448	0.3160	0.1435	0.2756	0.1288	0.2709	0.4629	0.2627
Flan-T5 Large + Biobert NER	0.3092	0.5356	0.5356	0.1232	0.2648	0.1373	0.2656	0.4541	0.2518
Flan-T5 Large + Medical NER	0.3160	0.5284	0.3091	0.1229	0.2641	0.1387	0.2643	0.4509	0.2534

ROUGE Score for Long Dialogue

	ROUGE 1			ROUGE 2			ROUGE L		
	f1	recall	precision	f1	recall	precision	f1	recall	precision
Baseline: LED_Large	0.5514	0.5006	0.6519	0.3112	0.2833	0.3668	0.3099	0.2804	0.3680
LED_Large + Keybert NER	0.5439	0.4782	0.6712	0.3060	0.2690	0.3776	0.3057	0.2671	0.3806
LED_Large + Biobert NER	0.5542	0.4874	0.6825	0.3177	0.2794	0.3916	0.3173	0.2777	0.3939
LED_Large + Medical NER	**0.5631**	0.5067	0.6799	**0.3306**	0.2980	0.3990	**0.3355**	0.3003	0.4084

4.1 Short Dialogue

In our study, we compared various FLAN-T5 models using ROUGE-1, ROUGE-2, and ROUGE-L metrics for recall, precision, and F1 scores to understand why the baseline FLAN-T5 Large (without NER) outperforms the NER-integrated versions.

The baseline model, trained on a broader dataset, generalizes better, resulting in higher recall, precision, and F1 scores. In contrast, FLAN-T5 Large models with NER, fine-tuned on specialized datasets (Bio, Medical, KeyBert), tend to overfit and perform worse on general datasets. This specialization limits their generalization ability, which is reflected in their slightly lower F1 scores compared to the baseline. Thus, the baseline FLAN-T5 Large model achieves better performance due to its superior generalization in comparison to NER-integrated counterparts.

4.2 Long Dialogue

For long dialogues, LED_Large performed well in summary generation. As shown in Table 1, incorporating NER outputs from BioBERT and Medical-NER, except for keywords, improved performance. The combination of Medical-NER with LED_Large was the best, achieving the highest ROUGE-1, ROUGE-2, and ROUGE-L F1, precision, and recall scores. Medical-NER helped focus on medically relevant content and filter out noise.

To further illustrate this improvement, we identified the sample with the largest increase in the ROUGE-1 F1 score between the baseline model and the LED_Large + Medical-NER model (0.3698 vs. 0.5648). In this case, the Medical-NER model successfully extracted the term 'pain medication' and included it in the prediction, whereas the baseline model failed to identify and incorporate this term in its summary.

Extracting keywords did not enhance summarization. Unlike BioBERT and Medical-NER, the keyword model failed to identify medically relevant entities. For example, for the same dialogue. Bio Ber and Medical NER extracted medical terms like ‘Huntington disease’, ‘memory loss’, but KeyBert extracted 'psychiatrist seeing long', 'going appointment neurologist' (see Appendix 5 for the full dialogue and NER comparison).

4.3 Error Analysis

When working with Flan-T5-large, particularly with the introduction of Named Entity Recognition (NER), we observed that the model became overly tailored to the training data, resulting in poor generalization. This was evident as the baseline Flan-T5-large outperformed the model variants that incorporated medical, bio, and KeyBert NER.

Additionally, the NER components occasionally misidentified entities. For example, in one instance, the text indicated that the patient had no history of cancer, but the summary incorrectly stated that the patient had cancer. Currently, we identify the NER entities and add them to the dialogue without accounting for the entity's status (e.g., present or not present), which can mislead the model. This highlights the need for improved entity recognition and better context understanding within the model.

Furthermore, the introduction of NER caused the model to overly focus on named entities, sometimes at the expense of overall text comprehension, leading to inaccuracies and misinterpretations. For instance, the model tends to generate repetitive summaries centered around the same entity.

