ABSTRACT

Title of Dissertation: A Human-centric Approach to NLP in Healthcare Applications

Han-Chin Shing

Dissertation Directed by: Professor Philip Resnik
Linguistics/UMIACS
Professor Douglas W. Oard
iSchool/UMIACS

The abundance of personal health information available to healthcare professionals can be a facilitator to better care. However, it can also be a barrier, as the relevant information is often buried in the sheer amount of personal data, and healthcare professionals already lack time to take care of both patients and their data. This dissertation focuses on the role of natural language processing (NLP) in healthcare and how it can surface information relevant to healthcare professionals by modeling the extensive collections of documents that describe those whom they serve.

In this dissertation, the extensive natural language data about a person is modeled as a set of documents, where the model inference is at the level of the individual, but evidence supporting that inference is found in a subset of their documents. The effectiveness of this modeling approach is demonstrated in the context of three healthcare applications. In the first application, clinical coding, document-level attention is used to model the hierarchy between a clinical encounter and its documents, jointly learning the encounter labels and the assignment of credits to specific documents. The second application, sui-

cidality assessment using social media, further investigates how document-level attention can surface "high-signal" posts from the document set representing a potentially at-risk individual. Finally, the third application aims to help healthcare professionals write discharge summaries using an extract-then-abstract multidocument summarization pipeline to surface relevant information.

As in many healthcare applications, these three applications seek to assist, not replace, clinicians. Evaluation and model design thus centers around healthcare professionals' needs. In clinical coding, document-level attention is shown to align well with professional clinical coders' expectations of evidence. In suicidality assessment, document-level attention leads to better and more time-efficient assessment by surfacing document-level evidence, shown empirically using a theoretically grounded time-aware evaluation measure and a dataset annotated by suicidality experts. Finally, extract-then-abstract summarization pipelines that assist healthcare professionals in writing discharge summaries are evaluated by their ability to surface faithful and relevant evidence.

A Human-centric Approach to NLP in Healthcare Applications

by

Han-Chin Shing

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Philip Resnik, Chair/Advisor
Professor Douglas W. Oard, Co-Advisor
Professor Guodong Gao, Dean's Representative
Assistant Professor Rachel Rudinger
Assistant Professor John P. Dickerson

Dedication

To Li-Ju Hong

Thank you for showing me what is possible.

Acknowledgments

My research journey as a student began and ended with improving healthcare with computers. However, the path from beginning to end was anything but straightforward. My undergraduate research focused on optimizing microfluidic biomedical devices with genetic algorithms. Through a surprising turn of events, I have since worked on a stamp placement system in the military, retrofitting word embeddings, and evidence combination of information retrieval systems. Somehow, I found my way back into healthcare: improving healthcare applications with natural language processing.

Among the many people that make this journey possible, perhaps my advisor, Philip Resnik, played the most crucial role. Not only did he give me a chance to learn firsthand how to do research and think about problems critically, but he also gave me the freedom and support to pursue my research interests. He provided invaluable advice to improve papers by reframing the arguments and problem structures (known as "Resnik the paper" among his advisees). My co-advisor, Doug Oard, taught me the importance of critical thinking. His ability to break down complex problems and identify key bottlenecks with plans to tackle them is always amazing to watch. His wealth of knowledge and the ability to connect the dots between different disciplines help me think about research in a broader context. I will undoubtedly miss the sometimes heated debates and conversations between the three of us. Hopefully, I manage to learn a fraction of their craft of research.

I wish to thank the members of my dissertation committee. Gordon Gao brought invaluable insights from a practical perspective, asking about the challenges of deployment. Rachel Rudinger asked precisely the right questions about the assumptions we made in

our models and evaluations, encouraging me to think about the potential limitations and implications. I enjoyed working with John Dickerson and his collaborators. His feedback helped me think about the potential societal impact of NLP in healthcare applications and how that may affect different groups of people.

Joining CLIP lab at UMD was perhaps the second-best decision I made as a Ph.D. student (we'll get to the best decision later). I wish to thank Hal Daumé III for helping me start the research on suicidality assessment, Jordan Boyd-Graber for encouraging me to think about what is special about suicidality assessment compared to other problems, and Marine Carpuat for introducing me to the field of natural language processing. I wish to thank Joe Barrow, Pedro Rodriguez, and Yogarshi Vyas for the countless discussions about research and espresso, and for their friendship and support. I wish to thank every member of the MATERIAL team. Working with Petra Galuščáková, Suraj Nair, and Elena Zotkina is a wonderful experience. I hope we can soon celebrate again at Looney's. I wish to thank Naomi Feldman, Denis Peskov, Jake Bremerman, Meir Friedenberg, Jo Shoemaker, Weiwei Yang, Khanh Nguyen, Pranav Goel, Alexander Hoyle, Kianté Brantley, Yow-Ting Shiue, Michelle Yuan, Mahmoud Sayed, Ahmed Elgohary, Amr Sharaf, Weijia Xu, Xing Niu, Shi Feng, Chen Zhao, Emily Gong, Trista Cao, and all members of CLIP lab and Linguistics for the great feedback and insightful discussions.

Outside of UMD, I have had the great fortune to work with many experts in their field. Suchi Saria at Johns Hopkins University introduced me to the field of machine learning for healthcare. At 3M Health Information Systems, I have the opportunity to work with Guoli Wang and many others in a production environment. I also have a wonderful experience at Amazon Comprehend Medical. Working with Parminder Bhatia, Chaitanya

Shivade, and Nima Pourdamghani allowed me to learn and grow as a researcher.

I owe my chances to pursue a career in research to my undergraduate advisors, Tian-Li Yu and Chih-Ting Lin. They sparked my interest in computer science and healthcare. It is their referrals and recommendations that help me come to UMD. My research at UMD is supported in part by the University of Maryland Strategic Partnership (MPower) seed grant, the National Institutes of Health, the AWS Machine Learning Research Award, the AI + Medicine for High Impact (AIM-HI) Challenge Award, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). Though I am grateful for the funding, the views and conclusions contained in this dissertation are those of the authors and should not be interpreted as necessarily representing the organizations and individuals mentioned above.

The constant support from my friends and family has been a beacon throughout the somewhat dreary world of academia. With their love and support, every rejection became more bearable, every success more sweet. Thank you, Joe, Micheal, Gwen, Aaron, Belle, Tina, Andy, Yunfeng, and Hank, for being good friends that I know I can always depend on. If I ever achieve anything, it is because of my amazing parents, Li-Ju Hong and TaiKang Shing. The unconditional love from my brother and sister, Hanyi and Han-Shin, my grandparents, 邢兆昌 and 林癸美, keeps me going in the face of adversity. The best decision I made in grad school, and perhaps life, is marrying my lovely wife, Sharon. It is impossible to thank you for the love and support you have given me in the last eleven years. I look forward to our journey through life's next chapter and beyond.

Table of Contents

Chapter 1: Introduction

With the advance of natural language processing (NLP) and machine learning, an unprecedented amount of information from text is suddenly unlocked and made available. This rise in information, both in quantity and quality, has led to improvements in many automated tasks. However, the increase in quantity poses a challenge to specific tasks that cannot, and *should not*, be fully automated without human intervention. NLP for healthcare applications is one such example and the focus of this dissertation. Specifically, we discuss the technological challenges and design decisions that arise from putting patients and healthcare professionals at the center of NLP for healthcare applications.

The relationship between patients and healthcare professionals is fundamental to care quality (Makoul et al., 2001; Fortin et al., 2012). An established relationship facilitates efficient information exchange between patients and healthcare professionals, contributing to improved care quality (Goold and Lipkin Jr, 1999). With information technology introduced into the mix, the ever-evolving relationship faces new challenges and opportunities. The collection, processing, integration, and interpretation of patient information and healthcare professionals' decision-making can all be influenced by the design and effectiveness of information technology (Weiner and Biondich, 2006).

Information technology, such as NLP for healthcare applications, can be a facilitator or a barrier to this information exchange (Weiner and Biondich, 2006). It can be a *facilitator*, in that it brings an unprecedented amount of information to both healthcare professionals and patients. This information can be indirectly related to the patients, such as increased public access to health-related information from published research. This dissertation will focus on information *directly* related to the patients, ranging from electronic health records (EHR) to naturally occurring language (e.g., social media) accumulated in everyday life.

The abundant information has the potential to help healthcare professionals, but it can also be a barrier. *Time*, is the limiting constraint. Healthcare professionals already lack time to manage both patients and their data (Weiner and Biondich, 2006; Sinsky et al., 2016). The increase in quantity does not guarantee an increase in care quality if healthcare professionals do not have time to process and interpret it, and in some cases leads to medical errors and negatively affects the mental well-being of the healthcare professionals (Tawfik et al., 2018; West et al., 2018). Design decisions must center around an efficient presentation of the information, surfacing the relevant content, and providing healthcare professionals effective means to manipulate and interpret the information. NLP has the potential to help mitigate the problem by prioritizing the information that aligns with healthcare professionals' needs.

This dissertation focuses specifically on NLP for healthcare and its roles in (1) processing patients' information and (2) assisting healthcare professionals by surfacing relevant information about the patients. In the following sections, we introduce the two main contributions of this dissertation. The first is modeling the patient as a set of documents; we show its application in clinical coding, suicidality assessment, and discharge summary composition. The second is prioritizing relevant information for healthcare professionals; we demonstrate the design of models and evaluations for surfacing relevant information in our work in suicidality assessment and discharge summary composition.

1.1 Modeling the Patient as a Set of Documents.

Textual information about patients is often extensive and complex. NLP has already played a role in helping healthcare professionals process mediated textual information about patients, such as clinical notes written by healthcare professionals in electronic health records (EHR) (Demner-Fushman et al., 2009; Wang et al., 2018, 2019). Clinical coding, for example, often involves machine learning models that assist healthcare professionals (in this case, clinical coders) in translating a clinical encounter, which is composed of a potentially large set of clinical notes, into a set of alphanumeric clinical codes for insurance and billing purposes (Resnik et al., 2006; Stanfill et al., 2010; Shing

et al., 2019). The well-recognized concerns of information overload in EHR have also lead to research and development of methods to summarize the extensive patient records, including our work on discharge summary composition described in Chapter 6 (Demner-Fushman and Lin, 2006; Farri et al., 2012; Pivovarov and Elhadad, 2015; Shing et al., 2021).

On the other hand, the use of non-mediated textual information, such as the patient's self-narrative in social media, is not as well studied. However, this non-mediated information has the potential to offer a valuable and complementary view for assessment. Coppersmith et al. (2017) introduce the concept of *clinical whitespace* in the context of mental health assessment, advocating filling the whitespace in between the relatively sparse clinical visits with dense language signals found in data sources such as social media. These naturally occurring language signals are not only complementary to mediated data but also have the value of being *in situ* behavior data (Resnik et al., 2020). Using NLP to triage suicidality risk based on social media data, for example, is an active area of research (Coppersmith et al., 2014, 2018; De Choudhury et al., 2016; Shing et al., 2018, 2020; Zirikly et al., 2019; Resnik et al., 2020; MacAvaney et al., 2021). Given an individual and their social media postings, the ultimate goal is to assist healthcare professionals in assessing, preventing, and monitoring the mental well-being of the patients.

These two complementary sources of information share some important characteristics. The most prominent is that they are both information at the patient level, although collected from different contexts and narrative standpoints. Another shared characteristic is that, in the lens of NLP, we can model the patient as *a set of documents*. In clinical coding, for example, a majority of clinical codes are assigned to the clinical encounter (a *set* of clinical notes) rather than a single clinical note.[1] For summarization systems of EHR, relevant patient information the healthcare professionals need can also scatter across different clinical notes (Farri et al., 2012). Relevant information thus needs to be extracted from the clinical encounter (again, a *set* of clinical notes) and then be aggregated to present it at the encounter level.

[1] Not all clinical coding is at the encounter-level. In outpatient radiology coding, for example, codes are typically assigned based on individual documents.

Similarly, suicidality assessment using social media aims to help healthcare professionals assess the risk of suicidality of a given individual by using the *sequence* of postings they made on social media. The risk is a property of the individual, not a specific posting. Importantly, not all postings made by an at-risk individual need to be an indication of suicide. Similarly, in the context of clinical coding, to assign a code for Chronic Obstructive Pulmonary Disease (COPD, ICD-10: J44.9), not all clinical documents need to contain evidence for COPD. Assigning a label to a set of documents does not imply that all documents within that set have evidence to support the assignment. In both cases, it is common that most documents do not contain evidence for label assignment.

This creates a challenge to the machine learning model, as other irrelevant documents can dilute relevant signals. On the other hand, we can mitigate the challenge if we know a priori which subset of the patient's documents contains signals. The observation that a subset of the documents suffices the inference hints at a hierarchical structure between the patient and their documents: on the higher patient level is the assignment of that inference, and the lower document level contains the evidence to support that inference. In Chapters 3, 5, and 6, we further discuss this hierarchical structure in the context of clinical coding, suicidality assessment, and discharge summary composition, respectively. Notably, this model does not support all assessment processes. Some assessment processes need to be global and require a holistic view to make the correct judgment, in contrast to when a more local view suffices. Thus, this dissertation will focus on the end of the spectrum when a *locality constraint* can be satisfied; correct judgments can be inferred based solely on a subset. In our suicidality work, our proposed evaluation measure makes another critical assumption – the document independence assumption. Here we assume that a document's relevance (defined as indicativeness of suicidal signals) is independent of other documents' relevance. That is, seeing a document does not change the relevance of the other documents. This assumption is often made, implicitly or explicitly, in most document retrieval tasks and many NLP tasks.

1.2 Prioritizing Relevant Information for Healthcare Professionals

Another challenge to healthcare applications, which we have hinted at earlier, is that many should not be fully automated. In clinical coding, a miss or false alarm can cost the patient or the insurance company thousands of dollars in economic terms alone. For clinical summarization, misses and false alarms (information in a summary not found in the source documents) have the potential to lead to medical errors. In the setting of suicidality assessment, the impact can potentially be worse. Thus, for many healthcare applications, it is desirable to involve humans in the loop to review the system's output. However, the introduction of human assessors gives rise to a problem of *resource-boundedness*: the amount of information, partly attributed to the success of machine learning, overwhelms what can be assessed by the human assessors in a given time. This resource-boundedness motivates a *prioritization* framing of the problem, surfacing the information that requires the most attention for a human to best utilize the limited human resources available.

What information should be prioritized and how it should be prioritized is problem-dependent. In this dissertation, we explore prioritization under two different settings. The first setting is healthcare professionals' assessment of suicidality using social media. In this setting, healthcare professionals are faced with an assessment of a potentially large pool of individuals. Arranging these individuals in a single priority queue, the healthcare professional can scan from top to bottom. This formulation of ranking is a well-studied problem in information retrieval (IR), with well-established evaluation measures like Expected Reciprocal Rank (ERR) and Normalized Discounted Cumulative Gain (NDCG) (Chapelle et al., 2009; Järvelin and Kekäläinen, 2002). In our setting, as argued above, the suicidality assessment of an individual follows a hierarchical structure; a sequence of postings represents an individual. Combining the ranking of documents with the ranking of the individuals whose documents describe, we arrive at a hierarchical ranking problem. In the context of suicidality assessment, we rank the individuals by their risk of suicide. Within the documents posted by an individual, we rank the documents for evidence that signals suicidality. This hierarchical ranking enables us better to utilize the limited resource – the assessor's time. By ranking individuals by their risk, we prioritize

the individuals who most need attention. By ranking documents within an individual, we hope to shorten the time it takes to assess that individual. Human assessors are hopefully more likely to find evidence of suicidal signals earlier.

A similar emphasis on time constraints can be found in the second setting: discharge summary composition from prior clinical documents in the encounter. In this setting, healthcare professionals are tasked to write a discharge summary, a semi-structured summary of the clinical encounter written for patient discharge. NLP has the potential to surface information relevant to the summary and can potentially assist healthcare professionals in writing discharge summaries. However, displaying the summary without knowing where the information comes from can hinder healthcare professionals' progress. If healthcare professionals suspect errors in the summary, they will need to spend time searching for evidence in the encounter that supports the information. Since time is the limiting resource, this potentially defeats the usefulness of the summarization system. Instead, the system should be built with *traceability* in mind; the summaries and their source should be displayed together such that we can easily trace back to the source documents for evidence. We can again describe this using the hierarchical structure between the patient and their documents. On the lower (document) level, extractive summarization systems can be used to extract and surface the relevant content. An abstractive summarization system can then collect the extracted content from all documents and merge them into a single summary for the higher (patient) level. Displaying the extracted content from each document together with the final summary provides a path of traceability. Healthcare professionals can reference the extracted content when they suspect errors in the summary; each extracted content item can also be traced back to their context in the source document for further review.

1.3 Clinical Coding at the Encounter Level

In Chapter 3, we describe our work on clinical coding at the encounter level. As previously mentioned, a clinical encounter can be seen as a set of clinical documents, with the labels assigned at the encounter level for the purpose of clinical code prediction. However,

the vast majority of the previous work focuses on a single representative document from the encounter, namely, the discharge summary. This creates two problems: (1) around 17.5% of the primary diagnoses can be missing from the discharge summary (Kripalani et al., 2007), and (2) for an outpatient encounter (where the patient is generally not admitted to the hospital), there might not even be a discharge summary, as the patient is generally not admitted in the first place, and thus cannot be discharged.

We address these two problems by observing a hierarchical structure in the clinical coding process: the code is assigned to the encounter, and at least one of the documents within the encounter will have evidence to support the code. At training time, however, no annotations are provided about which subset of documents contains the evidence. This hierarchical structure hints at a hierarchical solution: a hierarchical attention network (HAN, Yang et al., 2016). We extend HAN to learn document attention to aggregate document representations into an encounter representation and then use the encounter representation for code prediction. To test our hypothesis that modeling document attention improves the coding performance, we show that the model with document attention consistently outperforms a baseline without document attention.

Using a small but high-quality document-level test dataset annotated by professional clinical coders that is disjoint from the training, development, and testing datasets, we show that the document-level attention learned aligns well with the professionals' judgment of the source documents. This preliminary study inspires us to investigate further the potential of using document attention to surface evidence. In Chapter 5, we show how document-level attention can assist healthcare professionals in the context of suicidality assessment.

1.4 Suicidality Assessment using Social Media

In Chapter 5, we reframe suicidality assessment as a prioritization task under a time constraint. Similar to clinical coding, suicidality assessment using social media also follows a hierarchical structure. The degree of suicide risk is a property of the individual. At least one post may contain evidence of suicidal signals in the at-risk individual's social

media postings, and often with no annotations available on the post level. Going beyond our clinical coding work, however, we introduce an explicit time budget. As we argue above, the human assessor's time is the limiting resource. This time constraint suggests a hierarchical *ranking* reformulation: ranking the individuals by their risk of suicide and ranking the postings of an individual based on the likelihood of containing evidence.

To obtain a joint ranking of the individuals and their post rankings, we introduce the 3HAN model, a HAN model with three levels: attending from words to represent a sentence, attending from sentences to represent a document, and attending from documents to represent an individual. The individual's predictive risk score is used to rank the individual; the document attention learned without document annotations is then used to rank documents within an individual.

To evaluate the effectiveness of the hierarchical ranking under a time-budget, we extend the Time-Biased Gain (TBG, Smucker and Clarke, 2012) evaluation measure to Hierarchical Time-Biased Gain (hTBG) by extending the individual time estimation model with the cascading user model found in Expected Reciprocal Rank (ERR, Chapelle et al., 2009). We then show through axiomatic analysis that hTBG satisfies important characteristics one would want for a measure of suicidality assessment under a time budget.

In our experimentation, we show that hTBG can indicate the performance of individual ranking and document ranking. Furthermore, it shows that document ranking performance can have a non-negligible effect, as the time saved for assessing one individual means potential time for assessing others in need. Our 3HAN model also produces promising results, showing that better document ranking leads to a better individual ranking, leading to better hTBG scores.

