Overview of the Problem List Summarization (ProbSum) 2023 Shared Task on Summarizing Patients' Active Diagnoses and Problems from Electronic Health Record Progress Notes

Yanjun Gao[1], **Dmitriy Dligach**[2], **Timothy Miller**[3],
Matthew M. Churpek[1] and **Majid Afshar**[1]
[1]University of Wisconsin, [2]Loyola University Chicago,
[3]Boston Children's Hospital and Harvard Medical School
[1]{ygao, mchurpek, mafshar} @medicine.wisc.edu,
[2]ddligach@luc.edu,[3]Timothy.Miller@childrens.harvard.edu

Abstract

The BioNLP Workshop 2023 initiated the launch of a shared task on Problem List Summarization (ProbSum) in January 2023. The aim of this shared task is to attract future research efforts in building NLP models for real-world diagnostic decision support applications, where a system generating relevant and accurate diagnoses will augment the healthcare providers' decision-making process and improve the quality of care for patients. The goal for participants is to develop models that generated a list of diagnoses and problems using input from the daily care notes collected from the hospitalization of critically ill patients. Eight teams submitted their final systems to the shared task leaderboard. In this paper, we describe the tasks, datasets, evaluation metrics, and baseline systems. Additionally, the techniques and results of the evaluation of the different approaches tried by the participating teams are summarized.

1 Introduction

Electronic health records (EHR) document patients' daily progress and, therefore, are an essential component of hospital care (Brown et al., 2014). However, note-taking in the EHR, while necessary, suffers from note bloat and information overload, causing challenges for end-users (Liu et al., 2022). The cognitive burden on providers using EHR documents remains high and only increases in our digital era (Furlow, 2020; Hultman et al., 2019). Summarization tasks in clinical natural language processing (NLP) are a promising approach to overcome note bloat and help extract relevant, active diagnoses to help mitigate diagnostic errors. Augmented intelligence with computerized diagnostic decision support is a potential solution to improve care throughput and quality at the bedside (Demner-Fushman et al., 2009; Lederman et al., 2022).

Over the past two decades, many tasks in the field of clinical NLP (cNLP) have focused on information extraction, document classification (e.g., named entity recognition), and relation extraction, as noted in a recent scoping review of cNLP tasks (Gao et al., 2022b). As generative NLP systems have advanced significantly in recent years, the field has a growing opportunity on taking advantage of the latest advancements in generative large language models such as T5 (Raffel et al., 2020), Galactica (Taylor et al., 2022), and GPT (Floridi and Chiriatti, 2020). Tasks of clinical text generation include radiology report generation (Xue et al., 2018; Ben Abacha et al., 2021), clinical note generation (Yang and Yu, 2020; Krishna et al., 2021; Grambow et al., 2022), and patients' diagnoses and problems summarization (Gao et al., 2022d). For the BioNLP Workshop 2023, the three shared tasks for the BioNLP Workshop 2023 all focused on biomedical text generation, in line with the current research trend on generative NLP models. This paper presents the overview of the shared task for the BioNLP Workshop 2023, *Problem List Summarization*.

The Problem List Summarization (**ProbSum**) task, initially proposed by Gao et al. (2022d), aimed to facilitate the development of NLP models for downstream applications in diagnostic decision support systems. The goal of this shared task was to attract future research efforts in developing NLP systems for real-world decision support applications, where a system generating relevant and accurate diagnoses can augment the healthcare providers' decision-making process.

The ProbSum task utilized the daily progress notes, a documentation approach that is ubiquitous in the EHR with a specific format that is taught in medical schools. The progress note contains the following four major sections: Subjective, Objective, Assessment, and Plan (SOAP) (Weed, 1969). Participants developed models to create a list of relevant diagnoses or problems based on the information present in the Subjective, Objective, and

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 461–467
July 13, 2023

Assessment sections of the note. The Plan contained the diagnoses and problems that served as the ground truth labels.

Summarizing the list of problems and diagnoses from daily progress notes is a challenging task for several reasons. First, progress notes may contain old diagnoses that are no longer relevant or list diagnoses that are not described or mentioned elsewhere in the note. Second, the progress note may contain extra information needed for billing purposes or quality improvement metrics that do not contain diagnoses. Further, the data collected in the progress note may be embedded from other parts of the EHR, such as laboratory results or medications. The structured text and tabular data alongside unstructured text present a multimodal problem. Lastly, progress notes may contain medical jargon and abbreviations (e.g. "GI", "EGD" in Figure 1) that can be challenging for NLP models to understand and summarize. To accurately summarize patient problems, models must possess a deep understanding of the medical concepts and terminology involved, as well as the broader context of the patient's medical history and current condition.