5 Conclusion

Our study emphasizes the need for accurate medical dialogue summarization to improve healthcare delivery and reduce clinician burnout. Despite advancements in NLP, current methods often miss key medical terms, leading to incomplete summaries. Our project integrates advanced models and domain-specific adaptations to enhance summary precision and coherence.

We found the baseline Flan-T5 Large model outperformed NER-integrated variants due to its broader generalization. While NER components can aid entity recognition, they often led to overfitting on generalized datasets. For long dialogues, combining LED_Large with Medical-NER yielded the best results by focusing on medically relevant content. However, for short dialogues, NER integration was not beneficial, and baseline models performed better.

Exploring alternative entity recognition methods and more sophisticated context understanding could further improve model performance. Our evaluation using ROUGE, Precision, Recall, and F1 Score provides a robust assessment framework. Utilizing GPT or similar models could significantly enhance the accuracy of automated medical dialogue summaries, improving clinical workflows and patient care.

6 References

Chin-Yew Lin. (2004) “ ROUGE: A Package for Automatic Evaluation of Summaries.” In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jinhyuk Lee., et al. (2019) “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, arXiv:1901.08746.

Add Jacob Devlin., et al (2019) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv: 1810.04805v2, 24th May 2019.

Pengcheng He., et al.(2020) “DeBERTa: Decoding-enhanced BERT with Disentangled Attention”, arXiv:2006.034.

Iz Beltagy., et al. (2020) “Longformer: The Long-Document Transformer ”, arXiv:2004.05150v2.

Edward Hu., et al(2021) “LORA: Low-Rank Adaption of Large Language Models”, arXiv:2106.09685v2.

Eschenroeder, H. C., et al. (2021). "Associations of physician burnout with organizational electronic health record support and after-hours charting." Journal of the American Medical Informatics Association, 28(5): 960-966.

Hyung Won Chung., et al. (2022) “Scaling Instruction-Finetuned Language Models” arXiv:2210.11416.

Giorgi, J., et al. (2023). "WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models." ACL.

Wang, Y., et al. (2023). "Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method." aiXiv:22 May 2023.

Nair, V., et al. (2023). "Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models." arXiv preprint aiXiv:10 May 2023.

Wen-wai Yim., et al.(2023) “Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation" Nature Scientific Data, 05 September, 2023

Asma Ben Abach., et al.(2023) “An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. “
EACL, May 3-5, 2023, Dubrovnik, Croatia.

Wang, H., et al. (2024). "Adapting Open-Source Large Language Models for Cost-Effective, Expert-Level Clinical Note Generation with On-Policy Reinforcement Learning." arXiv preprint aiXiv:5 June 2024.

Hugging face model:

Medical NER:
Clinical-AI-Apollo/Medical-NER : https://huggingface.co/Clinical-AI-Apollo/Medical-NER

BioBert NER
alvaroalon2/biobert_chemical_ner:
https://huggingface.co/alvaroalon2/biobert_chemical_ner
alvaroalon2/biobert_diseases_ner:
https://huggingface.co/alvaroalon2/biobert_diseases_ner

LED base model:
allenai/led-base-16384: https://huggingface.co/allenai/led-base-16384