Chapter 4 describes the collection of an expert-annotated dataset for suicidality assessment from Reddit: the UMD Reddit Suicidality dataset, which we use for our work in Chapter 5. The dataset is annotated with three levels of annotation quality, ranging from suicidality experts, to crowdsourcers, to weak supervision – the action of posting on Reddit's SuicideWatch subreddit. For the subset annotated by the suicidality experts, we asked the experts to identify the single post that is most indicative of their assessment. For any individual we annotated, we also collected all their postings from other subreddits. A

version of this dataset was also used for the NAACL CLPsych 2019 shared task (Zirikly et al., 2019), and has since been shared with dozens of other researchers.

1.5 Learning to Compose a Discharge Summary from Prior Clinical Notes

Finally, we describe our work on learning to compose a discharge summary from prior clinical notes. Chapter 6 builds on our two main contributions: (1) modeling the patient as a set of documents, and (2) surfacing relevant information to healthcare professionals. In our suicidality work and clinical coding work, we show that modeling the patient as a set of documents can improve the predictive performance and surface relevant information. In Chapter 6, instead of predicting discrete labels like clinical codes or degree of suicidality risk, we focus on computer-assisted discharge summary composition from prior clinical documents of the encounter.

Physicians spend almost half of their time on clinical documentation (Shanafelt et al., 2016). This documentation burden is a significant driver to clinician burnout (Tawfik et al., 2018; West et al., 2018). One of the documents healthcare professionals write is the discharge summary, which is a semi-structured summary of the patient's clinical encounter (a *set* of documents) for the patient's discharge. This motivates our work on building systems to help healthcare professionals write discharge summaries. It extends our main contributions in two ways: (1) modeling the patient as a set of documents (i.e., multidocument summarization), similar to our clinical coding and suicidality assessment work, and (2) summarizing the encounter is a natural way to surface relevant medical information.

However, applications in the healthcare setting often come with important implications and should therefore involve healthcare professionals in the loop. This introduction of healthcare professionals, however, brings new challenges to the otherwise straightforward multidocument summarization task. Here, we identify three main challenges. (1) A clinical summarization system should support *traceability*: an ability to investigate the supporting evidence for the generated summary. (2) *Faithfulness* to the source documents is an important aspect of clinical summarization, and the evaluation should reflect that.

Finally, (3) the model needs to *scale* to multiple, potentially very long documents, as is sometimes the case for clinical encounters.

We address these challenges in Chapter 6. Recognizing the importance of *traceability* of clinical summarization systems, we propose an extract-then-abstract pipeline. We can describe the pipeline in the framework of the hierarchical structure between the patient and their documents. An extractive model first extracts relevant content from each document individually. An abstractive model then collects the extracted content from all documents and merges them into the final summary at the encounter level. This pipeline provides a path of traceability for the healthcare professionals. Healthcare professionals can reference the extracted content for information mentioned (or missing) in the summary. The extracted content can then be easily traced back to their context in the source document. The extract-then-abstract pipeline also helps with *scalability*, as extractive models are often more scalable than abstractive models.

A clinical summary needs to be *faithful* to the source. That is, the summarization system should not introduce new information not mentioned in the source documents. Following recent work on measuring faithfulness in summarization (Maynez et al., 2020; Zhang et al., 2020), we propose a set of *faithfulness-adjusted* measures based on matching medical mentions in Unified Medical Language System (UMLS, Bodenreider, 2004) Metathesaurus in Chapter 6.[2] In Chapter 6, these measures are used in conjunction with ROUGE, an informativeness measure conventionally used in summarization, to evaluate the summaries (Lin and Hovy, 2003; Maynez et al., 2020).

We derive our dataset from the MIMIC III v1.4 database (Johnson et al., 2016): a freely accessible critical care database consisting of a set of de-identified clinical data of patients admitted to the Beth Israel Deaconess Medical Center's Intensive Care Unit (ICU). The database includes structured data such as medications and laboratory results; and unstructured data such as clinical notes written by medical professionals. In Chapter 6, we focus on unstructured data.

Discharge summaries are semi-structured and can be broken down into different

[2]Note that the medical mention-based measure has its limits in approximating faithfulness. As these medical mentions do not capture negation or modifiers, they can only act as one type of measure of mention-level faithfulness. It is thus a proxy for a specific aspect of faithfulness. We describe this in further in Chapter 6.

sections, including past medical history, family history, chief complaints, and history of present illness. This allows us to compose discharge summaries one section at a time instead of composing the entire discharge summary. In Chapter 6, we test our extract-then-abstract pipeline on seven discharge summary sections.[3] Across the five models we test, including a BERT-based extractor (Liu and Lapata, 2019b), a reinforcement learning RNN-based extractor (Chen and Bansal, 2018), a BART abstractor (Lewis et al., 2019), a pointer generator (See et al., 2017), and a reinforcement learning RNN-based sentence rewriter (Chen and Bansal, 2018), we find that sentence-rewriting approaches, when supports traceability, perform consistently better on our faithfulness-adjusted measures. Interestingly, we observe that the BART abstractor can smooth out differences in extracted content that results from the choice of different extractors in the extract-then-abstract pipeline.

1.6 Contributions

Clinical Coding at the Encounter Level

- We identify an important label mismatch problem in the clinical coding hierarchy: clinical codes are assigned on a clinical encounter (*set of documents*), but most prior work focuses on prediction on a single document.
- We introduce document-level attention to learn how to aggregate the set of document representations into an encounter representation for prediction.
- On a 3M Health Information Systems internal dataset, we show that document-level attention leads to better encounter-level code prediction.
- On a small but professionally annotated dataset, we show that the document-level attention learned without document-level supervision matches professional medical coders' expectations. This suggests a potential use case of using document-level attentions to *surface evidence*.

[3]The full list of discharge summary sections we tested in Chapter 6 are (1) chief complaint, (2) family history, (3) social history, (4) medications on admission, (5) past medical history, (6) history of present illness, and (7) brief hospital course. These discharge summary sections were chosen based on their high prevalence in discharge summaries and their length diversity.

Suicidality Assessment using Social Media

- We collected the UMD Reddit SuicideWatch dataset by asking both crowdsourcers and suicidality experts to annotate a subset of the dataset's individuals.
- We reframed suicidality assessment as a hierarchical ranking problem: ranking both the individuals and their postings.
  - We identify the hierarchical structure in the suicidality assessment task: the risk of suicide is a property of an individual (*set of documents*), but the language evidence we intend to study is on a subset of the documents they posted.
  - Based on conversations with subject matter experts, we reframe suicidality risk assessment as a prioritization problem with a *limitation on the expert's time budget*.
- We introduce hierarchical Time-Biased Gain (hTBG) to measure the expected number of at-risk individuals found in a given time budget. We show both empirically and theoretically that hTBG is a desirable measure for suicidality assessment.
- We introduce 3HAN, a three-level hierarchical attention network that can jointly rank the individuals for their risk of suicidality and rank their documents for relevance to suicidality assessment, without document-level annotations.
- We show that a joint ranking model like 3HAN outperforms other plausible cascade baselines using hTBG.

Learning to Compose Discharge Summary from Prior Clinical Notes

- We introduce the task of discharge summary generation from the *set of prior documents* in the encounter.
- We derive our collection from a freely available dataset (MIMIC III); the task can potentially serve as a benchmark for clinical multidocument summarization.

- We introduce three faithfulness-based evaluation measures that are both reference-based and source-based: faithfulness-adjusted recall, faithfulness-adjusted precision, and incorrect hallucination rate.
- We propose an extractive-abstractive pipeline that supports the traceability of evidence and scales to multiple long documents.
- Results across seven commonly found discharge summary sections and five models show that a sentence-rewriting approach supporting traceability performs consistently better on our faithfulness-adjusted measures.

1.7 Outline

In the next chapter, we describe the background and related work to put our work in context. Chapter 3 discusses our work on clinical coding at the encounter level, the first example of the hierarchical structure between the patient's encounter and their clinical documents.

In Chapter 5, we propose a prioritization framing of suicidality assessment using social media and discuss the creation of the UMD Reddit Suicidality Dataset. Again, in this chapter, suicidality assessment follows a hierarchical structure similar to that of Chapter 3: the label assignment is on the at-risk individual, but the evidence is from their social media postings. Additionally, we introduce a resource limitation – time, owing to the importance of involving healthcare professionals in the loop.

In Chapter 6, we return to the clinical encounter. Instead of assisting clinical coders, we aim to help healthcare professionals write discharge summaries. Building on Chapters 3 and 5, we propose an extract-then-abstract pipeline: extracting relevant information at the document level and merging them at the patient's encounter level.

We conclude our work in Chapter 7. We also discuss limitations on the robustness of the results, effects on key stakeholders, modeling assumptions, and the need for user studies. We plan to explore modeling patients as a set of *distributed* documents and the traceability of abstractive summarization for future work.

Chapter 2: Background and Related Work

In this chapter, we put our work into context by describing related work. We first start with NLP in healthcare applications – NLP applied to the clinical domain and suicidality assessment, in particular. Section 2.1 gives a review of clinical NLP, and its role in healthcare, including clinical summarization and clinical coding. Section 2.2 gives a comprehensive review on applying NLP for suicidality assessment.

Section 2.3 describes the attention mechanism, especially hierarchical attention, and how it inspired our ELDAN model and our 3HAN model. In Section 2.4, we describe Multiple-Instance Learning (MIL) to compare and contrast their problem formulation with ours.

In Chapter 5, we propose a prioritization framing of suicidality assessment in the form of ranking, recognizing the limiting resource – healthcare professionals' time. Section 2.5 thus gives a review of rank-based and time-aware evaluations by describing them in a common framework.

2.1 Clinical NLP

NLP has the potential to unlock unstructured information in the clinical domain (Hripcsak et al., 1995). Most relevant to this dissertation, where we focus on a human-centric approach, is perhaps NLP's role in the Clinical Decision Support (CDS) systems. Hunt et al. (1998) define CDS systems as "any software designed to directly aid in clinical decision making in which characteristics of individual patients are matched to a computerized knowledge base for the purpose of generating patient-specific assessments or recommendations that are then presented to clinicians for consideration." While this definition focuses on presenting recommendations to clinicians, Demner-Fushman et al. (2009)

point out that CDS has extended beyond assisting clinicians to assist other stakeholders in healthcare. These include clinical coders, administrators (e.g., quality assessment of radiology reports, Dreyer et al., 2005; Duszak Jr et al., 2012), patients (e.g., explanation of medical terms to patients (Hardcastle and Hallett, 2007; Elhadad and Sutaria, 2007), patient-friendly documentation (Åhlfeldt et al., 2006)), students, and researchers (e.g., cohort selection using NLP, Edinger et al., 2012; Shivade et al., 2014; Stubbs et al., 2019).

Clinical NLP systems in CDS can be roughly categorized into general-purpose NLP architectures and specialized systems (Demner-Fushman et al., 2009). General purpose architectures, including LSP (Sager et al., 1987), MedLEE (Friedman et al., 1994), cTAKES (Savova et al., 2010), HiTEx (Zeng et al., 2006), and MediClass (Hazlehurst et al., 2005), often include components interacting with each other. By re-configuring these components and introducing specialized knowledge resources, these architectures can be applied to specialized tasks. Many clinical NLP systems are also directly developed for specialized tasks (Pons et al., 2016; Wu et al., 2020). Examples include, but are certainly not limited to, adverse event detection from clinical notes (Murff et al., 2003; Dandala et al., 2019), using question answering to support evidence-based practice (Demner-Fushman et al., 2008), and de-identification of clinical notes (Yadav et al., 2016; Heider et al., 2020). The clinical coding described in Chapter 3 and discharge summary composition described in Chapter 6 are also examples of these specialized systems, and we further discuss them in detail below.

Clinical Coding. Computer-Assisted Coding, or CAC, dates back at least to 1973, when Dinwoodie and Howell (1973) proposed a dictionary matching method based on clinical code descriptions. Since the 1990s, a growing literature has introduced natural language processing techniques to address the task of automatic clinical coding using unstructured text (Resnik et al., 2006; Pestian et al., 2007; Zhang et al., 2017; Mullenbach et al., 2018; Xie and Xing, 2018, and many others). Many of these studies, however, focus on limited categories of codes, such as variations of pneumonia, or only analyze specific subsets of clinical documents, such as radiology notes or discharge summaries (Stanfill et al., 2010). Progress on this problem using state of the art techniques has also been ham-

pered significantly by the broader research community's limited access to large, shareable datasets (Resnik, 2018).

Deep learning models have been applied to CAC, some exploiting attention mechanisms to support explainability (Shi et al., 2017; Baumel et al., 2018; Mullenbach et al., 2018; Xie and Xing, 2018). Crucially, however, these all look solely at the discharge summaries, which are assumed to condense information about a patient encounter. As we argue in Chapter 3, this can be problematic: Kripalani et al. (2007), reviewing 73 published studies investigating hospital communication and information transfer, find high rates of information missing from discharge summaries, notably 17.5% missing the main diagnosis. To our knowledge, our work is the first to investigate the hierarchical structure of the encounter as a whole.

Clinical Summarization. Most literature on clinical summarization focuses on extractive summarization, due to the risk involved in a clinical application (Demner-Fushman and Lin, 2006; Feblowitz et al., 2011; Pivovarov and Elhadad, 2015; Moen et al., 2016; Liang et al., 2019). For abstractive summarization, summarization of radiology reports has been a topic of interest in NLP research recently. Zhang et al. (2018) show promising results generating the assessment section of a chest x-ray radiology report from the findings and background section. MacAvaney et al. (2019) improved this model through the incorporation of domain-specific ontologies. However, such generated reports may not be clinically sound, and the models generate sentences inconsistent with the patient's background. Therefore, in subsequent work, Zhang et al. (2020) add a reinforcement learning based fact-checking mechanism to generate a clinically consistent assessment. Lee (2018) explores the generation of the *Chief Complaint* of emergency department cases from age group, gender, and discharge diagnosis code. Ive et al. (2020) follow a closely related approach of extracting keyphrases from mental health records to generate synthetic notes. They further evaluate the quality of generated synthetic data for downstream tasks. Work from Lee (2018) generates clinical notes by conditioning transformer-based models on a limited window of past patient data.

In Chapter 6, instead of focusing on purely extractive or abstractive clinical summa-

rization, we use an extractive-abstractive pipeline as a framework for clinical multidocument summarization. The extractive-abstractive pipeline first extracts relevant snippets from source documents and then merges them using an abstractive system. This framework is commonly found in other summarization settings that involve extensive documentation. For example, Jing and McKeown (1999) suggest that humans use a similar two-stage strategy to summarization long documents. Recent works on abstractive multidocument summarization also often include a coarse extractive model that limits the number of paragraphs before abstraction (Liu et al., 2018; Liu and Lapata, 2019a).

2.2 NLP for Suicidality Assessment

There is an extensive clinical literature on suicidality assessment (e.g., Batterham et al. (2015); Joiner Jr et al. (1999); Joiner et al. (2005)), but very little specifically looking at assessment of suicidality based on social media content. This is a new topic that has received very little study to date in the clinical literature, with prior work focusing on non-expert rather than healthcare professionals' judgments (Egan et al., 2013; Corbitt-Hall et al., 2016). Griffiths et al. (2010) present a review of randomized controlled trials involving Internet interventions for depression and anxiety disorders. Lind et al. (2017) offer a comprehensive discussion of crowdsourcing, using CrowdFlower, as a means for obtaining coding of latent constructs in comparison with content analysis.

Calvo et al. (2017), Guntuku et al. (2017), Resnik et al. (2020), and Harrigian et al. (2021) present reviews of NLP research in which social media are used to identify people with psychological issues who may require intervention, and Conway and O'Connor (2016) provide a shorter survey focused on public health monitoring and ethical issues, highlighting the annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych), initiated in 2014, as a forum for bridging the gap between computer science researchers and mental health clinicians (Resnik et al., 2014). Recent CLPsych shared tasks using data from the ReachOut peer support forums have provided opportunities for exploration of technological approaches to risk assessment and crisis detection (Milne et al., 2016; Milne, 2017); see also Yates et al. (2017); Losada et al. (2020); Goharian

et al. (2021); and the 2019 CLPsych workshop (Zirikly et al., 2019) that uses the dataset we collected in Chapter 4.

Although predictive modeling for risk assessment is a burgeoning area, a key challenge for work on mental health in social media is connecting the clinical side with available social media datasets (Ernala et al., 2019; Harrigian et al., 2021). Combining ground truth health record data with social media data is rare, with Padrez et al. (2015) representing a promising exception; they found that nearly 40% of 5,256 Facebook and/or Twitter users who were approached in a hospital emergency room consented to share both their health record and social media data for research.[1] Approximations of clinical truth are more common. For example, self-report of diagnoses in social media (Coppersmith et al., 2014), or observed user behaviors such as posting on SuicideWatch (De Choudhury et al., 2016). Coppersmith et al. (2015, 2016) employed the Twitter data collection method of Coppersmith et al. (2014) to discover Twitter users with self-stated reports of a previous suicide attempt in order to identify valuable signal and support automated classification.

In work similar to Chapter 4, Vioulès et al. (2018) applied a similar data collection approach to Coppersmith et al., searching Twitter for tweets containing key phrases based on risk factors and warning signs identified by the American Psychiatric Association and the American Association of Suicidology. They defined a four-category scale for distress and 500 tweets were annotated by researchers, with a subset of 55 validated by a psychologist. They achieved 69.1% and 71.5% chance-corrected agreement using Cohen's kappa and weighted kappa, respectively, with Fleiss kappa of 78.3% for the 55 tweets with three annotators; for automated classification they explored eight text classifiers and a variety of features, with their best performing combination for four-way classification achieving an F_1 of 0.518.

2.3 Attention Mechanisms

Many of the recent advances in predictive modeling for clinical NLP and mental health NLP are based on deep learning, an approach that allows the model to learn represen-

[1] Interestingly, participants agreeing to social media access were only slightly younger on average than those who declined (29.1 ± 9.8 versus 31.9 ± 10.4 years old.

tations and the relation between those representations. Attention mechanisms in deep learning allow the model to focus on specific "regions" of its input data, which has proven helpful in, among others, machine translation (Bahdanau et al., 2015), summarization (Rush et al., 2017; See et al., 2017; Liu and Lapata, 2019b), and sentiment analysis (Yang et al., 2016). It is also the central component in the recent Transformer-based architectures (Vaswani et al., 2017; Peters et al., 2018; Devlin et al., 2019; Lewis et al., 2019). Attention, especially in the context of NLP, has two main advantages: it allows the network to attend to meaningful parts of a sequence (either words or sentences), often leading to improved performance, and it provides insight into which parts of the sequence are being used to make the prediction.[2]

Building on Bahdanau et al. (2015), Yang et al. (2016), in the context of document classification, proposed a hierarchical attention mechanism based on dot product attention. They observe a hierarchical structure in a document: a document can be represented as a sequence of sentences, and a sentence can be represented as a sequence of words. Applying the hierarchical attention mechanism on both the word level and the sentence level, Hierarchical Attention Network (HAN) learns to pay attention to specific words in a sentence to form a sentence representation, and at the next higher level to weigh specific sentences in a document in forming a document representation.

Both our ELDAN model (Chapter 3, for clinical coding) and our 3HAN model (Chapter 5, for suicidality assessment) draw inspiration from Yang et al. (2016). In contrast to HAN, ELDAN and 3HAN move up the representational hierarchy, learning also to weight documents to form representations of encounter (ELDAN) and representations of individual (3HAN). In ELDAN, instead of building the representation from the word level, we directly use sparse document features provided by more traditional feature extraction methods grounded in subject matter knowledge and resources, e.g., UMLS. This allows us to alleviate the problem of out-of-vocabulary and uncommon abbreviation terms often found in clinical notes, therefore requiring less training data, which is beneficial in a setting where many codes are rare. In 3HAN, we apply three levels of hierarchical at-

[2]However, whether attention provides accurate explanation is debated. See discussion from Jain and Wallace (2019); Wiegreffe and Pinter (2019); Wallace (2019)

tention mechanisms: on the word level, on the sentence level, and, importantly, on the document level. The added document-level attention mechanisms in ELDAN and 3HAN allow us to represent an encounter or an individual as a set or sequence of documents, reflecting the hierarchical structure we observed in the individual assessment process.