(Subjective)
Chief Complaint: CC: Diarrhea Admission: GI Bleed 24 Hour Events: - GI performed EGD: diffuse erythema/ ulceration, in esophagus, stomach, duodenum.. . .

(Objective)
07:39 AM Vital signs Hemodynamic monitoring Fluid balance 24 hours Since AM Tmax: 38.1 C (100.6 Tcurrent: 37.7 C (99.8 HR: 122 (96 - 128) bpm . . .

(Assessment)
72yo M with h/o metastatic rectal CA and SVC syndrome on lovenox admitted with intractable diarrhea thought chemo now with BRBPR and hematemesis.

(Plan Subsections with Problem List Annotation)
GI Bleed/diarrhea : Given EGD results, his bleed appears to be from diffuse ulceration from the esophagus through the duodenum. As diarrhea has been mucuous with pink tinge, . . .

(Ground Truth Problem List)
GI Bleed; diarrhea; diffuse ulceration from the esophagus through the duodenum; Acute Renal Failure; SVC Syndrome; Rectal CA;

Figure 1: An example progress note with Subjective, Objective, Assessment sections and the Ground Truth Summary annotated from Plan sections. Annotated text spans are highlighted in color box . In the ground truth, annotated diagnosis/problem are concatenated by semicolons. We use . . . to denote the continuation of the notes due to space constraints.

2 Task Description

Healthcare providers, including physicians, write daily progress notes to record the patient's medical status, treatment, and progress while they are hospitalized. These notes contain essential information, such as vital signs, medications, procedures, and any significant developments in the patient's condition. Daily progress notes are crucial for monitoring the patient's overall health and ensuring that the healthcare team is informed of any changes or updates in their treatment.

Daily progress notes are formatted using the SOAP Format. The *Subjective* section of a SOAP format daily progress note comprises the patient's self-reported symptoms, concerns, and medical history. The *Objective* section consists of objective data collected by healthcare providers during observation or examination, such as vital signs (e.g., blood pressure, heart rate), laboratory results, or physical exam findings. The *Assessment* section summarizes the patient's overall condition with a focus on the most active problems/diagnoses for that day. Finally, the *Plan* section contains multiple subsections, each outlining a diagnosis/problem and its treatment plan. The aim of the task was to predict the list of problems and diagnoses outlined in the *Plan* section (Weed, 1969). Figure 1 presents an input example from a progress note with Subjective, Objective, and Assessment sections, and the Ground Truth problem/diagnosis List annotated from the Plan section.

3 Data Description

The task contained 768 hospital daily progress notes and 2783 diagnoses in the training set, which was previously described in more detail (Gao et al., 2022d). A critical care physician and clinical informatics expert annotated a test set of 237 daily progress notes with only one progress note per patient to avoid redundancy.

3.1 Data source

The progress notes were sourced from MIMIC-III, a publicly available dataset of de-identified EHR data from approximately 60,000 hospital critical care admissions at Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al., 2016)[1]. The progress note types from MIMIC-III included a total of 84 progress note types across

[1]Data Use Agreement is required for all participants

medical, surgical, cardiovascular, trauma, and neurology specialties. Other note types were excluded such as Nursing Progress Notes and Social Worker Progress Notes because they were not structured in the SOAP format.

3.2 Annotation

The corpora for Problem List Summarization followed the annotation framework and guidelines previously described in (Gao et al., 2022c). The annotation was performed in the INCEpTION annotation platform (Klie et al., 2018). The goal of the annotation was to label lists of relevant problems/diagnoses from the Plan subsections. The annotators first marked the text span of the Assessment and Plan subsections. For each Plan subsection, the annotators marked the text span for the Problem, separating the diagnosis/problems from the treatment or action plans. Since the annotation was conducted on individual progress notes, the guidelines specified that only the active and pertinent diagnoses should be considered, taking into account the patient's condition as described in those specific progress notes. Consequently, only the active and relevant diagnoses/problems were included in this shared task.

In the original dataset introduced by Gao et al. (2022a), there were 768 annotated progress notes for training and test set. We merged the original split into training data for this shared task, and had a critical care physician annotate another 237 progress notes for a new test set. We further parsed the annotation into CSV files that had five columns: FILE ID, Subjective Sections, Objective Sections, Assessment, and Summary (Ground Truth). The FILE ID corresponded to the unique IDs of clinical notes in the original MIMIC-III dataset. Participants could choose a combination of Subjective, Objective and Assessment as input. They were also free to incorporate any other domains of the MIMIC data that were related to the patient's progress note.