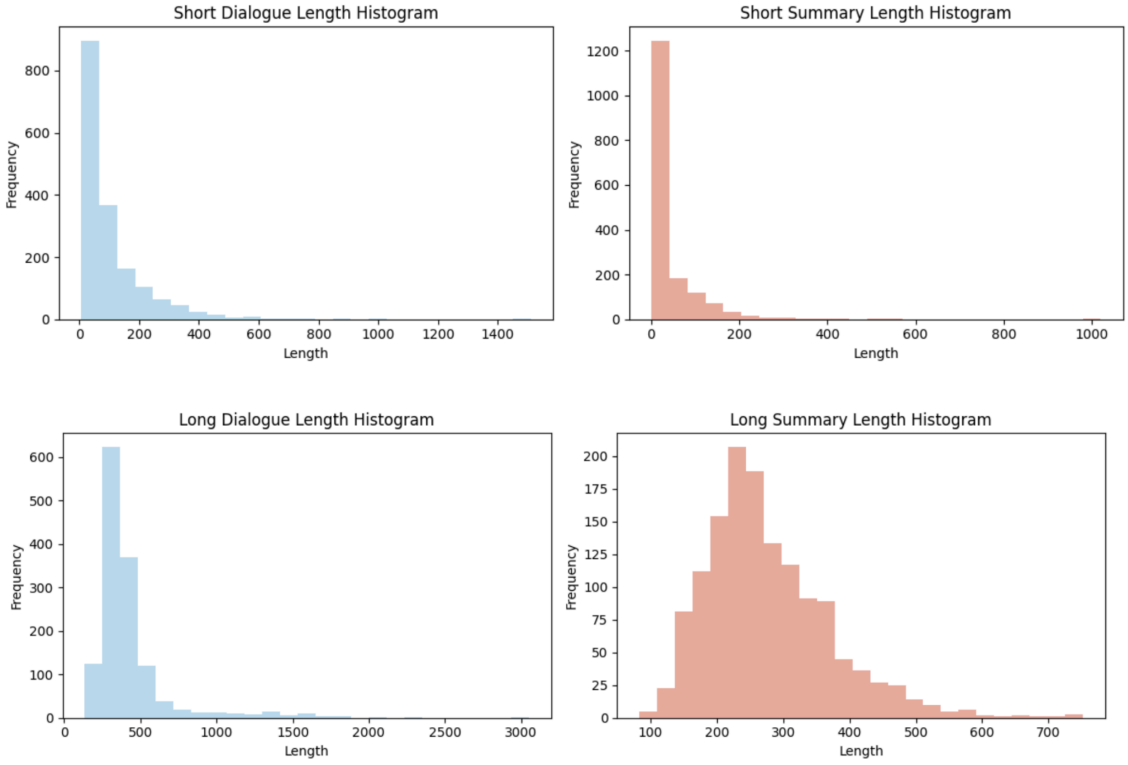
LED Large model:
allenai/led-large-16384: https://huggingface.co/allenai/led-large-16384

Flan_T5 model:
google/flan-t5-base: https://huggingface.co/google/flan-t5-base

Flan_T5 Large model:
google/flan-t5-large: https://huggingface.co/google/flan-t5-large

Appendix.

Appendix 1: Dialogue and summary length



Appendix 2 Model result for short dialogue

Table 2: Model results for LED short dialogue

ROUGE Score for Short Dialogue									
Baseline: LED_Large	**0.2101**	0.1458	0.6198	**0.0853**	0.0589	0.2479	**0.1351**	0.0914	0.4579
LED_Large + Keybert NER	0.1732	0.1141	0.6181	0.0689	0.0454	0.2474	0.1099	0.0703	0.4599
LED_Large + Biobert NER	0.1760	0.1139	0.6384	0.0686	0.0441	0.2499	0.1106	0.0695	0.4727
LED_Large + Medical NER	0.1714	0.1116	0.6181	0.0665	0.0431	0.2474	0.1095	0.0694	0.4649

These results in the table above show that the LED LORA large model variants with NER (KeyBERT, BioBERT, Medical NER) consistently performed worse than the baseline model in terms of ROUGE scores. The introduction of NER components did not improve performance, highlighting the challenge of integrating NER for short dialogue summarization effectively.