2.4 Multiple-Instance Learning

Besides the hierarchical attention mechanism, the hierarchical structure we observed in the assessment process has a similar problem formulation to that of Multiple-Instance Learning (MIL, Dietterich et al., 1997; Maron and Lozano-Pérez, 1997). In contrast to conventional supervised machine learning matching an instance X to a target Y, the focus of MIL is to match a set of $X = \{X_1, X_2, \cdots, X_n\}$ (known as a bag of instances) that is permutation-invariant (i.e., changing the order within the bag does not change the result) to a target Y. Another standard assumption often made is that a single instance of $\{X_1, X_2, \cdots, X_n\}$ in the bag being positive is enough to justify Y being positive. In contrast, a negative Y implies all $\{X_1, X_2, \cdots, X_n\}$ in the bag are negative (Dietterich et al., 1997; Carbonneau et al., 2018). We will call this the max-pooling assumption for convenience.

While most of the work on MIL uses a mean-pooling operator or a max-pooling operator to aggregate instances to a bag (Andrews et al., 2002; Settles et al., 2007; Feng and Zhou, 2017; Pinheiro and Collobert, 2015; Zhu et al., 2017), recent work has started to use attention-based operators to learn the different levels of contributions the instance may have to the final prediction (Ilse et al., 2018). Interestingly, the attention-based operator is almost identical to the dot product-based attention found in Bahdanau et al. (2015), Yang et al. (2016), and our document attention models ELDAN and 3HAN.

Chapter 3 and Chapter 5 differ from MIL in that our interest is modeling the individual assessment process. This leads to a hierarchical ranking formulation, in contrast to the MIL formulation of assigning a label to a set of instances. For both ELDAN and 3HAN, we investigate quantitatively how well document-level attention matches expert expectations. In Chapter 5, we introduce hTBG, which jointly evaluates, in MIL terminology, the

ranking of the bag and the ranking of the instances.

2.5 Ranking Evaluation

There is an extensive information retrieval literature on evaluation measures for ranked lists (Järvelin and Kekäläinen, 2002; Chapelle et al., 2009; Smucker and Clarke, 2012; Sakai, 2019). Many of these rank-based evaluation measures assume a user is working down a ranked list. This simple user model (assumption of how users interact with the system) leads to evaluation measures that generally reward placing highly relevant items near the top of the list, and are often relatively insensitive to mistakes made near the bottom. Carterette (2011) and Clarke et al. (2011), among others, point out that ranking measures can often be expressed as:

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\infty} g_k d_k, \tag{2.1}$$

where g_k corresponds to the gain (i.e. value) of placing the item at rank k, and d_k is a discount factor for the position k. Normalization $\mathcal{N}$ enables comparison across queries.

For example, one common formulation of Discounted Cumulative Gain (DCG, Järvelin and Kekäläinen, 2002), a well-known rank-based evaluation measure for graded relevance (i.e., items are annotated with multiple degrees of relevance), can be expressed as:

$$DCG@K = \sum_{k=1}^{K} \frac{2^{rel_k} - 1}{log_2(k+1)} \tag{2.2}$$

where it models the gain, g_k as $2^{rel_k} - 1$, with relevance, rel_k, being the relevance of the item at position k, and its discount, d_k as $\frac{1}{log_2(k+1)}$. The parameter K indicates that the ranked list is cut off at position K. With normalization, Normalized DCG, or NDCG can be defined as:

$$NDCG@K = \frac{DCG@K}{IDCG@K} \tag{2.3}$$

where the ideal DCG@K, $IDCG@K$, is the maximum achievable score with perfect

partial ranking, cutoff at position K.

Another example, Expected Reciprocal Rank (ERR, Chapelle et al., 2009) assumes that as the user works down a ranked list, they are more likely to stop after viewing a highly relevant item than after viewing an irrelevant one, as their information need is more likely to have been satisfied. This results in a cascade model of user behavior:

$$\sum_{k=1}^{\infty} \frac{1}{k} P\left(\text{user stops at k}\right) \tag{2.4}$$

where discount at position k, d_k is $\frac{1}{k}$ and gain at position k, g_k is defined as:

$$P\left(\text{user stops at k}\right) = R_k \prod_{i=1}^{k-1} \left(1 - R_i\right) \tag{2.5}$$

where $R_k = f(rel_k)$ is the probability to stop at position k, as a function of the relevance of the item at position k.

In the setting of Chapter 5, suicidality risk assessment, we care about how much gain (number of at-risk individuals found) can be achieved for a given time budget. Time-biased gain (TBG, Smucker and Clarke, 2012) measures this by assuming a determined user working down a ranked list, with the discount being a function of the time it takes to reach that position:

$$\text{TBG} = \sum_{k=1}^{\infty} g_k D\left(T\left(k\right)\right). \tag{2.6}$$

where $D(\cdot)$ is a function of time and $T(k)$ is the expected amount of time it takes a user to reach position k. For a detailed description of TBG, see Section 5.3.1.

However, neither TBG nor other ranking measures, to the best of our knowledge, can measure the *hierarchical* ranking found in the scenario that motivates our work in Chapter 5: ranking items (i.e. individuals) when each item itself contains a ranked list of potential evidence (their posts). In Chapter 5, we design a new metric, hierarchical time-biased gain (hTBG), to measure the hierarchical ranking by incorporating the cascading user model found in ERR into TBG.

Chapter 3: Assigning Clinical Codes at the Encounter Level by Allocating Attention to Documents

The vast majority of research in computer-assisted clinical coding focuses on coding at the document level, but a substantial proportion of clinical coding in the real world involves coding at the level of the patient's clinical encounters, each of which is typically represented by a potentially large set of documents.

Recall that in Chapter 1, we describe a hierarchy between the patient and their documents. This chapter introduces encounter-level document attention networks, which model the encounter as a set of documents and use hierarchical attention to explicitly take the hierarchical structure between the patient's encounter and their documents into account. Experimental evaluation demonstrates improvements in coding accuracy as well as facilitation of human reviewers in their ability to identify which documents within an encounter play a role in determining the encounter level codes.[1]

3.1 Computer-Assisted Coding

Clinical coding translates unstructured information about diagnoses, treatments, procedures, medications and equipment into alphanumeric codes for billing purposes. Coding is challenging and expensive, requiring high-expertise professionals, and even experienced coders frequently disagree with each other (Resnik et al., 2006). Increasingly, **c**omputer-**a**ssisted **c**oding (CAC) is used to help address these issues by automatically suggesting clinical codes, generally within a workflow that supports subsequent human review to ensure that codes are correct or to make revisions.

[1]This chapter contains content from: **Shing, Han-Chin**, Guoli Wang, and Philip Resnik. "Assigning Medical Codes at the Encounter Level by Paying Attention to Documents." In ML4H, Machine Learning for Health Workshop at NeurIPS. 2019.

The vast majority of relevant literature focuses on automatic code assignment at the document level, such as radiology reports (e.g., Farkas and Szarvas, 2008) or discharge summaries (e.g., Perotte et al., 2013; Stanfill et al., 2010). However, in many settings codes are assigned not to individual documents, but to an entire clinical *encounter*, such as a patient visit to a hospital. Encounter-level documentation often involves multiple documents (O'Malley et al., 2005), and the relationship between the encounter-level codes and the unstructured information in the documents is indirect — so the standard approach, treating coding as a well understood kind of text classification problem (e.g., Pang et al., 2002; Wang and Manning, 2012; Yang et al., 2016), does not map naturally to document *collections*.

This can be problematic: Kripalani et al. (2007) find high rates of information missing from discharge summaries, which are the focus of most prior research (Stanfill et al., 2010; Shi et al., 2017; Baumel et al., 2018; Mullenbach et al., 2018; Xie and Xing, 2018). Notably, discharge summaries miss 17.5% of the main diagnosis, which would therefore need to be identified from other documentation in the encounter. In addition, for outpatient encounters, discharge summaries are rarely a part of the record.[2] One obvious solution, using document-level models and then merging their predictions into encounter codes, immediately runs up against a lack of training data: clinical coders do not identify which documents are the "source" for each encounter code. In addition, merging document-level codes involves non-trivial interactions. For example, specific codes suppressing more general codes (O'Malley et al., 2005).

In this chapter, we instead focus on training an *encounter-level* model directly. One straightforward approach would be to aggregate (via sum or average) all document features into a single encounter feature set, but this would be noisy, as the signal of the targeted clinical code is diluted when irrelevant documents are also included. It also fails to address the crucial problem of *interpretability*: human coders reviewing auto-suggested codes need to relate proposed encounter codes back to document-level evidence.[3] We therefore introduce a new approach to encounter-level coding, observing that its structure

[2]The patient is generally not admitted to the facility, and thus will not be discharged.

[3]Interpretability is also important from a technical perspective, to identify problems in the prediction model.

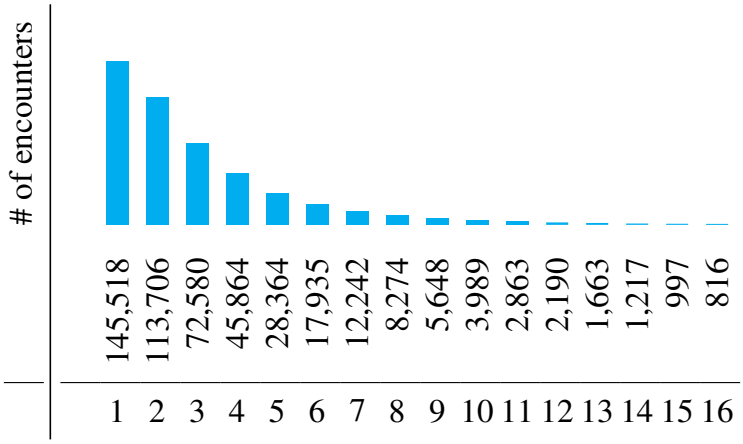

Table 3.1: Histogram of the number of documents in an encounter.

is essentially hierarchical, progressing from textual evidence up to documents, and from there to entire encounters. Our Encounter-Level Document Attention Network (ELDAN) applies the key insights of hierarchical attention networks (HAN, Yang et al., 2016), enabling the model to identify which documents are most relevant in encounters as driven by the encounter-level task. We obtain positive results for encounter-level labeling in comparison to a strong, realistic baseline, and also show that the resulting weighting helps coders identify which documents are likely sources for a code. In Chapter 5, we further explore how the learned importance of documents (i.e., document attention) can be used to surface evidence to help healthcare professionals .

3.2 Datasets

We used outpatient procedure (CPT) coding production data internal to 3M Health Information Systems, a leading provider of CAC solutions, sampled from multiple hospital sites. Our dataset includes 463,866 coded encounters containing 1,390,605 documents, with 31% of encounters containing a single document; in the remainder, there are an average of 3.91 documents per encounter. Table 3.1 shows a histogram of encounters that contain a specific number of documents. We generated a random 80-10-10 training/tuning/evaluation split by encounter ID. Coding exists only at the encounter level, with no indication of which codes are associated with which document(s). To minimize

the risk of inappropriate protected health information (PHI) transmission even internally within 3M HIS, once documents were selected, they were immediately converted from their original form to feature vectors. These features include UMLS CUIs (Concept Unique Identifiers) and 3M HIS internal numeric concept identifiers, as well as words or phrases (775,330 unique features) for all downstream machine learning development and experimentation. No PHI contributed to the features used to represent documents.

In addition, to assess the value of document-level attention in identifying which documents are responsible for encounter codes (for facilitating human code review) we extracted a separate dataset from production data, consisting of 393 encounters. To eliminate possible leakage across experiments these do not overlap with the first set. For each encounter, experienced clinical coders annotated *document*-level codes corresponding to the encounter-level coding. Specifically, coders were instructed to read through all the documents contained in the encounter, and assign a code from the encounter level to the document if (and only if) it contains sufficient evidence for assigning the code. Note this means the same code can be assigned to multiple documents within the encounter.

3.3 Model: Encounter-Level Document Attention Network

Encounter-level coding can be considered as a multi-label classification problem. We decompose the problem into multiple one-vs-all binary classification problems, each targeting one code, which adds flexibility for use cases where codes of interest could vary across sites or even dynamically, and also facilitates comparing code-specific document attention learned from the model to document annotations labeled by clinical coders, in our evaluation in Section 3.4.

The overall architecture of Encounter-Level Document Attention Networks (EL-DAN, Figure 3.1) consists of three parts: (1) a document-level encoder that turns sparse document features into dense document features using an embedding layer followed by two fully connected layers, (2) a document-level attention layer that draws inspiration from Yang et al. (2016), and (3) an encounter-level encoder using a fully connected layer.

We first introduce notation and then describe the three parts in more detail. Let

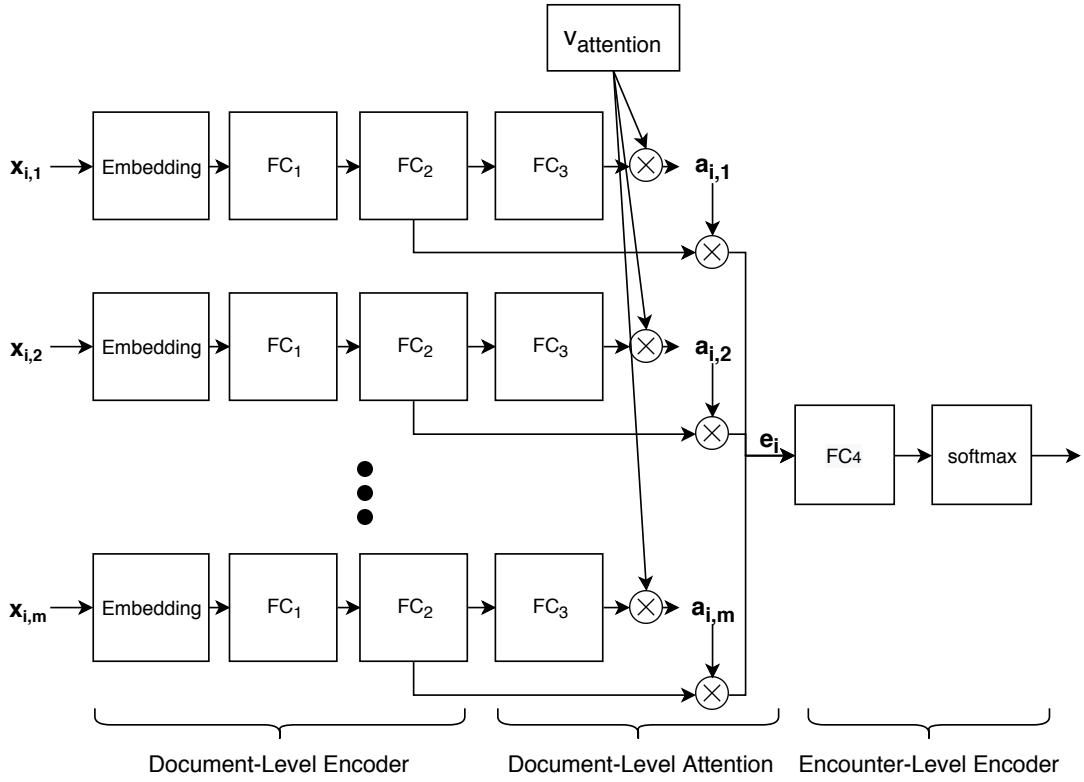

Figure 3.1: Architecture of Encounter-Level Document Attention Network (ELDAN)

the set of encounters be $E = \{e_1, e_2, \cdots, e_n\}$, and their corresponding labels be $Y = \{y_1, y_2, \cdots, y_n\}$, where $y_i \in \{-1, 1\}$ represents whether the encounter e_i contains the targeted medical code c_t. Each encounter e_i consists of multiple documents, and the number of documents that an encounter contains can vary across encounters. Finally, let $x_{i,j}$ and $d_{i,j}$ be the sparse and dense feature vectors that represent document j in encounter i, respectively.

Document-Level Encoder. The goal of the document-level encoder is to transform a sparse document representation, $x_{i,j}$, into a dense document representation, $d_{i,j}$. The sparse document representation, $x_{i,j}$ is first passed into an embedding layer, to map the 775,330-dimensional sparse document representation into a 300-dimensional vector. It is then followed by two fully-connected layers, FC_1 and FC_2, to produce a dense document representation, $d_{i,j}$. Specifically,

$$h_{i,j,0} = W_{Embedding} x_{i,j} \tag{3.1}$$

$$h_{i,j,1} = tanh\left(W_{FC_1} h_{i,j,0} + b_{FC_1}\right) \tag{3.2}$$

$$d_{i,j} = tanh\left(W_{FC_2} h_{i,j,1} + b_{FC_2}\right) \tag{3.3}$$

where W represents the weight matrix, b represents a bias vector, and tanh is the hyperbolic tangent. $h_{i,j,0}$ and $h_{i,j,1}$ are hidden representations of document j in encounter i.

Document-Level Attention. When a clinical code is assigned to an encounter, it does not imply that all its documents contain evidence for that code. Directly summing or averaging all the encounter's dense document representations, $\{d_{i,1}, d_{i,2}, \cdots, d_{i,m}\}$, will typically capture irrelevant information, diluting the signal for the presence of the code. Instead, ELDAN computes a weighted average, where more relevant documents receive more attention. This is calculated by comparing the dense document representation, $d_{i,j}$, to a learnable attention vector, $v_{attention}$, after passing through a fully connected-layer and a non-linear layer. Specifically,

$$u_{i,j} = tanh\left(W_{FC_3} d_{i,j} + b_{FC_3}\right) \tag{3.4}$$

$$a_{i,j} = \frac{exp\left(u_{i,j}^{\top} v_{attention}\right)}{\sum_{j=1}^{m} exp\left(u_{i,j}^{\top} v_{attention}\right)} \tag{3.5}$$

$$e_i = \sum_{j=1}^{m} a_{i,j} d_{i,j} \tag{3.6}$$

where $a_{i,j}$ is the normalized attention score for document j in encounter i, and e_i is the encounter representation of encounter i. As shown in Equation 3.5, the transformed document representation $u_{i,j}$ is compared with the learnable attention vector $v_{attention}$ using dot product, and further normalized for the weighted averaging step in Equation 3.6.

Encounter	ELDAN's Document Attention	Human Coders
enc_1	$[\mathbf{doc_1}, doc_2, doc_3]$	$[\mathbf{doc_1}, doc_2, doc_3]$
enc_2	$[\mathbf{doc_4}]$	$[\mathbf{doc_4}]$
enc_3	$[doc_5, \mathbf{doc_6}, doc_7, doc_8]$	$[doc_5, \mathbf{doc_6}, doc_7, doc_8]$

Table 3.2: An illustration of how ELDAN's document attention predictions are evaluated using source documents labeled by human coders. $Green$ (the shading under Human Coders) indicates the "source" documents for the encounter-level code (truth), and $Gray$ (the shading under Eldan's Document Attention) indicates the documents with high attention (prediction). The bolded documents are the true positives. In this example, the precision is $\frac{tp}{tp+fp} = \frac{3}{3+2} = \frac{3}{5}$. The recall is $\frac{tp}{tp+fn} = \frac{3}{3+1} = \frac{3}{4}$. The document-level F_1 score is thus $\frac{2}{3}$.

Encounter-Level Encoder. Once we have the encounter representation e_i, we can predict whether the encounter contains the targeted medical code. Specifically,

$$P(\hat{y}_i) = softmax\left(W_{FC_4} e_i + b_{FC_4}\right) \tag{3.7}$$

Finally, we compare with the ground truth label of encounter i using negative log-likelihood to calculate a loss $-log\left(p(\hat{y}_i = y_i)\right)$ on encounter i, where y_i is the ground-truth label.