For the final test set, we found the top 10 most diagnoses are: Anemia, auricular fibrillation, acute renal failure, Septicemia, Sepsis, Respiratory failure, Hypertension, Congestive heart failure, thrombocytopenia and Pneumonia.

4 Evaluation

As the evaluation metric, we employed ROUGE-L, a measurement that captures the Longest Common Subsequence (LCS) between a generated summary and its corresponding reference summary (Lin, 2004). In what follows, we report ROUGE-L Precision (RL-P), ROUGE-L Recall (RL-R), and ROUGE-L F-score (RL-F). Although we acknowledge the limitations of using ROUGE as an evaluation approach, as it does not assess semantics, the field of automated evaluation for clinical note summarization is still in its early stages, with most research relying on human evaluation besides conventional metrics like ROUGE (Gao et al., 2022d; Otmakhova et al., 2022). Therefore, for the purpose of the shared task, which requires ranking systems, we have chosen to use ROUGE as our evaluation metric.

5 Participation

We released the training data and test data through PhysioNet[2], which serve as the data owners of the MIMIC-III data and the executors of the Data Use Agreement with participants. During System Evaluation Phase, we created a CodaLab competition to manage the system submission and leaderboard (Pavao et al., 2022). To be specific, the CodaLab was configured to receive system prediction outputs in a text file format, and subsequently ran the ROUGE-L evaluation script to produce scores. Each user was permitted to submit a maximum of 30 runs. At the time of writing this paper, eight teams from the United Kingdom, India, Spain, South Korea, Ireland, Australia, and Mexico submitted a total of 164 runs. Along with the system output text file, we asked participants to provide a brief description of their methodology; however, some users submitted their results without providing us with any system details.

6 Results

6.1 Baseline

According to the findings of Gao et al.(2022d), the ProbSum task's baseline result was established by applying pre-trained sequence-to-sequence language models, specifically T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). T5 produced the most favorable set of results. To achieve the best ROUGE-L F-score of 18.80, T5 was subjected to domain adaptive pre-training (Gururangan et al., 2020) and concept masking, which involved randomly masking the concepts identified by the

[2] https://physionet.org/

concept extractor and continually training T5 on MIMIC data. Gao et al. (2022d) further found that T5 produced the best results when Subjective and Assessment sections were used as inputs. However, inputting Objective sections could result in noisy output.

6.2 Main results

Team	RL-P	RL-R	RL-F
CUED	41.69	30.51	32.77
PULSAR	44.30	27.18	31.14
Pune Institute of Comp. Tech	41.39	22.96	27.44
Universitat Politècnica de Catalunya	32.86	24.79	25.92
Sungkyunkwan Univ.	34.44	19.55	22.39
Deakin Univ.	28.66	19.11	20.84
Univ. of Limerick	39.44	15.18	19.67
Universidad Nacional Autónoma De México	11.88	13.37	11.83

Table 1: Best-performing submission results from the eight participation teams. We report ROUGE-L Precision (RL-P), Recall (RL-R) and F-score (RL-F).

Table 1 displays the best results submitted by eight participating teams, with Team CUED (University of Cambridge) achieving the highest F1 score of 32.77. Team PULSAR (University of Manchester) closely followed with a score of 31.44. The team from Pune Institute of Computer Technology held the third spot with an F1 score of 27.44. The F1 score reported by Team Universitat Politècnica de Catalunya was 25.92, while Team Sungkyunkwan University and Team Deakin University reported F1 scores of 22.39 and 20.84, respectively. Team University of Limerick reported an F1 score of 19.67, which is near the baseline. Team Universidad Nacional Autónoma De México reported the lowest F1 score of 11.83.

6.3 Methods overview

We asked participants to submit a post-participation survey that further described their approaches, and received four teams' response. The survey contained questions regarding the best-performing systems and input setting, as listed in Figure 2. Four of the eight teams submitted their answers to the survey questions: team CUED, PULSAR, Universitat Politècnica de Catalunya and Univ. of Limerick.

The collected surveys revealed that all teams were utilizing transformer based language models, like T5 and BERT (Devlin et al., 2019). T5 checkpoints trained on medical resources were popular choices for methods, such as Clinical-T5 (Lehman et al., 2023). The medical resources employed for training, apart from MIMIC, included Unified Medical Language System (UMLS) (Bodenreider, 2004), BC5CDR (BioCreative V CDR corpus) (Li et al., 2016), and i2b2 2010 dataset (Uzuner et al., 2011). All teams employed the Subjective and Assessment sections as inputs, while two teams used Objective sections as additional inputs. There were no teams that included a member who was a medical provider or had consulted with medical experts.