Table 3: Model results for Flan-T5 short dialogue - finetuned

ROUGE Score for Short Dialogue - finetuned									
Baseline: Flan-T5 Base finetuned	**0.2978**	0.5387	0.2580	0.1220	0.2577	0.0984	**0.2625**	0.4635	0.2314
Flan-T5 Base + Keybert NER finetune	0.2868	0.5313	0.2460	**0.1271**	0.2571	0.1042	0.2497	0.4552	0.2162
Flan-T5 Base + Biobert NER finetune	0.2855	0.5226	0.2518	0.1200	0.2438	0.0999	0.2480	0.4444	0.2209
Flan-T5 Base + Medical NER finetune	0.2890	0.5299	0.2468	0.1190	0.2460	0.0990	0.2498	0.4479	0.2164

For all Flan-T5 models, we initially used a learning rate of $1\times10-3$, typically reasonable but we found that this value led to model overfitting, while a much smaller rate resulted in poor model performance. To address this issue, we incorporated dropout and slow decrease in learning rate value to fine-tuned even better models.

Appendix 3 LED_base and LED_large ROUGE score comparison

ROUGE Score for Long Dialogue

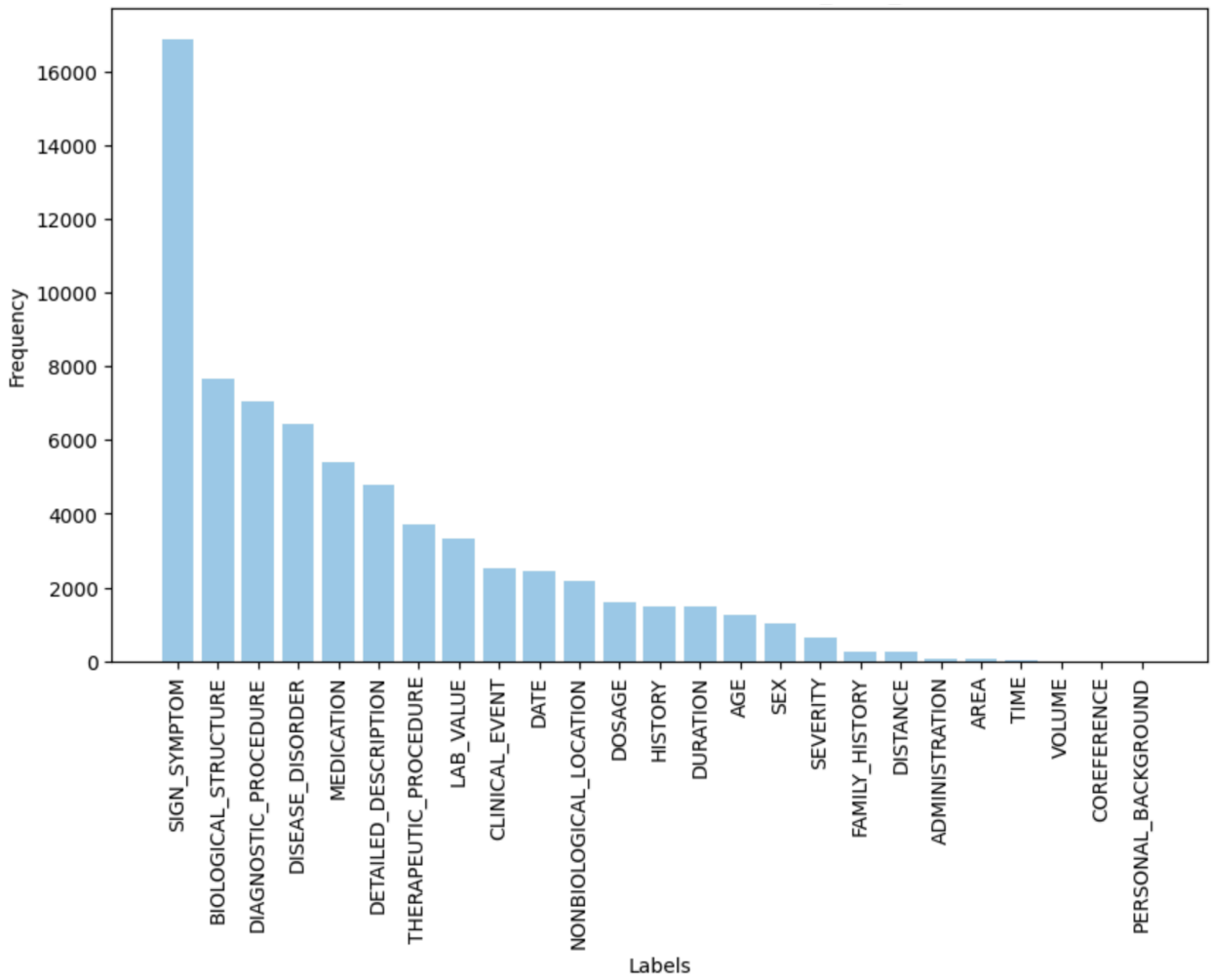
	ROUGE 1 (f1)	ROUGE 2 (f1)	ROUGE L (f1)
LED_Base*	0.4156	0.1791	0.2176
LED_Large	**0.5514**	**0.3112**	**0.3099**