3.4 Experiments

Our first validation experiment tests ELDAN's effectiveness for predicting encounter-level codes. The second looks at the value of document-level attention from ELDAN as a prediction of which documents in the encounter can be considered the "source" for the encounter-level codes.

3.4.1 Evaluating Encounter-Level Code Prediction

We train two ELDAN models. One is a standard ELDAN model. The other, which we refer to as ELDAN+TRANSFER, includes a simple but effective enhancement for handling rare codes, since, when the code is rare, training a deep one-vs-all network can be challenging. To address this issue, we use a naïve transfer learning technique that initializes the embedding layer ($W_{Embedding}$) with that of a trained model on a more frequent

code. See Section 3.4.3 for more details. We call this naïve, as it is clearly not optimal nor novel, but the results demonstrate a potentially promising direction for training classifiers for rare clinical codes in settings where a single multi-label classifier may be less desirable for other reasons, as discussed in Section 3.3. We measure performance in standard fashion using the F_1 score.

We regard Yang et al. (2016)'s non-attention hierarchical network (HN-AVE in their paper) as a strong baseline since, in experiments across six document classification datasets, they demonstrated that it substantially outperformed a range of typical baselines lacking hierarchy; these included, for example, bag of words, SVM, LSTM, and CNN classifiers. Analogously, we define ELDN (encounter level document network) as a baseline that simply averages documents rather than using attention.[4]

3.4.2 Relevant-Document Prediction against Human Judgments

To evaluate the extent to which document attention learned by ELDAN matches human clinical coders' judgments about the documents relevant for coding the encounter, we apply our trained models to our second dataset. Recall that this is a separate set of 393 encounters for which a team of experienced clinical coders annotated codes at the document level. We calculate *document-level F_1-score* by treating document attention learned from ELDAN as the prediction of which documents are the "source", and comparing this to clinical coders' ground truth — see Table 3.2 for an illustration. Note that this is different from the encounter-level F_1 scores used to evaluate encounter-level code prediction.

To determine which *documents* are predicted to contain targeted codes and therefore are relevant for human code review of the encounter-level coding, we pass the annotated dataset through the ELDAN model trained for encounter-level code prediction, with no further tuning or training. We then use a selection strategy that takes the attention scores of all the documents in an encounter and marks all documents that are strictly larger than half the maximum attention score as containing the targeted code. Since a baseline to compare with document-level attention can be non-trivial to implement, in the spirit of having a

[4]Note that most prior methods for clinical coding base the prediction on a single discharge summary (which is rarely present in outpatient encounters), and are thus not applicable as baselines in our setting.

chance-adjusted measure, we compare with a baseline obtained by randomly generating attention scores from a uniform distribution on the documents within an encounter, then following the same selection strategy as in ELDAN's document attention selection. The chance baseline is run 500 times to reduce the noise level.

3.4.3 Training Details

Our 80-10-10 dataset split results in 371,092 encounters for training, 46,387 encounters for development/tuning, and 46,387 encounters for testing. Note that no document-level annotations are available. We train models implemented with PyTorch (Paszke et al., 2017) on the 150 most frequent codes, using minibatch stochastic gradient descent (Sutskever et al., 2013) with a minibatch size of 64, a learning rate of 0.01, and a momentum of 0.9. Since we are in an imbalanced setting (some medical codes can be extremely rare, see Fig. 3.4), we randomly resampled the training data by assigning different probabilities to the positive and negative classes so that the ratio of positive encounters to negative encounters is close to $1:6$. These hyperparameters were selected based on our results on the development set. No resampling is done for the development set and test set.

For naïve transfer learning, models are trained from the most frequent code to the least frequent. The model for the most frequent code is trained from scratch just like ELDAN. For all the other models, the weight of the (n)-th most frequent model's embedding layer ($W_{Embedding}$, see Equation 3.1) is first initialized (but not fixed) by that of the $(n-1)$-th most frequent model prior to training.

3.5 Results and Discussion

Results Evaluating Encounter-Level Code Prediction. ELDAN numerically outperforms the baseline for 17 of the most frequent 20 codes (Table 3.3). Comparing across 150 codes, ELDAN also outperforms ELDN.[5] To show the trend across the full range of codes we macro-average every 10 codes from most frequent to least frequent (Table 3.4).

[5]Statistical significant using paired t-test across 150 codes at $p < 0.05$

CPT Codes	#Docs	Prevalence	ELDN	ELDAN	ELDAN +TRANSFER
43239	3.13	4.15%	84.59	**86.21**	84.93
45380	2.78	3.56%	72.68	**75.14**	74.02
45385	2.75	2.44%	71.33	**72.33**	70.31
66984	2.51	1.90%	92.15	92.87	**93.00**
45378	2.40	1.89%	62.67	65.45	**67.57**
12001	2.20	1.60%	**46.96**	44.74	43.62
12011	2.35	1.19%	41.03	42.12	**43.30**
29125	2.85	1.05%	52.32	**56.50**	54.10
10060	2.09	1.00%	45.15	48.73	**52.25**
69436	3.01	0.96%	83.30	85.18	**88.32**
12002	2.60	0.92%	25.53	28.36	**28.43**
59025	1.86	0.92%	**73.82**	69.00	67.73
11042	3.20	0.88%	64.38	63.45	**66.86**
47562	4.36	0.80%	70.74	76.25	**77.67**
62323	2.10	0.79%	61.17	57.07	**64.25**
Average	2.62		58.02	60.40	**61.26**

Table 3.3: Encounter-level F_1-scores of the 20 most frequent CPT codes. #Docs is the average number of documents found in the encounters that contain the code; prevalence is the percentage of all encounters that contain that code.

ELDAN with or without naïve transfer learning consistently outperforms ELDN, even for extremely rare codes ($< 0.1\%$). As codes become rarer, ELDAN+TRANSFER tends toward outperforming ELDAN more substantially; see increasing trend for ΔELDAN. This improvement can be explained by viewing the embedding layer as a vector space model that maps sparse features that are extracted from the document (such as medical concepts, UMLS CUIs) to a dense representation, which can be effective for bootstrapping the training of rare codes.

Results Evaluating Relevant-Document Prediction against Human Judgments. Table 3.5 shows document-level F_1-scores for the most frequent 20 encounter-level codes, with surprisingly strong results: 100% F_1-score on 7 out of 19 available codes.[6] However, even chance performance could be good if the number of possible documents to assign credit to is very small. As an extreme case, performance for code 51072 is evaluated on two encounters, each of which contains only a single document (Table 3.5), though this is atypical. Therefore we compare to the chance baseline. ELDAN is consistently better

[6]Note that as the dataset is smaller and disjoint from the training dataset, codes can be missing (such as code 59025).

Average	Prevalence	ELDN	ELDAN	ELDAN +TRANSFER	ΔELDAN
1st to 10th	1.97%	65.22	66.93	**67.14**	0.22
11st to 20th	0.78%	50.82	53.87	**55.38**	1.50
21st to 30th	0.51%	55.93	**63.07**	62.23	-0.85
31st to 40th	0.40%	44.93	51.92	**55.24**	3.32
41st to 50th	0.30%	32.08	38.61	**39.35**	0.74
51st to 60th	0.26%	33.83	38.80	**39.10**	0.30
61st to 70th	0.23%	28.37	35.05	**36.62**	1.56
71st to 80th	0.21%	25.66	30.62	**32.93**	2.31
81st to 90th	0.18%	34.92	42.03	**43.26**	1.23
91st to 100th	0.16%	24.54	29.06	**31.32**	2.25
101st to 110th	0.14%	25.15	33.17	**34.57**	1.40
111st to 120th	0.12%	24.87	31.74	**32.84**	1.09
121st to 130th	0.11%	18.14	24.10	**28.09**	3.99
131st to 140th	0.10%	20.39	28.53	**32.21**	3.68
141st to 150th	0.08%	26.93	33.13	**40.94**	7.82

Table 3.4: Macro average of encounter-level F_1 scores for every 10 codes (from most to least frequent). $\Delta\text{ELDAN} = \text{ELDAN+TRANSFER} - \text{ELDAN}$.

except for one code, usually by a large margin.[7] These results support the conclusion that ELDAN's document attention is effective in identifying signal from "source" documents for the targeted code — crucially, without training on document-level annotations.

3.6 Effectiveness of Document-level Attention

In this chapter, we have introduced a new approach to encounter-level coding that explicitly takes the hierarchical structure of the patient's encounter and their documents into account. Experimental validation of the model shows that *document-level attention* improves coding accuracy against a strong baseline. It also supports the conclusion that the assignment of document-level attention would provide value in helping human coders to identify document-level evidence for encounter-level codes during review.

These results inspired us to further investigate document-level attention. We explore this further in the context of suicidality risk assessment. In Chapter 4, we first collect a dataset for risk assessment of suicidality via online postings. Chapter 5 investigate how document-level attention, learned jointly with an individual's risk of suicidality, can be used to re-rank documents, ultimately saving healthcare professionals' assessment

[7] Improvement is significant at $p < .05$ using a one-sample t-test comparing the population mean of average F_1 over the 500 chance baseline runs against the document-level F_1 obtained using the document attention model.

CPT Codes	#enc	#doc	#source	Attention	Chance	Diff
43239	8	19	9	88.89	59.22	29.67
45380	5	11	5	90.91	56.47	34.44
45385	6	13	8	85.71	67.52	18.20
66984	7	13	7	100.00	68.65	31.35
45378	10	20	11	90.91	67.44	23.47
12001	1	3	1	100.00	45.63	54.37
12011	3	8	3	57.14	54.30	2.85
29125	2	9	4	72.73	50.91	21.81
10060	4	9	6	100.00	71.65	28.35
69436	7	18	8	87.50	60.54	26.96
12002	4	13	6	92.31	56.02	36.29
59025	0	0	0	-	-	-
11042	5	23	16	58.06	64.89	-6.82
47562	1	5	3	100.00	57.62	42.38
62323	5	11	7	87.50	69.85	17.65
64483	3	8	4	100.00	58.07	41.93
43235	6	18	6	83.33	45.19	38.15
20610	5	9	5	100.00	72.25	27.75
49083	10	27	13	85.71	60.21	25.50
51702	2	2	2	100.00	100.00	0.00

Table 3.5: Document-level F_1-score calculated by comparing document attention from ELDAN and human coders on 20 CPT codes. #enc is the number of encounters that contain the code. #doc is the number of documents within those encounters. #source is the number of documents being labeled by human coders as the source documents for the code. Attention (from ELDAN) and Chance both report document-level F_1-score, and Diff is the difference between them.

time. Using a new evaluation measure, hTBG, we show that document-level attention can potentially lead to faster and better suicidality assessment.

Chapter 4: Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings

Chapter 3 introduces a hierarchical structure between patients and their encounter documents in the context of clinical coding. Results suggest that document attention can potentially facilitate human review by surfacing relevant documents. We further investigate this hierarchical structure and how document-level evidence can be surfaced in another setting: assessing suicide risk via online postings.

This chapter describes the creation of the University of Maryland Reddit Suicidality Dataset, which will later be used for our experimentation of document attention in Chapter 5. In this setting, the risk of suicide is a property of the individual, but the evidence is found in the documents posted on their social media. The resulting dataset contains three sets of disjoint individuals with an increasing level of annotation quality: rule-based *weak supervision*, *crowdsource* workers, and suicidality assessment *experts*. For the dataset with expert annotations, we additionally obtain information on which document most supports their judgment. Evaluation of risk-level annotations by experts yields what is, to our knowledge, the first demonstration of reliability in risk assessment by healthcare professionals based on social media postings.[1]

4.1 Suicidality Assessment via Online Postings

The majority of assessment for suicide risk takes place via in-person interactions with clinicians, using ratings scales and structured clinical interviews (Batterham et al., 2015; Joiner Jr et al., 1999; Joiner et al., 2005). However, such interactions can take place

[1]This chapter contains content from: **Shing, Han-Chin**, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. "Expert, crowdsourced, and machine assessment of suicide risk via online postings." Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018.

only after patient-clinician contact has been made, and only when access to a clinician is available. This is no small challenge in many places — in the U.S., for example, nearly 122 million people live in federally designated mental health provider shortage areas in 2020, where access to a provider can be difficult even when the person (or someone close to them) knows that clinical help is needed (Bureau of Health Workforce, 2020).

At the same time, people are spending an increasing amount of their time online, and online discussions related to mental health are providing new opportunities for people dealing with mental health issues to find support and a sense of connection: examples include Koko, ReachOut, 7cups, SuicideWatch on Reddit, and others.[2] Although many such discussions are peer-to-peer, site moderators often play a crucial role, identifying users who post material indicating imminent risk and the need for intervention.

An emerging subset of the artificial intelligence and language technology communities has been making progress toward automated methods that analyze online postings to flag mental health conditions, with the goal of being able to screen or monitor for suicide risk and other conditions (Calvo et al., 2017; Resnik et al., 2014; Milne et al., 2016; Milne, 2017; Losada et al., 2020; Goharian et al., 2021). Some sites have been taking advantage of these methods to add automation to their moderation, in the form of a pipeline from algorithmic risk assessment to human moderator review to preventive action.

With all of these technology-driven developments taking place so quickly, it is easy to forget that a *healthcare professional*'s assessment of suicidality from online writing is a new and largely unstudied problem. To what extent is level of suicide risk discernible from online postings? How are traditional training and experience in assessment brought to bear in the absence of interaction with the person being assessed? And as online writing can often be extensive and thus time-consuming to assess, how can technology make healthcare professionals' assessment more efficient?

To investigate how healthcare professionals assess online writing, we collect a dataset of risk assessment for online postings using data from Reddit (reddit.com), an online site for anonymous discussion on a wide variety of topics. We focus specifically on users who have posted to a discussion forum called *SuicideWatch*, which, as its name

[2] koko.ai, au.reachout.com, 7cups.com, reddit.com/r/SuicideWatch, respectively.

suggests, is dense in postings by people who are considering taking their own lives.[3] We leave the discussion on how technology can make a healthcare professional' assessment more efficient, as well as a hierarchical ranking of the problem formulation, to Chapter 5.

4.2 The UMD Reddit Suicidality Dataset

We describe the creation of a dataset consisting of individuals who posted on SuicideWatch, that, by virtue of posting to the forum, were by definition considered potentially at risk. This is, however, very noisy since posting on SuicideWatch does not necessarily imply suicidal ideation.[4] A subset of individuals was thus assessed independently by four healthcare professionals who specialize in suicidality assessment. Together with their assessment of the individual's suicidality risk, these experts also provide annotations for which document of the individual most supports their judgments. In addition to experts, crowdsource workers assessed a larger set of individuals based on the same instructions.

In the following sections, we describe a rule-based collection strategy and the annotation instructions for experts and crowdsourcers.

4.2.1 Data Collection by Weak Supervision

Our approach to data collection, which we term WEAK SUPERVISION, is inspired by Coppersmith et al. (2014), who introduced an innovative way to solve for the absence of clinical ground truth when studying mental health in social media. Their approach is to use heuristic rules, or weak superivisions (Ratner et al., 2016), to identify individuals who have produced an overt signal in social media, indicating they *might* be a positive instance of the relevant condition, and then manually assessing the signal to filter out candidates for which the signal does not appear genuine. Coppersmith et al. (2014) applied this on Twitter by seeking variations of the statement *I have been diagnosed with X*, (where X is *depression*, *PTSD*, or other conditions), and then manually filtering out tweets for which the statement was in jest or otherwise not a true indication. For example, *The Red Sox lost*

[3] Titled forums on Reddit are called *subreddits*, but for clarity and generality we sometimes adopt the more common term *discussion forum*.

[4] For example, seeking help for a friend or offering support could not be evidence of suicidal ideation.

their third game in a row. I've just been diagnosed with depression. They also collected controls who had not made such statements.

The Coppersmith et al. approach does not yield clinical ground truth, since there is no way to verify an actual diagnosis, nor any way to determine that a control instance might not actually be positive for the condition. However, obtaining clinical data presents extremely challenging procedural burdens[5], and shared datasets for healthcare and mental health are thus typically orders of magnitude smaller than datasets supporting research in other domains (Spasic and Nenadic, 2020; Harrigian et al., 2021).

The WEAK SUPERVISION "signal" we use for an individual's candidate positive status with respect to suicidality is their having posted in the `/r/SuicideWatch` subreddit, a forum providing "peer support for anyone struggling with suicidal thoughts, or worried about someone who may be at risk".[6] We began with a snapshot of every publicly available Reddit posting from January 1, 2008 through August 31, 2015, with partial data from 2006-2007, comprising approximately 42G of compressed data.[7] Eliminating individuals who had fewer than ten total posts across all of Reddit, we had 11,360 individuals who had posted in SuicideWatch for a total of 1,556,194 posts. For these individuals we extracted not only their SuicideWatch posts, but *all* their Reddit posts available in the snapshot. Through random sampling, we further selected 1,097 individuals, of which 934 ultimately were included for further human annotation. These huamn annotated individuals are then excluded from the WEAK SUPERVISION dataset to prevent data leakage, leaving a final number of 10,263 potential at-risk individuals. We also aggregated the data from an equal number of control individuals who had not posted in any of the mental health subreddits identified by Pavalanathan and De Choudhury (2015), nor in the `/r/schizophrenia` subreddit.[8]

[5] Access to healthcare data in the U.S. is governed by the Healthcare Insurance Portability and Accountability Act, or HIPAA.

[6] https://www.reddit.com/r/SuicideWatch/

[7] The corpus: https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/. See Gaffney and Matias (2018) for caveats. Note that more recent data is available, see https://files.pushshift.io/reddit/

[8] Our full set: addiction, alcoholism, Anger, bipolarreddit, BPD (Bederline Personality Disorder), depression, DPDR (depersonalization, derealization), EatingDisorders, feelgood, getting_over_it , hardshipmates, mentalhealth, MMFB (MakeMeFeelBetter), panicparty, psychoticreddit, ptsd, rapecounseling, schizophrenia, socialanxiety, StopSelfHarm, SuicideWatch, survivorsofabuse, traumatoolbox.

4.2.2 Annotation Instructions

Having posted on SuicideWatch does not necessarily imply the individual has suicidal ideation. In this section, we describe the instructions for our annotators – experts and crowdsourcers. It is infeasible to ask annotators to read through all of an individual's postings, where the number can be in the thousands. For purposes of annotation, we limit postings to those on SuicideWatch for each of the 934 individuals, although at training and test time, we use all postings. This challenge of annotation mirrors the challenge of assessment. However, in a setting where a healthcare professional is assessing a potentially at-risk individual using social media postings, they may not have weak supervision signals to help limit the individual's postings to be assessed. In Chapter 5, we revisit this challenge by using document attention to surface postings that are more likely to contain signals of suicidality.

To facilitate crowdsourced as well as expert annotation, we divided sequences of more than five SuicideWatch posts for a single individual into multiple annotation units containing up to five posts each, yielding a total of 982 annotation units. For example, an individual with 12 SuicideWatch posts would yield three annotation units of their first 5 posts, next 5 posts, final 2 posts. In order to determine individual-level risk, we consider an individual to have the highest risk associated with any of their annotation units.