- Method descriptions
- What is the input to your best performing system? (Options: Subjective, Objective, Assessment)
- Did you use language model?
- Did you use medical doctors' expertise? (One of your team member is medical doctor, or your team has consulted with medical doctor)
- Did your model use other resources (e.g. during pretraining) besides MIMIC? (E.g. PubMed text, UMLS)

Figure 2: Post-participation Survey Questions

Team CUED achieved the best performance by using an ensembled clinical T5 model. On the other hand, Team PULSAR used Flan-T5 (Chung et al., 2022) and GPT2XL (Radford et al., 2019). Team Universitat Politècnica de Catalunya trained a BERT-based Named Entity Recognition (NER) BIO tagger (NER) system to extract keywords from input and composed the output. Finally, team University of Limerick developed a hybrid summarization system that utilized the Pagerank algorithm (Langville and Meyer, 2006) and QuickUMLS (Soldaini and Goharian, 2016) on a concept graph to extract the most important sentences. They further used a fine-tuned T5 model with concept masking to generate an abstractive summary of the generated summary.

7 Discussion

The 2023 ProbSumm BioNLP Workshop revealed that transformer-based language models trained on medical resources are capable of generating summaries from medical texts.

While this finding was consistent with previous research that showed the effectiveness of these models in various NLP tasks, including text summarization, the ProbSumm task proved to be very difficult and is far from solved with a best F1 score of 32.77.

None of the teams involved in the task included a medical provider or sought consultation from medical experts while developing their models. This could be attributed to practical limitations such as time and resource constraints. However, incorporating the expertise of medical professionals during the development process may provide valuable insights into the clinical implications of the generated summaries, resulting in the creation of more clinically relevant and useful models. Given the effectiveness of incorporating human feedback in various NLP tasks (Ouyang et al., 2022), we recommend future research to explore the performance of involving medical experts in the development and evaluation through a human-in-the-loop approach.

8 Conclusion

The ProbSum task, part of the BioNLP Shared Task 1A, focused on summarizing patients' diagnoses from daily progress notes. Eight teams from 7 countries submitted the final systems for evaluation, and the top-performing system has achieved substantial gains from the baseline with a new state-of-the-art result of 32.77. The experience gained by participants can inform the development of more effective strategies for medical text summarization, which could have significant implications for augmenting computerized diagnostic decision support systems at the bedside and potentially translate into better patient care.

References

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

PJ Brown, JL Marquard, B Amster, M Romoser, J Friderici, S Goff, and D Fisher. 2014. What do physicians read (and ignore) in electronic progress notes? *Applied clinical informatics*, 5(02):430–444.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Bryant Furlow. 2020. Information overload and unsustainable workloads in the era of electronic health records. *The Lancet Respiratory Medicine*, 8(3):243–244.

Y. Gao, J. Caskey, T. Miller, B. Sharma, M. Churpek, D. Dligach, and M. Afshar. 2022a. Tasks 1 and 3 from Progress Note Understanding Suite of Tasks: SOAP Note Tagging and Problem List Summarization. PhysioNet.

Yanjun Gao, Dmitriy Dligach, Leslie Christensen, Samuel Tesch, Ryan Laffin, Dongfang Xu, Timothy Miller, Ozlem Uzuner, Matthew M Churpek, and Majid Afshar. 2022b. A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association*, 29(10):1797–1806.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M Churpek, and Majid Afshar. 2022c. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources & Evaluation*, volume 2022, page 5484. NIH Public Access.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew MM Churpek, and Majid Afshar. 2022d. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991.

Colin Grambow, Longxiang Zhang, and Thomas Schaaf. 2022. In-domain pre-training improves clinical note generation from doctor-patient conversations. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 9–22.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Gretchen M Hultman, Jenna L Marquard, Elizabeth Lindemann, Elliot Arsoniadis, Serguei Pakhomov, and Genevieve B Melton. 2019. Challenges and opportunities to improve the clinician experience reviewing electronic progress notes. *Applied clinical informatics*, 10(03):446–453.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Amy N Langville and Carl D Meyer. 2006. *Google's PageRank and beyond: The science of search engine rankings*. Princeton university press.

Asher Lederman, Reeva Lederman, and Karin Verspoor. 2022. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. *Journal of the American Medical Informatics Association*, 29(10):1810–1817.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *arXiv preprint arXiv:2302.08091*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022. "note bloat" impacts deep learning-based nlp models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Julia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Lawrence L Weed. 1969. *Medical records, medical education, and patient care: the problem-oriented record as a basic tool*. Press of Case Western Reserve University.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer.

Zhichao Yang and Hong Yu. 2020. Generating accurate electronic health assessment from medical graph. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 3764. NIH Public Access.