* Subset: Ramdonly select 30% of the test data

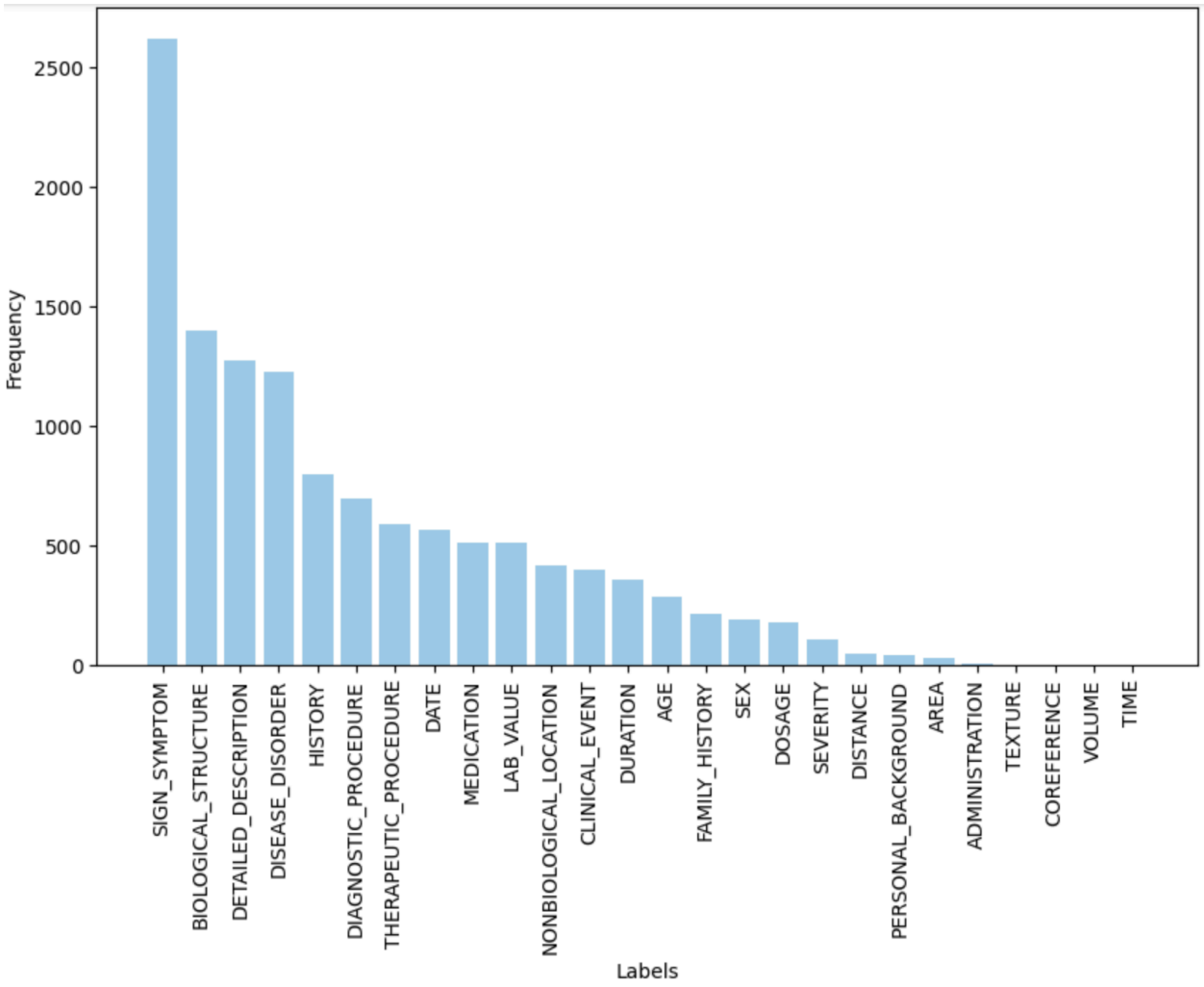
Appendix 4: Medical-NER

Full list: Activity, Administration, Age, Area, Biological_Attribute, Clinical_Event, Color, Coreference, Date, Detailed_Description, Diagnostic_Procedure, Disease_Disorder, Distance, Dosage, Duration, Family_History, Frequency, Height, History, Lab_Value, Mass, Medication, Nonbiological_Location, Occupation, Other_Entity, Other_Event, Outcome, Personal_Background, Qualitative_Concept, Quantitative_Concept, Severity, Sex, Shape, Sign_Symptom, Subject, Texture, Therapeutic_Procedure, Time, Volume, Weight

Frequency of Medical NER Label for Train set of Long Dialogue



Frequency of Medical NER Label for Train set of Short Dialogue



Appendix 5: Sample NER extracted from models

Dialogue	BioBert Unique_NER	BioBert All_NER	Keyword
Doctor: I have your referral here, from your primary physician? She said that you have a history of Huntington disease and that you have been experiencing some memory loss and confusion. Patient: Yes. That's right! I also have high blood pressure. Doctor: Are you on medication for your high blood pressure? Patient: Yes, I am. Doctor: Can you tell me more about the memory problems you have been experiencing? Patient: My memory has been getting worse. I can never find anything, or people keep moving my things. Guest_family: She is not recognizing us more and more these days. And it takes her longer to realize what is going on. I made an appointment