We defined a four-way categorization of risk, adapting Corbitt-Hall et al. (2016) (who provided lay definitions based on risk categories in Joiner Jr et al. (1999)): **(a) No Risk (or "None")**: I don't see evidence that this person is at risk for suicide; **(b) Low Risk**: There may be some factors here that could suggest risk, but I don't really think this person is at much of a risk of suicide; **(c) Moderate Risk**: I see indications that there could be a genuine risk of this person making a suicide attempt; **(d) Severe Risk**: I believe this person is at high risk of attempting suicide in the near future.[9]

We then defined two sets of annotator instructions. The *short* instructions, intended only for a subset of experts, simply presented the above categorization and asked them to follow their training in assessing patients with suicide risk. A *long* set of instructions

[9]These correspond roughly to the *green*, *amber*, *red*, and *crisis* categories defined by Milne et al. in CLPsych ReachOut shared tasks (Milne et al., 2016; Milne, 2017).

was similar in intent to those of Corbitt-Hall et al. (2016), but whereas their instructions focused on three risk factors (*thoughts of suicide*, *planning*, and *preparation*), we identified four families of risk factors: *thoughts* includes not only explicit ideation but also, for example, feeling they are a burden to others or having a "fuck it" (screw it, game over, farewell) thought pattern; *feelings* includes, for example, a lack of hope for things to get better, or a sense of agitation or impulsivity (mixed depressive state, Popovic et al., 2015); *logistics* includes, for example, talking about methods of attempting suicide (even if not planning), or having access to lethal means like firearms; and *context* includes, for example, previous attempts, a significant life change, or isolation from friends and family.[10]

In both sets of instructions, expert annotators were additionally asked to label the post that most strongly supports the judgment, and they were told that choices should never be downgraded: if an earlier post suggests a person is at severe risk ("I'm going to kill myself"), and a later post suggests the risk has decreased ("I've decided not to kill myself"), the higher risk should be chosen, and the severe-risk post should be identified as the basis for the judgment.

4.2.3 Expert Annotation

We selected 242 individuals at random to create a set of 245 annotation units that were labeled independently by four volunteer experts in assessment of suicide risk.[11] These included a suicide prevention coordinator for the Veteran's Administration; a member of the National Suicide Prevention Lifeline's Standards, Training and Practices Sub-Committee; a doctoral student with expert training in suicide assessment and treatment whose research is focused on suicidality among minority youth; and a clinician in the Department of Emergency Psychiatry at Boston Children's Hospital. Two of these experts received the detailed long instructions, and the other two were given the short instructions.

Table 4.1 shows Krippendorff's α (Krippendorff, 2004) pairwise among the experts,

[10] See Appendix A.2 for the long instruction.

[11] Random selection was from the set of crowdsource-annotated individuals obtained in Section 4.2.4, ensuring that all expert annotations would be accompanied by crowdsourced annotations. Recall that an individual's label is the highest-risk label assigned for any of that individual's annotation units, if there are more than one. The original EXPERT dataset had 245 individuals; we exclude three owing to errors in processing.

Krippendorff α	exp_L1	exp_L2	exp_S1	exp_S2
exp_L1	1	0.837	0.804	0.823
exp_L2	-	1	0.808	0.831
exp_S1	-	-	1	0.768
exp_S2	-	-	-	1

Table 4.1: Krippendorff's α pairwise among experts. exp_L1 represents the first expert who follows the long instruction.

indicating the set of instructions they used as (S)hort or (L)ong. The average of 0.812 satisfies the conventional reliability cutoff for chance-corrected agreement (> 0.8, Krippendorff (2004)), which is to our knowledge the first result demonstrating inter-rater reliability by experts for suicide risk based on social media postings. Inter-rater reliability for the pair receiving short instructions was substantially lower (0.768), demonstrating the value of our detailed rubric based on explicitly identified risk factors.

We generated consensus individual-level labels based on the expert annotations using Dawid-Skene (Dawid and Skene, 1979; Passonneau and Carpenter, 2014, implemented with Stan[12]), a well known model for inferring consensus labels from multiple noisy annotations. Using Dawid-Skene, we generate consensus for the pairs receiving long instructions (*Long Experts*), short instructions (*Short Experts*), and consensus among all four experts. This results in an EXPERT dataset consisting of 242 individuals.

4.2.4 Crowdsourced Annotation

For the CROWDSOURCE dataset, we created a task on CrowdFlower (formerly crowdflower.com, now appen.com) using the long instructions. We restricted participation to high performance annotators (as determined by the CrowdFlower platform) and who also agreed with our annotations on seven clear test examples. Although we began with 1,097 individuals to annotate, crowdsourcer participation tailed off at 934.[13] After discarding any annotation unit labeled by fewer than three annotators, our data consists of 865 individuals and 905 annotation units. We used CrowdFlower's built-in consensus label as the

[12]https://mc-stan.org/docs/2_27/stan-users-guide/data-coding-and-diagnostic-accuracy-models.html

[13]We conjecture that, with fewer jobs left available, annotators were less inclined to go through the detailed instructions and test because there was less for them to get paid for.

	# Posts	10-20	20-40	40-60	60-100	100-200	200-500	500-1,000	$\geq$ 1,000
WS	Control	4,674	3,023	1,140	965	620	257	57	13
	Positive	2,390	2,362	1,328	1,465	1,396	935	236	61
Crowdsource	No Risk	31	42	25	27	18	12	4	0
	Low Risk	19	22	5	11	2	4	0	0
	Moderate Risk	46	45	19	14	9	7	1	0
	Severe Risk	80	79	37	19	28	12	3	0
Expert	No Risk	3	7	2	5	7	8	3	0
	Low Risk	6	11	5	11	8	7	1	1
	Moderate Risk	23	19	12	26	13	14	5	3
	Severe Risk	7	2	5	9	10	4	4	1

Table 4.2: Number of individuals with the number (range) of posts (in all of Reddit, not just SuicideWatch), by dataset and risk category. WS stands for WEAK SUPERVISION.

crowdsourced label for each unit.[14] Krippendorff's α for inter-annotator agreement of the crowdsourcers for individual labels is 0.554.

To prevent data leakage, we further exclude individuals in the CROWDSOURCE dataset who have also been annotated by experts, reducing the 934 individuals to a final number of 621 individuals. However, these annotations are not wasted. These overlapping individuals between CROWDSOURCE and EXPERT are used to calculate annotation disagreement in Section 4.3.

4.2.5 Dataset Statistics

The final dataset contains three subsets with disjoint individuals. The first, WEAK SUPERVISION dataset, includes 10,263 potential at-risk individuals and 10,759 control individuals; they are respectively considered to be indirectly positively and negatively labeled. The second set is the CROWDSOURCE dataset, including 621 individuals annotated by crowdsourcers with four risk levels: *No Risk*, *Low Risk*, *Moderate Risk*, and *Severe Risk*. The last is the EXPERT dataset, including 242 individuals with the same four risk levels, by four suicide risk assessment experts. Along with the level of risk for each individual, the annotators for EXPERT dataset also designated the single post that most strongly supported each of their low, moderate, or severe risk labels. An at-risk individual's number of

[14] See *Confidence Score* https://success.appen.com/hc/en-us/articles/202703305-Getting-Started-Glossary-of-Terms

	Long Experts	Short Experts	Crowdsourcers
All Experts	0.8367	0.7173	0.5047

Table 4.3: Macro F_1 scores for consensus human predictions on the 242 individuals labeled by both experts and crowdsourcers, using all-experts consensus as ground truth.

posts can range from 10 to 1,326. See Table 4.2 for a detailed breakdown of the number of posts per individual across datasets and risk categories.

4.3 Annotation Disagreements

To investigate the quality of annotation across and within groups of crowdsourcers and experts, we begin by treating annotation as a prediction task performed by humans. Table 4.3 shows the macro F_1 score using all-experts consensus labels as ground truth, with different human consensus values as the prediction. These pattern as one would expect, decreasing from experts with long instructions, to experts with short instructions relying on (varied) training, and we hypothesize that the much lower performance of crowdsourcers arises both because they have less training than experts, and because they are less mission-driven in their motivations and therefore are likely to feel a lower commitment to the task.

Nonetheless, it is worth noting that there is clear value in the crowdsourced annotations. Table 4.4 shows a confusion matrix measuring crowdsourcers' consensus against the all-experts consensus, and it appears that most of the disagreements involve crowdsourcers erring on the side of caution, misclassifying more than half of the low-risk individuals as having higher risk, and misclassifying a large number of moderate risk individuals (no imminent threat of a suicide attempt) as having severe (imminent) risk. In settings where the goal is to flag individuals for more careful review and possible intervention, false positives seem likely to be the preferred kind of error.[15]

Table 4.5 shows the confusion matrix for experts receiving short versus long instructions, which may be illuminating for scenarios in which trained healthcare profes-

[15] Performance differences between experts and non-experts require more study. For example, Homan et al. (2014) found that two novice annotators were *more* likely to assign their expert's "low distress" tweets to the "no distress" category. Conversely, on a related but coarser-grained categorization task, Liu et al. (2017) find "some evidence that multiple crowdsourcing workers, when they reach high inter-annotator agreement, can provide reliable quality of annotations".

		Crowdsourcers			
		None	Low	Moderate	Severe
All Experts	None	29	1	1	5
	Low	11	13	20	6
	Moderate	6	11	47	51
	Severe	1	1	8	34

Table 4.4: Counts of agreement and disagreement cases between experts' consensus (All Experts) and crowdsourcers' consensus (Crowdsourcers).

		Short Experts			
		None	Low	Moderate	Severe
Long Experts	None	36	1	1	0
	Low	5	16	34	3
	Moderate	1	0	56	14
	Severe	0	0	17	61

Table 4.5: Counts of agreement and disagreement cases between experts using long instruction (Long Experts) and short instruction (Short Experts).

sionals perform assessment using social media posts but do not take the time to apply the long-instructions rubric or do not do so consistently. We observe the same trend toward erring in the direction of false positives, and it is notable that *no* severe-risk individuals (based on the long-instruction consensus) are assigned to no risk or even low risk by the short-instructions consensus.

4.4 Privacy and Anonymization

Our research involving the University of Maryland Reddit Suicidality Dataset has undergone review by the University of Maryland Institutional Review Board with a determination of Category 4 Exempt status under U.S. federal regulations. For this dataset, (a) the original data are publicly available, and (b) the originating site (Reddit) is intended for anonymous posting.

Individual accounts on Reddit are fundamentally anonymous: when creating a Reddit account, only a username and password need to be supplied, with e-mail address optional (Reddit, 2018). Since individuals might have chosen to include potentially iden-

tifying information in their usernames, we go a step further and replace usernames with unique numeric identifiers.[16]

In addition, the dataset used in this chapter has undergone automatic de-identification using named entity recognition to aggressively identify and mask out potential personally identifiable information, such as person names and organizations, in order to create an additional layer of protection (Zirikly et al., 2019). In an assessment of de-identification quality, Zirikly et al. (2019) manually reviewed a sample of 200 randomly selected posts (100 from the SuicideWatch subreddit and 100 from other subreddits), revealing zero instances of personally identifiable information.

Following Benton et al. (2017), we treat the data (even though de-identified) as sensitive and restrict access to it, we use obfuscated and minimal examples in the dissertation and presentations, and we do not engage in linkage with other datasets.

4.5 From Classification to Prioritization

This chapter has created a dataset for research on risk assessment for suicidality based on social media, which includes expert ratings for 242 individuals and crowdsourced ratings for 621 individuals. We found that inter-rater agreement among experts is very good, with consistency particularly encouraged using detailed instructions specifying classification criteria. We also looked at differences in consistency when ratings are provided by experts using their own experience and judgment rather than following detailed instructions.

Since the creation of this dataset in 2018, this dataset since has been shared, through a collaboration with the American Association of Suicidology (AAS), with more than 35 teams internationally.[17] The dataset has also been used in a shared task in the 2019 Computational Linguistics and Clinical Psychology Workshop (CLPsych 2019) held at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (Zirikly et al., 2019).

Most of the teams approach the dataset in a classification framework. That is, the

[16] For example, a hypothetical individual could choose the username `maryjanesmith1973.collegepark`, identifying name, birth year, and location.

[17] See dataset availability and governance plan: http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html

predictive system attempts to predict the category of suicidality risk by taking the sequence of postings as inputs. The predicted risk is then compared with the expert assessment to calculate micro or macro F_1 scores. While these approaches make sense, they do not directly address the scarcity of mental health assessment resources – healthcare professionals' time, in particular. For healthcare professionals to assess an individual consisting of a potentially large number of documents takes time; assessing many individuals take even more time. Recognizing this challenge, we introduce a reformulation of the problem from classification to prioritization in the next chapter. Instead of evaluating the categorical agreement, we evaluate the ranking of the individuals based on their suicidality risk. Furthermore, we also investigate document attention's ability to surface documents likely containing signals of suicidality to save healthcare professionals' time.

Chapter 5: A Prioritization Model for Suicidality Risk Assessment

Chapter 4 describes the creation of a dataset for assessing suicidality risk using social media, which follows a similar structure to the clinical coding problem in Chapter 3: the level of inference is at the individual level, but the evidence is found in a subset of the individual's documents. In this chapter, recognizing healthcare professionals's time constraint, we introduce a reformulation of the problem from classification to prioritization.

Recall that in Chapter 1, we argue that many applications in a healthcare setting should not be automated without healthcare professionals' intervention. The need to involve healthcare professionals introduces a resource limitation.[1] *Time*, is the limiting resource. In this chapter, we reframe suicide risk assessment from social media as a ranking problem whose goal is maximizing detection of severely at-risk individuals given the *time* available. Building on measures developed for resource-bounded document retrieval, we introduce a well-founded evaluation paradigm. Using the expert-annotated test collection in Chapter 4, we demonstrate that meaningful improvements over plausible cascade model baselines can be achieved using a document attention-based approach (similar to ELDAN in Chapter 3) that jointly ranks individuals and their social media posts.[2]

5.1 The Need to Prioritize

Mental illness is one of the most significant problems in healthcare: in economic terms alone, by 2030 mental illness worldwide is projected to cost more than cardiovascular disease, and more than cancer, chronic respiratory diseases, and diabetes combined (Bloom et al., 2012). Suicide takes a terrible toll: in 2016 it became the second leading cause of

[1] In this setting, healthcare professionals may refer to suicide prevention specialists, mental health clinicians, or psychiatrists.

[2] This chapter contains content from: **Shing, Han-Chin**, Philip Resnik, and Douglas W. Oard. "A prioritization model for suicidality risk assessment." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

death in the U.S. among those aged 10-34 and fourth among those aged 35-54. Overall rates are nearly twice as high for most rural areas (where access to help is more likely to be a challenge) than most urban areas (Hedegaard et al., 2018). Prevalence statistics suggest that roughly 409 of the 7,711 authors who submitted to the 58th annual meeting of the Association for Computational Linguistics (ACL) in 2020 have since had serious thoughts of suicide, 116 have made a plan, and 46 have actually made attempts.[3]

The good news is that NLP and machine learning are showing strong promise for impact in mental health. Traditional methods for predicting suicidal thoughts and behaviors have failed to make progress for fifty years (Franklin et al., 2017), but with the advent of machine learning approaches (Linthicum et al., 2019), including text analysis methods for psychology (Chung and Pennebaker, 2007) and the rise of research on mental health using social media (Choudhury, 2013), algorithmic classification has reached the point where it can now dramatically outstrip performance of prior, more traditional prediction methods (Linthicum et al., 2019; Coppersmith et al., 2018). Further progress is on the way, as the community shows increasing awareness and enthusiasm in this problem space (e.g., Milne et al., 2016; Losada et al., 2020; Zirikly et al., 2019; Goharian et al., 2021).

The bad news is that moving these methods from the lab into practice will create a major new challenge: identifying larger numbers of people who may require clinical assessment and intervention will increase stress on a severely resource-limited mental health ecosystem that cannot easily scale up.[4] This motivates a reformulation of the technological problem from classification to *prioritization* of individuals who might be at risk, for clinicians or other suitably trained staff as downstream users.

Perhaps the most basic way to do prioritization is with a single priority queue that the user scans from top to bottom. This "ranked retrieval" paradigm is common for Information Retrieval (IR) tasks such as document retrieval. The same approach has been applied to ranking people based on their expertise (Balog et al., 2012), or more generally to ranking entities based on their characteristics (Balog, 2018). Rather than evaluating

[3] Approximately: ACL is international, but these figures use prevalence statistics for U.S. adults in 2019 (Elinore F. McCance-Katz, SAMHSA, 2020): 5.3% had serious thoughts, 1.5% made a plan, 0.6% made attempts.

[4] 122 million Americans live in areas with mental healthcare provider shortages (Bureau of Health Workforce, 2020). That number reflects an increase of about 9 million people between September 30, 2019 and December 31, 2020.

categorical accuracy, ranked retrieval systems are typically evaluated by some measure of search quality that rewards placing desired items closer to the top (Voorhees, 2001). Most such measures use only item position, but we find it important to also model the *time* it takes to recognize desired items, since in our setting the time of qualified users is the most limited resource.

In this chapter, we do so by building on Time-Biased Gain (TBG, Smucker and Clarke, 2012), an IR evaluation measure that models the expected number of relevant items a user can find in a ranked list given a time budget. We observe that in many risk assessment settings (e.g., Yates et al., 2017; Coppersmith et al., 2018; Zirikly et al., 2019), the available information consists of a (possibly large and/or longitudinal) set of documents, e.g., social media posts, associated with each individual, of which possibly only a small number contain a relevant signal.[5] This hierarchical structure, which is similar to that of Chapter 3, combined with prioritization, gives rise to a formulation of our scenario as a nested, or *hierarchical*, ranking problem. In this hierarchical ranking, individuals are ordered by priority, but each individual's documents must also be ranked. Accordingly, we introduce hierarchical Time-Biased Gain (hTBG), a variant of TBG in which individuals are the top level ranked items, and expected reading time is modeled for the ranked list of documents that provides evidence for each individual's assessment.

In addition, we introduce a prioritization model that jointly optimizes the nested ranking task using a three-level hierarchical attention network (Yang et al., 2016); this model also addresses the fact that in our scenario, as in many other healthcare-related scenarios, relevance obtains at the level of individuals rather than individual documents (see Chapter 3). Using a test collection of Reddit-posting individuals who have been assessed for suicide risk by healthcare professionals based on their posts (Chapter 4), we demonstrate, using hTBG, that our joint model substantially outperforms cascade model baselines in which the nested rankings are produced independently.

[5] Our dataset, for example, has one severe risk individual with 1,326 postings, of which only two are "signal" posts identified by experts. See Table 4.2 for detailed statistics.

5.2 Prediction Model

We began by motivating risk assessment via social media as a person-centered, time-limited prioritization problem, in which the technological goal is to support downstream healthcare professionals or other assessors in identifying as many people at risk as possible. This led to the conclusion that systems should not only rank individuals but, for each individual, rank their posts.

Next, we need a system that can produce such nested rankings of individuals and their posts. Ideally such a system should be able to train on only individual-level, not document-level, labels, since suicide risk is a property of individuals, not documents, and document labels are more difficult to obtain. In addition, such a system should ideally produce additional information to help the downstream healthcare professional — if not justification of its output, then at least highlighting potentially useful information.

The ELDAN model in Chapter 3 fits this need. It handles a similar hierarchical structure and supports training on the individual level (encounter level in Chapter 3) without document-level annotations. The document-attention of ELDAN has also been shown to match professional medical coders' expectations. We thus introduce 3HAN, a hierarchical attention network (Yang et al., 2016) that extends up to the level of individuals, who are represented as sequences of documents. This architecture is similar to the ELDAN we proposed in Chapter 3 for coding clinical encounters; it obtained good predictive performance and we also showed that, despite concerns about the interpretation of network attention (Jain and Wallace, 2019), hierarchical document-level attention succeeded in identifying documents containing relevant evidence. 3HAN differs from ELDAN in that it builds representations hierarchically from the word level, as opposed to pre-extracted conceptual features such as those used in ELDAN, and it takes document ordering into account using a bi-directional GRU (Bahdanau et al., 2015).

Specifically, our model has five layers (Figure 5.1). The first is a word-embedding layer that turns a one-hot word vector into a dense vector. The second to fourth layers are three Seq2Vec layers with attention that learn to aggregate, respectively, a sequence of word vectors into a sentence vector, a sequence of sentence vectors into a document

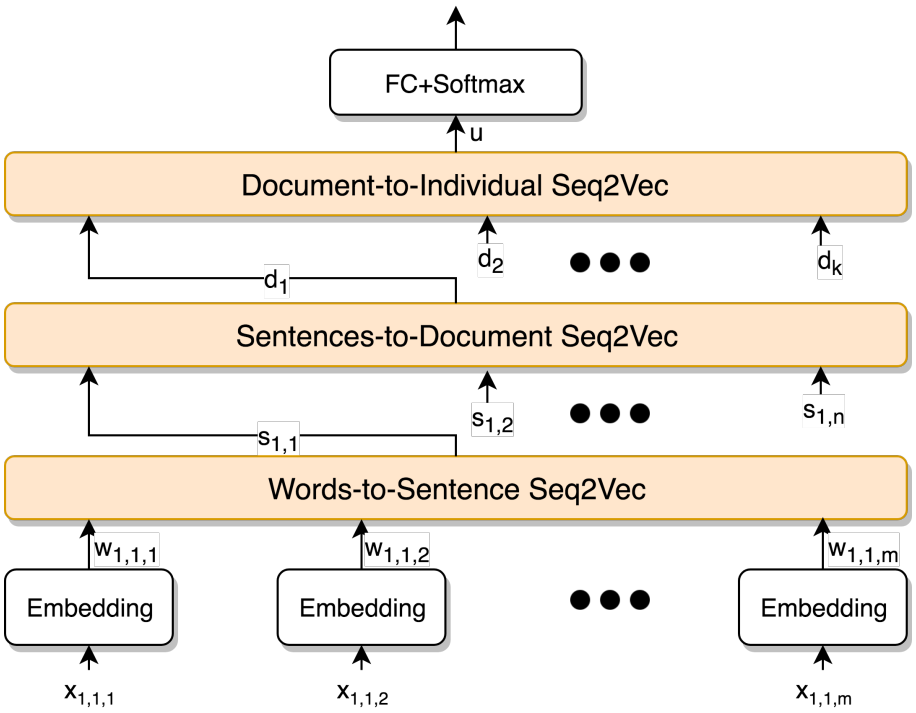

Figure 5.1: An illustration of the three-level Hierarchical Attention Network (3HAN) model

vector, and a sequence of document vectors into an individual vector (hence 3HAN). The final layer is a fully connected layer followed by softmax.

We detail our Seq2Vec layer in the context of aggregating a sequence of document vectors to an individual's vector, though the three Seq2Vec layers are the same. See Figure 5.2 for an illustration. Document vectors $\{d_{i,j}\}_{j=1}^{m}$ are first passed through a bi-directional GRU layer. The outputs, after passing through a fully-connected layer and a non-linear layer, are then compared to a learnable attention vector, $v_{\text{attention}}$. Specifically,

$$g_{i,j} = \text{Bi-GRU}(d_{i,j}) \tag{5.1}$$

$$r_{i,j} = \tanh\left(W g_{i,j} + b\right) \tag{5.2}$$

$$a_{i,j} = \frac{e^{r_{i,j}^{\top} v_{\text{attention}}}}{\sum_{j'=1}^{m} e^{r_{i,j'}^{\top} v_{\text{attention}}}} \tag{5.3}$$

$$u_i = \sum_{j=1}^{m} a_{i,j} g_{i,j} \tag{5.4}$$

where $a_{i,j}$ is the normalized document attention score for the j-th vector, and u_i is the

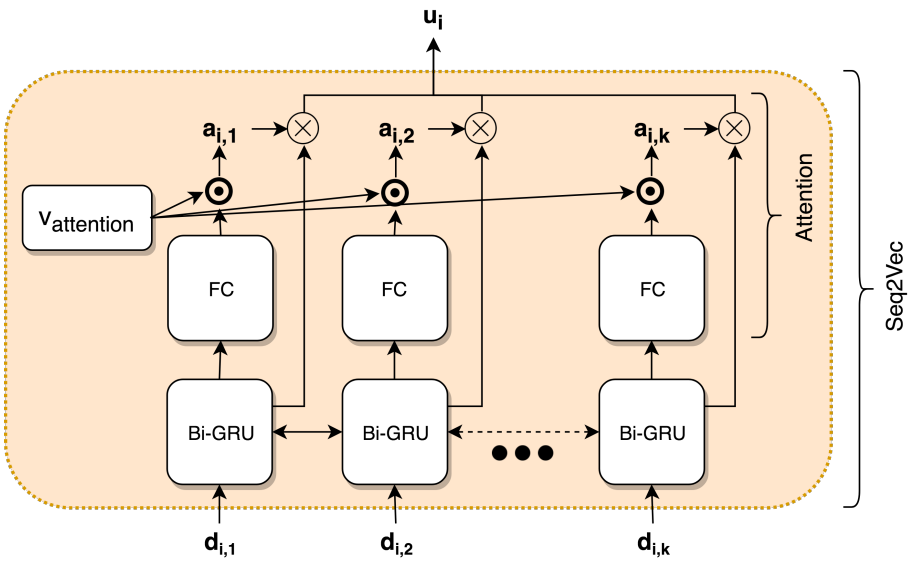

Figure 5.2: Seq2Vec with Attention used in the 3HAN model

final aggregated individual vector. As shown in Equation 5.3, the transformed vector $r_{i,j}$ is compared with the learnable attention vector $v_{\text{attention}}$ using a dot product, and further normalized for the weighted averaging step in Equation 5.4. The Seq2Vec layer that uses document attention to aggregate a sequence of documents into an individual vector is directly parallel to the document-level attention layer of ELDAN from Chapter 3. The main difference between the two is that 3HAN's document attention additionally models sequential information by using a bi-directional GRU.

Once we have the individual vector u_i, we can predict the risk label of the individual by passing it through a fully-connected layer and a softmax. Specifically,

$$P(\hat{y}_i) = softmax\left(W_{FC}u_i + b_{FC}\right) \tag{5.5}$$

Finally, we compare with the ground truth label y_i of individual i using negative log-likelihood to calculate a loss:

$$loss_i = -log\left(P\left(\hat{y}_i = y_i\right)\right). \tag{5.6}$$

5.3 A Measure for Risk Prioritization

Section 5.1 argues for formulating risk assessment as a prioritization process where the assessor has a limited time budget. This motivates the need for a hierarchical ranking that jointly ranks individuals and their documents – thus the 3HAN model in Section 5.2. The need to evaluate that hierarchical ranking under a limited time constraint leads to four desired properties in an evaluation measure:[6]

- **Risk-based**: Individuals with high risk should be ranked above others.
- **Head-weighted**: Ranking quality near the top of the list, where assessors are more likely to look, should matter more than near the bottom.
- **Speed-biased**: For equally at-risk individuals, the measure should reward ranking the one who can be assessed more quickly closer to the top, so that more people at risk can be identified within a given time budget.
- **Interpretable**: The evaluation score assigned to a system should be meaningful to assessors.

Among many rank-based measures that satisfy the *risk-based* and *head-weighted* criteria, TBG directly accounts for assessment time in a way that also satisfies the *speed-biased* criterion (see Theorem 5.3.1). Furthermore, the numeric value of TBG is a lower bound on the expected number of relevant items — in our case, high-risk individuals — found in a given time budget (Smucker and Clarke, 2012), making it *interpretable*. After introducing TBG, in Section 5.3.2 we develop *hierarchical* Time-Biased Gain (hTBG), an extension of TBG, to account for specific properties of risk assessment using social media posts.[7]

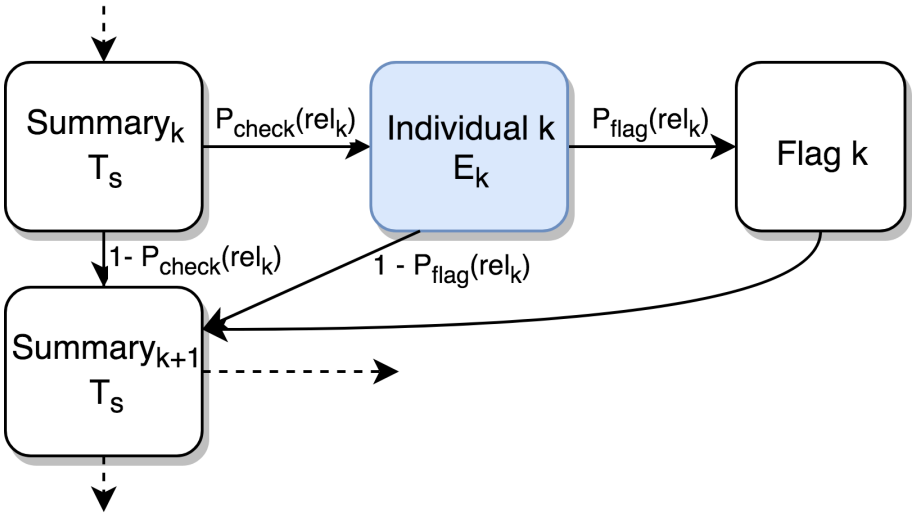

Figure 5.3: User model for Time-Biased Gain (TBG)

5.3.1 Time-Biased Gain

TBG was originally developed in IR for the case of a user seeking to find a relevant document, but here we frame it in the context of risk assessment (Figure 5.3). TBG assumes a determined user (say a healthcare professional) examining a ranked list of individuals in the order presented by the system. For each individual, the healthcare professional first examines a *summary* and then decides whether to check relevance via more detailed examination, or to move on. Checking requires more time to make an assessment of whether the individual is indeed at-risk. TBG is a weighted sum of gain, g_k, and discount, $D(\cdot)$, a function of time:

$$\text{TBG} = \sum_{k=1}^{\infty} g_k D\left(T\left(k\right)\right). \tag{5.7}$$

[6]Throughout, *assessor* or *user* signify a healthcare professional or other human assessor, and *individual* is someone being assessed.

[7]TBG and hTBG code: https://github.com/sidenver/hTBG

$T(k)$ is the expected amount of time it takes a user to reach position k:

$$T(k) = \sum_{i=1}^{k-1} t\,(i) \tag{5.8}$$

$$t(i) = T_s + P_{\text{check}}\,(\text{rel}_i)\,E_i \tag{5.9}$$

where $t(i)$ is expected time spent at position i. Breaking down $t(i)$, T_s is the time it takes to read a summary and decide whether to check the individual; if yes (with probability $P_{\text{check}}(\text{rel}_i)$), E_i is expected time for detailed assessment, calculated as a function of the individual's total word count W_i:

$$E_i = T_\alpha W_i + T_\beta \tag{5.10}$$

where T_α and T_β scale words to time. The discount function $D(t)$ decays exponentially with half-life h:

$$D(t) = 2^{-\frac{t}{h}} \tag{5.11}$$

where h is the time at which half of the healthcare professionals will stop, on average. The expected stop time (or mean-life) is $\frac{h}{ln(2)}$. Finally, the gain, g_k is:

$$g_k = P_{\text{check}}(\text{rel}_k) P_{\text{flag}}(\text{rel}_k) \mathbb{1}_{[rel_k=1]} \tag{5.12}$$

where $P_{\text{check}}(\text{rel}_k)$ is the probability of checking the individual after reading the summary at position k, and $P_{\text{flag}}(\text{rel}_k)$ is the probability of flagging that individual as high risk. Gain thus accrues only if a healthcare professional actually finds a high-risk individual, making TBG (and thus the following hTBG) a measure for binary relevance judgment.

The decay function in Equation 5.11 monotonically decreases with increasing time (and thus rank), so TBG satisfies the *head-weighted* criterion. Table 5.1 shows the parameters used in Smucker and Clarke (2012), which were estimated from user studies using data from TREC 2005 Robust track.

Particularly of interest in a time-limited assessment, we can prove that TBG (and thus hTBG) is *speed-biased*:

Parameter	Description	Value
$P_{\text{check}}(\text{rel}_i)$	Prob. to check, given the relevance of summary	0.64, if $\text{rel}_i = 1$ 0.39, if $\text{rel}_i = 0$
$P_{\text{flag}}(\text{rel}_i)$	Prob. to flag, given the relevance of individual	0.77, if $\text{rel}_i = 1$ 0.27, if $\text{rel}_i = 0$
T_s	Seconds to evaluate a summary	4.4
$T_\alpha W + T_\beta$	Seconds to judge W words	$0.018W + 7.8$

Table 5.1: Parameters used for TBG and hierarchical TBG.

Theorem 5.3.1 (TGB satisfies the speed-biased criterion)**.** *Swapping an at-risk individual of longer assessment time ranked at k with an equally at-risk individual of shorter assessment time ranked at $k + r$, where $r > 0$, always increases TBG.*

Proof. See Appendix B.2.1 □

5.3.2 Hierarchical Time-Biased Gain

TBG assumes that detailed assessment involves looking at *all* available evidence (Equation 5.10). However, in our setting, an individual may have a large or even overwhelming number of social media posts. One severe risk individual in the UMD Reddit Suicidality dataset (Chapter 4), for example, has 1,326 posts in Reddit, the vast majority of which would provide the assessor with no useful information. Therefore, we need to prioritize the documents to be read and a way of estimating when the user will have read enough to make a decision.

In general, healthcare professionals engage in a sensemaking process as they examine evidence, and modeling the full complexity of that process would be difficult. We therefore make two simplifying assumptions: (1) that there is a high-signal document that suffices, once read, to support a positive relevance judgment, and (2) that the healthcare professional will not read more than some maximum number of documents. These assumptions align well with those of Expected Reciprocal Rank (ERR), whose cascading user model assumes that as the user works down a ranked list (in our case, the ranked documents posted by a single individual), and that they are more likely to stop after viewing a

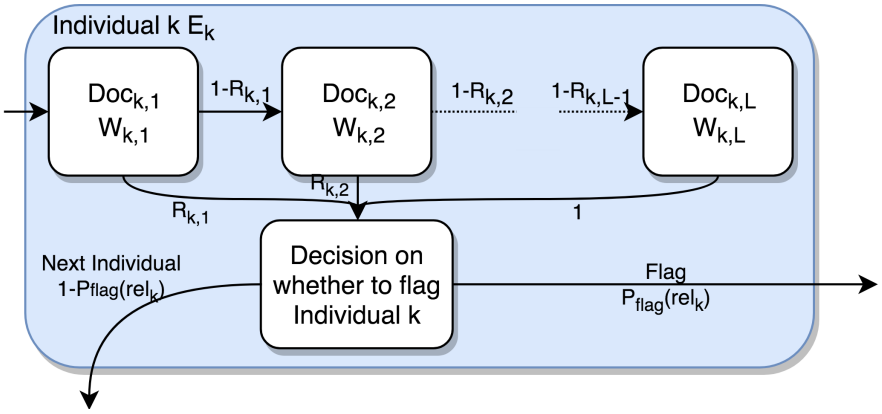

Figure 5.4: hTBG's model for calculating expected assessment time for an individual, replacing shaded box in Figure 5.3.

highly relevant document than after viewing an irrelevant one, as their information need is more likely to have been satisfied Chapelle et al. (2009). This results in a cascade model of user behavior: ERR $= \sum_{k=1}^{\infty} \frac{1}{k} P\,(\text{stop at k})$, in which $P\,(\text{stop at k}) = R_k \prod_{i=1}^{k-1} (1 - R_i)$, where $R_k = f(rel_k)$ is the probability of stopping at position k as a function of relevance.

This suggests replacing Equation 5.10 with the following expected time estimate for detailed assessment of an individual:

$$E_i = T_\alpha \sum_{l=1}^{L} \left(W_{i,l} \prod_{m=1}^{l-1} (1 - R_{i,m}) \right) + T_\beta \tag{5.13}$$

where $R_{i,l}$ is the probability of stopping at the l-th document for individual i, and $W_{i,l} > 0$ is the cost (in our case, word count) of reading the l-th document for individual i. Note that for the special case that no relevant document exists, $\forall i, l \in N, R_{i,l} = 0$), hTBG reduces to TBG. See Figure 5.4 for an illustration of E_i for hTBG. For the derivation of Equation 5.13 from ERR's cascading user model, see Appendix B.2.3.

5.3.3 Optimal Values for TBG and hTBG

Calculation of the optimal value for a measure is often important for normalization, though not always easy; in some cases it can be NP-hard (Agrawal et al., 2009). Another popular approach is to normalize by calculating the metric with an ideal collection. For example, Smucker and Clarke (2012) calculate the normalization factor of TBG by

assuming a collection with an infinite number of relevant documents, each of which lack any content. In our case, however, we are actually interested in an optimal value achievable for a given test collection: the optimal values of TBG and hTBG are properties of the bottleneck that occurs due to the user's limited time-budget. We find that:

Theorem 5.3.2 (Optimal TBG)**.** *The optimal value of TBG under binary relevance is obtained if and only if (1) all at-risk individuals are ranked above not-at-risk individuals, and (2) within the at-risk individuals, they are sorted based on time spent in ascending order.*

Proof. See Appendix B.2.1 □

Theorem 5.3.2 makes sense, as any time spent on assessing a not-at-risk individual is time not spent on assessing other potentially at-risk individuals. Preference in assessing individuals with shorter assessment time also increased the chance of assessing more individuals in the given time budget.

Minimum Individual Assessment Time. To calculate optimal hTBG, we need to minimize individual assessment time. A natural question to ask, then, is whether a result similar to Theorem 5.3.2 holds for the individual assessment time of hTBG in Equation 5.13. By swapping paired documents, we can use proof by contradiction to show that:

Theorem 5.3.3. *Minimum individual assessment time is obtained if the documents are sorted in descending order by* $\frac{R_{i,l}}{W_{i,l}}$.

Proof. See Appendix B.2.2 □

Theorem 5.3.3 shows a surprisingly intuitive trade-off between how relevant a document might be, and how much time (proportional to word counts) the expert needs to take to read it: highly relevant documents with short reading time are preferred.

Observe that Theorem 5.3.1 (speed-biased criterion) and Theorem 5.3.2 both apply to hTBG, as the two theorems only concern the ranking of individuals, not documents, and hTBG is an extension of TBG to measure the effects of document ranking. Using

Theorem 5.3.3 and Theorem 5.3.2, calculation of optimal TBG and hTBG values is simply a matter of sorting. For TBG, time complexity is $O(n\log(n))$, where $n \leq K$ is the number of at-risk individuals in the test collection. For hTBG, worst-case time complexity is $O(n\log(n) + nm\log(m))$, where $m \leq L$ is the maximum number of relevant documents per individual.

5.4 Experimentation

As we have shown in Section 5.3.2, hTBG provides a natural way to measure 3HAN: 3HAN's prediction can rank individuals, and within each individual we can use 3HAN's document attention to rank their social media posting. However, hTBG is not limited to just measuring 3HAN. Given an individual-level ranker and a document-level ranker, we can calculate hTBG scores by comparing it with a test collection that has both individual-level and document-level annotations.

In this section, we reintroduce the UMD Reddit Suicidality dataset from Chapter 4 as our test collection. We then show how we can evaluate 3HAN and different combinations of individual-level rankers and document-level rankers on the test collection using hTBG. Training details for all models can be found in Appendix B.3.

5.4.1 Test Collection

In our experimentation, we use the University of Maryland Reddit Suicidality Dataset described in Chapter 4 as our test collection.[8] The test collection consists of individuals represented by a (potentially long) sequence of social media postings, where the risk of suicidality is assigned to the individual. However, the annotation is done using their postings. For a detailed description, please refer to Chapter 4. Here we give a summary of the test collection.

Recall from Chapter 4 that the full dataset has three subsets with disjoint individuals, annotated with increasing level of annotation quality. The first, which we term the WEAK SUPERVISION dataset, annotated using heuristic rules, includes 10,263 po-

[8] See Appendix B.1 for IRB and ethical considerations.

tential positive individuals and 10,759 potential control individuals. The second set is the CROWDSOURCE dataset, including 621 individuals annotated by crowdsourcers with four risk levels: *No Risk*, *Low Risk*, *Moderate Risk*, and *Severe Risk*. The last is the EXPERT dataset, including 242 individuals with the same four-level annotation, by four suicide risk assessment experts. In addition to the level of risk for each individual, the expert annotators also designated the single post that most strongly supported each of their low, moderate, or severe risk labels.