with the Neurologist for next week. Doctor: Who is her Neurologist? Guest_family: Doctor Townsend. Doctor: Are you seeing any other specialists for your condition? Patient: Yes, I see Doctor Smith. Doctor: What kind of doctor is Doctor Smith? Patient: She is a psychiatrist. I have been seeing her for a long time. Doctor: Oh okay. So, she is not a psychiatrist at our facility? Patient: No.	‘Huntington disease', 'memory loss', 'confusion', 'high blood pressure', 'memory problems'	'Huntington disease', 'memory loss', 'confusion', 'high blood pressure', 'high blood pressure', 'memory problems'	'disease experiencing memory', 'psychiatrist seeing long', 'going appointment neurologist', 'huntington', 'things guest_family recognizing'
	Medical Unique_NER	**Medical All_NER**	
	Huntington disease', 'memory loss', 'confusion', 'high blood pressure', 'memory problems', 'memory find', 'moving', 'Neurologist', 'next week'	'Huntington disease', 'memory loss', 'confusion', 'high blood pressure', 'high blood pressure', 'memory problems', 'memory find', 'moving', 'Neurologist', 'next week'	

Appendix 6 Unique NER and frequency based NER ROUGE score comparison

ROUGE Score for Long Dialogue

	ROUGE 1 (f1)	ROUGE 2 (f1)	ROUGE L (f1)
LED_Large + unique Biobert NER*	0.4820	0.2364	0.2358
LED_Large + all Biobert NER	**0.5542**	**0.3177**	**0.3173**
LED_Large + unique medical NER*	0.5400	0.2931	0.2936
LED_Large + all medical NER	**0.5631**	**0.3306**	**0.3355**

* Subset: first 30% of the test data