5.4.2 Evaluating with TBG and hTBG

As TBG and hTBG are measures designed for binary relevance judgements, we map the *Severe Risk* category to *at-risk*, and everything else to *not-at-risk*.[9] For word counts, we directly use the token counts in documents. We use the parameters that Smucker and Clarke (2012) estimated for TBG in user studies (Table 5.1). As discussed in Section 5.3.2, we assume there exists a maximum number of documents the healthcare professional can read for each individual. We set that number to 50 for the calculation of hTBG; if no relevant document exists in the top 50 documents, we consider that individual a miss and set the gain to zero.[10]

To rank individuals using our classification models, we use a standard conversion method to convert four-class probability to a single score:

$$\sum_{\text{rel}_i}^{R} P\left(\hat{y}_i = \text{rel}_i\right) \text{score}_{\text{rel}_i} \tag{5.14}$$

where R is $\{\text{No}, \text{Low}, \text{Moderate}, \text{Severe}\}$, and $\text{score}_{\text{rel}_i}$ is the real number that maps to the risk-level of the individual i. We use $\{\text{No} = 0, \text{Low} = 1, \text{Moderate} = 2, \text{Severe} = 4\}$ as our mapping — *No Risk* can plausibly be treated the same as a post with no annotation (e.g. a control individual), and exponential scaling reflects our emphasis on finding high risk individuals.

The hTBG metric also requires a stopping probability for each document, $R_{i,l}$.

[9] Since the label definitions distinguish severe from moderate by focusing on the risk of an attempt *in the near future*, this binary distinction is aligned with recent work in suicidology that focuses specifically on characterizing "the acute mental state that is associated with near-term suicidal behavior" (Schuck et al., 2019).

[10] All parameters were frozen prior to testing.

Assuming that the more severe the risk associated with a document is, the more likely the assessor is to stop and flag the individual, on the EXPERT dataset where we have document-level annotations, we can estimate the expected stopping probability as:

$$R_{i,l} = 1 - \prod_{c=1}^{C} \left(1 - \frac{\text{score}_{\text{rel}_{i,l,c}}}{\text{score}_{\text{max}}}\right) \tag{5.15}$$

where C annotators annotated the post as most strongly supporting their judgment. $\text{Score}_{\text{rel}_{i,l,c}}$ is a mapping from the document-level risk by annotator c to a real number, with the same mapping used in Equation 5.14. $\text{Score}_{\text{max}} = 4$ is the maximum in that mapping.

To reflect different time budgets, we report results with the half-life parameter ranging from 1 to 6 hours, which correspond to expected time budgets from 1.4 to 8.7 hours.[11]

5.4.3 Models for Ranking Individuals

hTBG allows us to measure the hierarchical ranking (individual and their documents) produced by 3HAN. To compare 3HAN with other baselines using hTBG, we first describe three individual-level rankers that rank individuals based on their suicidality risk. In the next section, we will describe the document-level rankers. All models are pretrained on the WEAK SUPERVISION dataset, fine-tuned on the CROWDSOURCE dataset, and test on the EXPERT dataset.

3HAN. 3HAN is first pretrained on the binary WEAK SUPERVISION dataset. The model is then further tuned on the four-class CROWDSOURCE dataset by transferring the weights (except for the last fully-connected prediction layer) over. We initialized and fixed the word embedding using the 200-dimensional Glove embedding pretrained on Twitter (Pennington et al., 2014).[12]

3HAN_Av. 3HAN Average is trained the same way as 3HAN, except that the last Seq2Vec layer (the layer that aggregates a sequence of document vectors to an individual

[11] The expected stop time (or mean-life) is $\frac{h}{ln(2)}$

[12] We experimented with trainable Glove embedding as well as BERT, but saw little to no improvement in performance using cross-validation.

vector) is averaged instead of using attention, which can be achieved by fixing $a_{i,j} = \frac{1}{m}$ in Equation 5.3. This is similar to the HN-AVE baseline in Yang et al. (2016). Note that 3HAN_AV cannot rank documents, as it lacks document attention.

LR. A logistic regression model is trained on the CROWDSOURCE dataset. The feature vector for an individual is computed by converting documents into document-level feature vectors, and then averaging them to obtain an individual-level feature vector. For each document, we concatenate four feature sets: (1) bag-of-words for vocabulary counts larger than three, (2) Glove embedding summing over words, (3) 194 features representing emotional topics from Empath (Fast et al., 2016), and (4) seven scores measuring document readability.[13] This model is included as a conventional baseline from suicide risk assessment, where the features used are similar to some systems found in the NAACL CLPsych 2019 shared task (Zirikly et al., 2019).

5.4.4 Models for Ranking Documents

Recall from the previous section, comparing 3HAN with baseline using hTBG requires individual-level rankers and document-level rankers. However, ranking documents in a setting where document-level annotations are missing, as is our case here, is challenging. Here we describe three document-level rankers that do not need document-level annotations.

3HAN_Att. Document attention learned jointly with 3HAN. As a side effect to training our 3HAN model, we learn document attention scores, see Equation 5.3. This score can then be used to rank documents in terms of their relevance to the judgement. This availability of document ranking, despite a lack of document-level annotations, is a significant advantage of hierarchical attention networks, since fine-grained document-level annotations are difficult to obtain on a large scale. Sentence- and word-level attention are a further advantage, in terms of potentially facilitating user review (see Figure 5.5).

[13] Flesch-Kincaid Grade Level, Flesch Reading Ease, Dale Chall Readability, Automated Readability Index (ARI), Coleman Liau Index, Gunning Fog Index, and Linsear Write.

Individual Ranker	Document Ranker	Half-life h 1 hr	3 hrs	6 hrs
LR	FORWARD	7.51	10.05	10.89
3HAN_AV	FORWARD	7.76	10.15	10.94
3HAN	FORWARD	7.40	9.98	10.84
LR	BACKWARD	8.75	11.70	12.68
3HAN_AV	BACKWARD	9.65	12.09	12.89
3HAN	BACKWARD	9.73	12.17	12.95
LR	3HAN_ATT	9.44	12.05	12.88
3HAN_AV	3HAN_ATT	10.16	12.35	13.04
3HAN	3HAN_ATT	**10.39**	**12.49**	**13.12**
Optimal hTBG		19.78	20.39	20.54

Table 5.2: hTBG scores with three different time budgets, all combinations of individual and document rankers.

Forward and Backward. Ranking an individual's documents in either chronological order (FORWARD) or reverse chronological order (BACKWARD) is an obvious default in the absence of a trained model for document ranking, important baselines for testing whether a document ranking model actually adds value.

5.5 Results and Discussion

Our model, 3HAN+3HAN_ATT (the only joint model) achieves the best performance on hTBG compared to all other combinations of individual rankers and document rankers across three different time budgets (Table 5.2). The diffference in hTBG is statistically significant except when compared to 3HAN_AV+3HAN_ATT.[14] However, using 3HAN_ATT to rank documents implies that you have already trained 3HAN. Therefore, a more reasonable combination to compare with is 3HAN_AV+BACKWARD, which we outperform by a significant margin.

Overall, the effect of document ranking is larger than the effect of individual ranking. Notably, the FORWARD document ranker always yields the worst performance. BACKWARD, on the other hand, is surprisingly competitive. We hypothesize that this may be an indication that suicidal ideation worsens over time, or perhaps of the unfortu-

[14] Paired bootstrap resampling test, repeated 1000 times, $p < 0.05$.

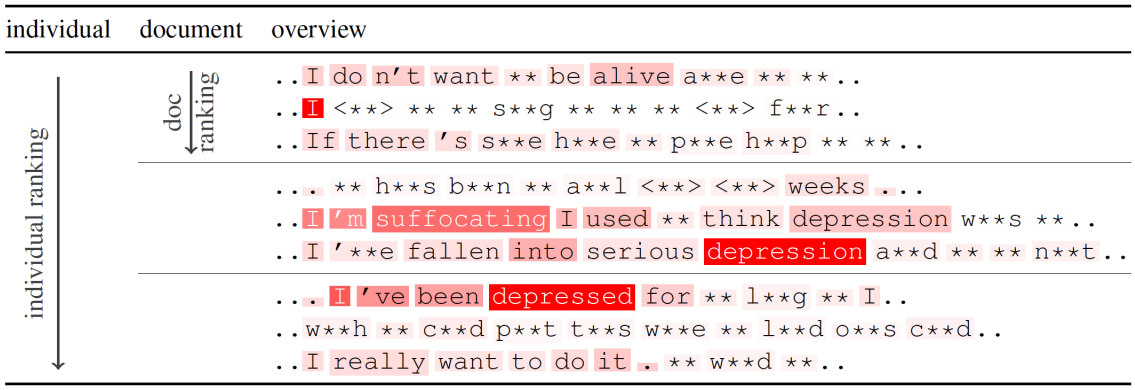

Figure 5.5: Illustration of an assessment framework in which individuals are ranked by predicted suicide risk based on social media posts, posts are ranked by expected usefulness for downstream review by a healthcare professional, and word-attention highlighting helps foreground important information for risk assessment. Real Reddit posts, obfuscated and altered for privacy. Note that we are only showing the top-three documents from the three highest-risk individuals, but in reality there can be thousands of documents and thousands of at-risk individuals.

nate event of suicide attempts following posting a *Severe Risk* document. This motivates the importance of prioritizing the reading order of documents: being able to find evidence early in suicide assessment leaves more time for other individuals, and will reduce the probability of misses.

Document ranking alone does not decide everything, as 3HAN+BACKWARD outperforms LR+3HAN_ATT. It is the combination of 3HAN and its document attention that produces our best model. This makes sense, as 3HAN, while learning to predict the level of risk, also learns which documents are important to the prediction.

Figure 5.5 shows the top 3 documents in a summary-style view for each of the highest ranked 3 individuals, with word-level attention shown using shading. Words with lower attention scores are obfuscated; others are altered to preserve privacy. The top-ranked individuals are annotated as *Severe Risk*, *Moderate Risk*, and *Moderate Risk*, respectively. For the two individuals annotated as *Moderate Risk*, one out of the four experts annotated them as *Severe Risk*. We suspect this is partially due to training on the CROWDSOURCE dataset, where crowdsourcers tend to err on the side of cautions and assign a higher risk category (see Section 4.3). In fact, crowdsourcers annotated the second highest rank individual as *Severe Risk*. Also, recall that we do not fine-tune on the

Ranker	hTBG	TBG	NDCG@20
3HAN+3HAN ATT.	**12.49**	11.46	70.90
3HAN AV.+BACKWARD	12.09	11.40	68.28
LR+BACKWARD	11.70	10.98	69.44
Optimal	20.39	19.75	100.00

Table 5.3: TBG and NDCG@20 listed to compare with hTBG. Both hTBG's and TBG's half lives are set at 3 hrs, and maximum document cutoff is set at 50.

EXPERT dataset, since in a realistic scenario, expensive annotations from experts usually are not available for training.

Most top-3 ranked posts mention plans or past experiences of suicide. Some of the highest-ranked postings from these individuals are not posted on SuicideWatch, but on other mental health-related forums. For example, an individual describes having long-term depression and isolation and a single pregnant mother with a history of depression and anxiety.

Previously Existing Measures. For previously existing measures (e.g., TBG, NDCG, Järvelin and Kekäläinen, 2002), document ranking has no effect, and thus these are not suitable measures in our scenario. However, we include results here for reference (Table 5.3). Since 3HAN_AV and LR cannot rank documents, it is impossible to calculate hTBG, so we report results on the chronologically backward ranking strategy. NDCG@20 is NDCG score cut off at 20 (see related work in Section 2.5), chosen based on the optimal hTBG value.

5.6 Summary

In this chapter, we extend our findings from Chapter 3 and demonstrate the effectiveness of document-level attention in the context of suicidality assessment. Using the UMD Reddit Suicidality dataset we collected in Chapter 4 and hTBG, we show that document-level attention leads to improved individual-level ranking and reduces the assessment time by surfacing documents that are likely to contain suicidal signals.

As mentioned in Chapter 1, Chapters 3 and 5 give rise to two important charac-

teristics: (1) modeling the individual (or an individual's clinical encounter) as a set of documents, and (2) surfacing relevant information from that set of documents. In Chapter 6, we return to the clinical encounter of Chapter 3. However, instead of assisting clinical coders with clinical coding, our aim is now to assist healthcare professionals with writing discharge summaries from prior clinical documents. We explore these two characteristics in this new task by surfacing content relevant to the discharge summary from the (potentially large) set of documents in the clinical encounter.

Chapter 6: Learning to Compose Discharge Summaries from Prior Clinical Notes

The records of patients can be extensive and complex, thus placing a premium on tools that can help healthcare professionals efficiently identify key facts about a patients. Clinical coding and suicidality risk assessments assisted by computer discussed in Chapters 3 and 5 are examples of such tools. They infer discrete labels about the patients and provide evidence for those inferred labels. Two shared characteristics emerge: (1) the patient is modeled as a set of documents, and (2) evidence supporting the model inference is surfaced from the extensive documents about the patient. These source of information include naturally occurring language written by the patient, like social media postings, and electronic health records (EHR) written by clinical practitioners.

In this chapter, we focus on the problem of computer-assisted discharge summary composition from a clinical encounter. This problem shares a similar characteristic to the previous chapters. We can model a patient's clinical encounter as a set of documents; we can then model the writing of the discharge summary, which is a summary of the clinical encounter typically written at the time of discharge, as a task of surfacing and composing relevant information from the encounter. In contrast to the previous chapters, we focus on producing natural language text (the discharge summary) instead of discrete labels. Conceptually, this leads us to model the problem as a multidocument summarization task.

Summaries in this setting need to be *faithful*, *traceable*, and *scalable* to multiple long documents, motivating the use of extract-then-abstract summarization cascades. We introduce two new measures, faithfulness and hallucination rate, for evaluation in this task, which complement existing measures for fluency and informativeness. Results across seven commonly found sections of the discharge summary and five models show

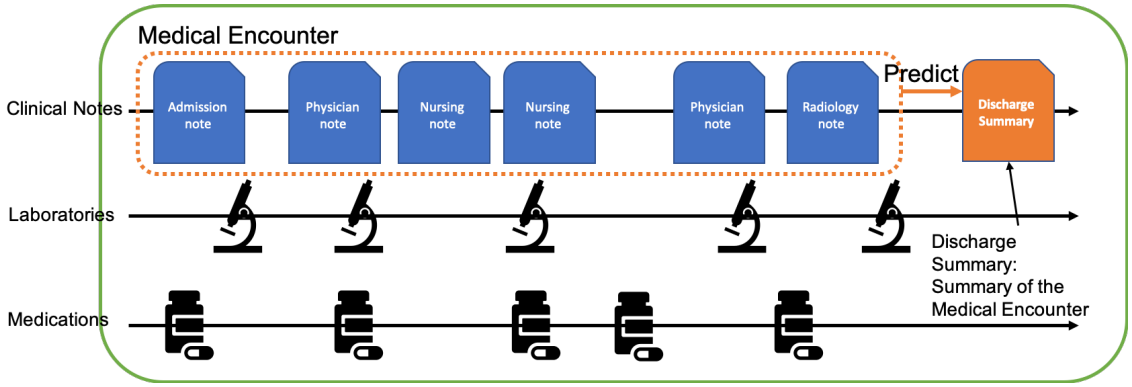

Figure 6.1: A clinical encounter is an interaction between a patient and a healthcare provider, which may contains hundreds of clinical notes.

that a summarization architecture that supports traceability yields promising results, and that a sentence-rewriting approach performs consistently better than more complex summarization models on the measure used for faithfulness (faithfulness-adjusted F_3) over a diverse range of generated sections.[1]

6.1 Discharge Summary in A Clinical Encounter

Clinical notes in the EHR are used to document the patient's progress and interactions with healthcare professionals for other healthcare professionals further downstream. These notes contain rich and diverse information, including but not limited to admission notes, nursing notes, radiology notes, and physician notes (Figure 6.1). The information downstream healthcare professionals need, however, is often buried in the sheer quantity of text. Finding the information can be time-consuming; time that is already in short supply for the healthcare professionals to attend to the patients (Weiner and Biondich, 2006; Sinsky et al., 2016), which can contribute to the worsening physician burnout crisis (Tawfik et al., 2018; West et al., 2018).

In this chapter, we focus on a specific type of clinical note: the *discharge summary*. The discharge summary is meant to summarize the clinical encounter, typically written at the time of patient discharge. Recall from Chapter 3, a clinical encounter (Figure 6.1)

[1]This chapter contains content from: **Shing, Han-Chin**, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. "Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes." In arXiv 2021.

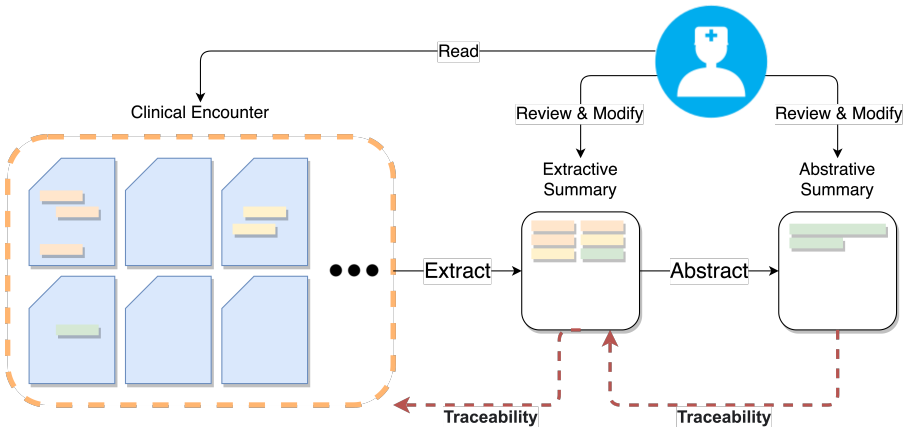

Figure 6.2: An extractive-abstractive summarization pipeline is shown with healthcare professionals in the loop. The recall-oriented extractor extracts relevant sentences from clinical documents; the abstractor smooths out irrelevant or duplicated information. Healthcare professionals can review and modify extracted content and abstracted summary, and each summary stage can be traced to its source.

documents an interaction between a patient and a healthcare provider (e.g., a visit to the hospital). In some cases, it contains hundreds of clinical notes written by healthcare professionals. Discharge summaries are also semi-structured; each section in the discharge summary (e.g., past medical history, brief hospital course, medications on admission) represents a different aspect of the encounter.

Writing a discharge summary for the encounter is, therefore, no small task. Tools that can assist healthcare professionals in writing them are placed at a premium since they hold the potential to expedite the healthcare professionals' workflow and reduce human labor. It is well-documented that healthcare professionals write some sections (e.g., past medical history, family history) by manually searching for and copying relevant content from prior clinical notes, and then rewrite the copied content to remove redundancy and into the format of the discharge summary section (Hirschtick, 2006; Pivovarov and Elhadad, 2015). Natural language processing can potentially expedite this process already in use by healthcare professionals. Manual selection of relevant content can be assisted by using a content selection module, extracting information related to the specific discharge summary selection. Rewriting and removing duplicated extracted information can be framed as an abstractive summarization task.

However, it is essential to note that this process should not be fully automated with-

out human intervention, especially in a clinical setting. The content extracted should be displayed together with their context in the source documents so that healthcare professionals can further review and modify the content by deleting irrelevant content or adding missing content. The abstractive summary generated from the selected extracted content should also only act as a starting point to be further modified by healthcare professionals. Figure 6.2 demonstrates an example of the interaction between these systems and humans. By building a system to extract and compose discharge summary sections from prior clinical notes (i.e., notes written before the discharge summary) in the same encounter, we can display the information in a format healthcare professionals can further review and modify. This is in part similar to the computer-assisted coding process we describe in Chapter 3 and the use of speech recognition to assist medical transcription (David et al., 2009), where technologies are not meant to replace humans, only to assist them.

Training and evaluating these systems with a human in the loop, however, is difficult. In this chapter, we choose to evaluate these systems by focusing on the special case where humans do not review or correct the systems' output. This allows us to model the problem of discharge summary composition as an extract-then-abstract task. In Section 6.10, we discuss the limitations of this modeling choice.

Under the extract-then-abstract framework, we identify three main challenges of discharge summary composition: (1) the *traceability* of the system that allows healthcare professionals to trace the summary to their source, (2) the *faithfulness* of the summary to the source documents, and (3) the *scalability* of the system to the extensive clinical encounter. All three challenges need to be properly addressed before a discussion about deployment can happen. Thus, this chapter focuses on measuring and understanding how existing state-of-the-art summarization systems perform on these challenges. Additionally, we propose an extractive-abstractive summarization pipeline that addresses the *traceability* challenge and the *scalability* challenge. For the third challenge, *faithfulness*, we introduce a faithfulness-adjusted evaluation measure that is based on matching *medical mentions* such as those specified in the Unified Medical Language System (UMLS, Bodenreider, 2004), inspired by recent work on faithfulness in summarization (Maynez et al., 2020; Zhang et al., 2020; Nan et al., 2021).

6.2 Traceability, Faithfulness, and Scalability

In this section, we identify three main challenges in the discharge summary composition problem under the context of the eventual goal to create a tool to help healthcare professionals create discharge summaries more efficiently.

Traceability. A summary should be displayed with a mean for the healthcare professionals reviewing the summary to inspect and understand where the information comes from. In this respect, extractive summarization has a clear advantage over abstractive summarization, as the source of the extracted content can be easily traced and displayed in context. However, abstractive summarization does benefit from more fluent generation and thus the potential to function as a writing aid to alleviate the clinicians' documentation burden. The challenge lies in how to design the system such that the summary can be traced to its source documents.

Faithfulness. Like any model supporting clinical decision making, measuring and understanding the faithfulness of the model output is important. In the context of clinical summarization, we follow the definition of Maynez et al. (2020), and define a faithful summary as a summary without any information not found in the source documents. For abstractive summarization systems, since they are trained and evaluated to generate fluent output, faithfulness can be a challenge to these models. Addressing this problem is an active area of research (Maynez et al., 2020; Zhang et al., 2020).

A faithful summary might not be useful, however. Any snippets extracted from the source documents are by definition faithful but might not be relevant to healthcare professionals' needs. In Section 6.4, we discuss how to combine faithfulness and informativeness in our faithfulness-adjusted measures.

Scalability. Summarizing an encounter (modeled as a set of documents), the quantity of text available can easily exceed the memory limit of the model. This memory limitation is especially challenging for modern transformer-based architectures that typically require large GPU-memory. Parts of the text that do not fit in memory can contain relevant clinical

information for summarization. Attempting to train an abstractive model to generate a summary without the source information available can encourage the model to generate content that is unfaithful to the input document; this is a dangerous outcome in the context of clinical summarization.

6.3 Extract and then Abstract

These challenges are common in summarization. In particular, one of the main challenges in multi-document abstractive summarization is to summarize a large number of documents. While significant progress has been made to scale abstractive models (Beltagy et al., 2020; Zaheer et al., 2020), recent work still involves first using an extractive model (e.g., tf-idf based cosine similarity (Liu et al., 2018), logistic regression (Liu and Lapata, 2019a)) to limit the number of paragraphs before abstraction.

Here we propose a similar extractive-abstractive summarization pipeline. However, in a clinical context, we wish to place more weight on the extractor rather than rely on the abstractor to summarize a large quantity of text. This decision is motivated by the fact that extractive models are inherently better at being faithful to the source, as they do not introduce explicit novel information.[2] Furthermore, by definition, an extractive summary can be traced back to the source, making them ideal candidates for clinical summarization (Pivovarov and Elhadad, 2015).

Our extractor-abstractor pipeline involves two stages (Figure 6.2). The first stage functions as a recall-oriented extractive summarization system to extract relevant sentences from prior documents. The extracted sentences are then passed through post-processing steps that remove duplicated sentences and arrange them to form an extractive summary. The second stage is an abstractive summarization system that aims to take the extractive summary from the previous step and smooths out irrelevant or duplicated information. We describe the details of implementations and how to scale this pipeline to very long text in Section 6.8.

Another advantage of this pipeline is that it provides a clear path of traceable fall-

[2]Implicit hallucination can still happen. For example, the two extracted sentences: "I gave prof a gift" and "I passed the test" put together without the original context can incorrectly imply causation.

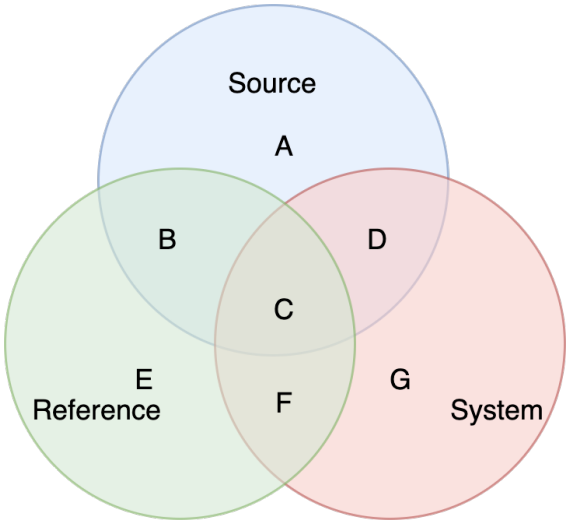

Figure 6.3: Relationship between source documents, reference summary, and system-generated summary.

back (Figure 6.2). Healthcare professionals can reference the extractive summary if they find the abstractive summary problematic or if the abstractor model has low confidence. The extractive summary can also be further traced back, as the extracted sentences came from the source documents. We can thus display the extracted sentences in context or use the extractor as a highlighter.

6.4 Measuring Faithfulness

Following prior work, we report ROUGE-n ($n = \{1, 2\}$) to measure token n-gram overlap as a proxy for informativeness, and ROUGE -L (longest common subsequence, with possible gaps) as a proxy for fluency (Lin and Hovy, 2003; Maynez et al., 2020). However, as Schluter (2017) and Cohan and Goharian (2016) have argued, ROUGE alone is insufficient and possibly misleading for measuring informativeness, specifically when it comes to faithfulness and factualness.

In a summarization setting, a faithful summary refers to a summary that does not contain information from outside of the source, as we have defined in Section 6.2. On the other hand, a factual summary allows information not presented in the source, as long as the information is factually correct. In the setting of clinical summarization, we argue that faithfulness is far more important. Novel information appearing in a summary that

has no support from the source, whether factual or not, can affect the transparency of the model.

A downside of this definition of faithfulness, however, is that it does not take reference summaries into account. Any extracted sentences (e.g., the first three sentences) from the source are always faithful by definition. Such extraction, however, might not be a summary relevant to this task. Figure 6.3 helps us illustrate the relationship between source documents, reference summary, and system-generated summary using a Venn diagram.[3]

A desirable summary, especially in a clinical setting, is *faithful* to the source and *relevant* as measured by the reference summary. In Figure 6.3, this region corresponds to $B + C$, the ideal set of information a clinical summarization system should target. Based on this observation, we define *Faithfulness-adjusted Precision* as $\frac{C}{System}$ and *Faithfulness-adjusted Recall* as $\frac{C}{B+C}$. Intuitively, faithfulness-adjusted precision measures how much information in the system-generated summary is both relevant and faithful. Similarly, faithfulness-adjusted recall measures the ratio of faithful and relevant information that has been included by the system. In a clinical setting, recall is often more important than precision; it is better to over-extract and have healthcare professionals ignore or remove the irrelevant content than have missing content. While our extractive-abstractive pipeline provides a series of *fallbacks* that allows healthcare professionals to inspect what could be missing by looking at the context of the extracted sentences, under-extraction can still happen. We therefore report a recall-oriented measure to combine the two above measures: *Faithfulness-adjusted* F_β, where we set $\beta = 3$. In this setting, faithfulness-adjusted recall is three times more important than faithfulness-adjusted precision (Van Rijsbergen, 1979).[4]

Hallucination is perhaps the leading concern when applying abstractive summarization system in a clinical setting. If one defines hallucination as a system generating content that is not faithful to the source, we can identify hallucination as the region $F+G$. G, the information that is present in neither the source nor the reference, is particularly

[3]Here we are showing a single reference summary, but in reality, the reference summary available is just one possible manifestation of all possible, potentially equally valid summaries (Nenkova and Passonneau, 2004). Our discussion can be extended to multiple reference summaries by treating each one independently in the calculations and averaging them to report the final scores.

[4]We plan to explore the values of β in consultation with healthcare professionals in future work.

Reference Summary	past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 # stable angina on long acting nitrate
System Summary	# hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 nc occupation : changes to medical and family history :
Source Documents	borderline prolonged p-r interval . intraventricular conduction delay . prior anteroseptal myocardial infarction with ongoing anterolateral and lateral myocardial ischemia . compared to tracing # 2 there is no significant change . tracing # 3 sinus rhythm . prolonged a-v conduction . left axis deviation . intraventricular conduction delay . *[.....continue with 55763 more words and 1448 more unique medical mentions]*

Table 6.1: A example calculation of faithfulness-adjusted measures. # is a symbol clinicians used to indicate an item in a list. Highlighted words are medical mentions identified by scispaCy. To calculate faithfulness-adjusted recall and precision, we need to identify the region B+C, indicated here by blue – medical mentions in the reference summary that is also found in the source documents. Orange indicates a medical mention in the reference summary that is not found in the source documents. We can then compare the medical mentions in the system summary to the blue mentions to calculate C, the faithfulness-adjusted true positive, indicated by green. Red indicates faithfulness-adjusted false positive. Thus, faithfulness-adjusted precision $= \frac{C}{System} = \frac{5}{8}$; faithfulness-adjusted recall $= \frac{C}{B+C} = \frac{5}{7}$. In this case, all false positive mentions can be found in the source, so incorrect hallucination rate $= \frac{G}{System} = \frac{0}{8}$.

problematic. We therefore measure *Incorrect Hallucination Rate* as $\frac{G}{System}$.

UMLS-based Medical Mentions as a Proxy to Information Overlap. However, an important underlying assumption of these measures is that the regions in Figure 6.3 are quantifiable. While there are many possible proxies one can use for these regions, as a starting point, we use a *medical mention* recognition system in scispaCy, a competitive model compared to state-of-the-art models across nine datasets (Neumann et al., 2019). The scispaCy model is trained on the MedMentions dataset (Murty et al., 2018) to match text span (*medical mention*) that can be linked to a medical concept in the Unified Medical Language System (UMLS, Bodenreider, 2004) Metathesaurus, an integrated biomedical terminology database. These medical mentions cover a wide range of vocabulary, including but not limited to Current Procedural Terminology; Chemical Biology and Drug Development Vocabulary; and International Classification of Diseases. After transforming the clinical text into a set of medical mentions, the cardinalities of the sets and their overlaps can then be used to calculate the above measures. See Table 6.1 for an example of using medical mentions to calculate faithfulness-adjusted measures.

6.5 Related Work

Recall from Section 2.1, we discuss related work on clinical summarization. Here we discuss related work on summarization faithfulness.

Faithfulness in Summarization. Recognizing the limitations of the existing measures and the danger of hallucination in summarization systems, faithfulness in summarization has gained attention recently (Kryscinski et al., 2020; Cao et al., 2017). Recent work on faithfulness evaluation in summarization involves using textual entailment (Maynez et al., 2020) or question answer generation (Arumae and Liu, 2019; Wang et al., 2020). For radiology summarization, Zhang et al. (2020) proposed using a radiology information extraction system to extract a pre-defined set of 14 types of medical information tailored to radiology reports.

In this chapter, we use the overlap of UMLS-based medical mentions as a proxy to a key aspect of information overlap. We argue that the domain of clinical encounter summarization is very different from the domains of most textual entailment tasks or question answer generation tasks. The domain is often much more specific, allowing us to use UMLS-based medical mentions as a proxy. However, it is not as specific as radiology summarization (Zhang et al., 2020), where a set of 14 pre-defined types of information (e.g., airspace opacity, pneumonia, cardiomegaly) can be succinctly identified.

6.6 Dataset

We derive our dataset from the MIMIC III database v1.4 (Johnson et al., 2016): a freely accessible, English-language, critical care database consisting of a set of de-identified clinical data of patients admitted to the Beth Israel Deaconess Medical Center's Intensive Care Unit (ICU). The database includes structured data such as medications and laboratory results and unstructured data such as clinical notes written by medical professionals. For this chapter, we focus on the unstructured data.

The challenge for adapting the MIMIC III database for our purpose, however, is that MIMIC III is incomplete. Due to the way that MIMIC III was collected, not all clinical

Dataset	Input	Ouput	Sample Size
Gigaword	10^1	10^1	10^6
CNN/DailyMail	10^2–10^3	10^1	10^5
WikiSum	10^2–10^6	10^1–10^3	10^6
Our Dataset	10^4–10^5	10^0–10^3	10^3

Table 6.2: Size comparison of summarization datasets. For detailed stats of the output sections of our dataset, see Table 6.4.

notes are available; only notes from ICU, radiology, echocardiogram, electrocardiogram (ECG), and discharge summary (Johnson and Shivade, 2020) are guaranteed to be available. It is important to note that the incompleteness is not a property of the problem we are trying to address; it is a property of MIMIC III. We mitigate the incompleteness issue by focusing on the subset of encounters that contain at least one admission note (a clinical note written at the time of admission) as a proxy for completeness. This leaves us about 10% of the total encounters, or around 6,000 encounters.

We identify seven discharge summary sections as our targets for summarization: (1) chief complaint, (2) family history, (3) social history, (4) medications on admission, (5) past medical history, (6) history of present illness, and (7) brief hospital course. These medical sections were chosen based on their high prevalence in discharge summaries and their length diversity (see Table 6.4).

Discharge Summary Section Extraction. To extract the seven discharge summary sections from the discharge summary, we first use a regular expression-based approach to identify the discharge summary section headers' variants from the training set. We then use a *rule-based extraction* approach to collection the discharge summary section: collecting the content from the target discharge summary section header and stop right before the next section header in the discharge summary. About one hundred randomly selected extracted discharge summary sections are manually examined to ensure no missing content or over-extraction. For each of these discharge summary sections, we then collect all prior clinical notes (according to the chart date timestamp in MIMIC III) as their source documents. On average, the source documents consist of 64 documents and 36,357 words. Table 6.2 shows a comparison with other datasets.

After the rule-based extraction, we split the encounters training, validation, and test sets (80/10/10) based on the patient's *subject id* to prevent data leakage. If the rule-based extraction returns nothing, the encounter is excluded. See Table 6.4 for the statistics of sample size.

6.7 Ethical Considerations

Deidentification. Our dataset is derived from the publicly available database MIMIC III v1.4 (Johnson et al., 2016). Johnson et al. (2016) deidentified the database in accordance with the Health Insurance Portability and Accountability Act (HIPAA) standard. This standard requires removing all eighteen identifying data elements, including patient name, telephone number, address, and dates. These fields are replaced with placeholders. A constant (but different per patient) offset is applied to shift the dates. Patients over 89 years old were mapped to over 300, in compliance with HIPAA.

Although under U.S. federal guidelines, secondary use of fully deidentified, publicly available data is exempt from institutional review board (IRB) review (45 CFR § 46.104, "Exempt research"), we still consider the dataset sensitive. We are careful to treat it as such. During training and error analysis, we of course do not attempt to identify individuals, and when qualitative analysis is shown, we double-check to avoid showing potentially identifiable information.

Population. In MIMIC III, out of the 38,161 patients, 71.34% are White, 7.69% Black, 2.38% other, 2.37% Asian, and the rest unknown. Most of the patients in MIMIC III were older adults, with the most common age group being 71–80, followed by the 61–70 age group. (Dai et al., 2020).

6.8 Models and Experiments

As explained in Section 6.3, our proposed pipeline involves an extractive summarization component and an abstractive summarization component. This section identifies a set of existing extractors and abstractors across a diverse range of different approaches to

understand what models are suitable for discharge summary composition. To understand the robustness of these approaches, we train and test these models across seven discharge summary sections with a diverse range of length and content.

Extractors. Since our goal is to summarize an encounter conditioned on a target discharge summary section, we focus our attention on supervised extractors. Supervised extractive summarization is framed as a sentence extraction problem. Each sentence is encoded into a representation used to determine whether the sentence should be included in the extracted summary. RNN or transformer-based attention are often used to encode the surrounding sentences as context.

RNN+RL$_{\text{ext}}$: Chen and Bansal (2018) proposed a method to use reinforcement learning to fine-tune a pretrained RNN sentence extractor. By modeling the next sentence to extract (including the extra "end-of-extraction" sentence) as the action space, the current extracted sentences as the state space, and by using ROUGE between the reference summary sentence and the rewritten extracted sentence (rewritten by a separate pretrained abstractor) as the reward, the authors re-purpose the sentence extractor to extract sentences from the source documents and reorder them as they might appear in the summary.

PRESUMM$_{\text{ext}}$: Liu and Lapata (2019b) proposed Presumm, a family of summarization models. Here we are especially interested in the extractive summarization variant that uses a modified pretrained BERT model (Devlin et al., 2019) to encode sentences to determine whether the sentence should be included in the extracted summary. While the model has been shown to achieve competitive results, applying a BERT encoder to very long text can be challenging in terms of memory limitations. Thus, we apply a split-map-reduce framework, where the long text is split into smaller units during training and inference. After inference, each smaller unit's extracted sentences are then concatenated back together in the same order as appeared in the original source. Since the model only assigns scores to sentences, we select the score cutoff threshold on the validation set using ROUGE-L scores, and apply that cutoff on the test set. Additionally, we select another recall-oriented score cutoff using ROUGE-L F_3 scores (calculated by placing more

weights on ROUGE-L recall). This version of the summary is termed PRESUMM$_{\text{ext-F}_3}$, which is used inside the extractor-abstractor pipeline. In contrast, PRESUMM$_{\text{ext}}$ (cutoff selected with ROUGE-L) can be examined as a standalone extractive summarization system.

Abstractors. In our extractive-abstractive pipeline, abstractors play a role in rewriting the extracted sentences to the reference summary. Here we include two abstractor variants:[5]

RNN+RL$_{\text{abs}}$: This is similar to RNN+RL$_{\text{ext}}$. However, after each sentence is extracted, it is immediately rewritten by passing through a pretrained sentence-level abstractor. The goal is to rewrite each extracted sentence to the format of what might appear in the reference summary. This sentence-rewriting approach has the disadvantage of only having a local view when rewriting (thus no merging of information). However, the advantage is that the memory limitation of sentence-level rewriting does not grow with the number of sentences, so it can be applied to longer summaries.

BART: Lewis et al. (2019) propose BART as a transformer variant that uses a bidirectional encoder similar to BERT and an autoregressive (left to right) decoder similar to GPT (Radford et al., 2019). The model has competitive performance for summarization, and thus is our choice for transformer-based abstractor. In contrast to the sentence-rewriting approach of RNN+RL$_{\text{abs}}$, we train BART to rewrite all extracted sentences together to the summary.

Baselines. Since discharge summary composition is a new task, there are no baselines from prior work. Following prior work on summarization (See et al., 2017; Liu and Lapata, 2019b), we include two extractive baselines: (1) ORACLE$_{\text{ext}}$: Extraction by using the reference summary; for each sentence in the reference summary, greedily select the source sentence in the source document that yields the maximum ROUGE-L score. (2) RULE-BASED$_{\text{ext}}$: apply the same rule-based section extraction method in Section 6.6 that was used to construct the dataset. Instead of applying it to the discharge summary, we

[5] We also experimented with a pointer-generator (See et al., 2017), but we found that BART consistently outperforms pointer-generator, so we leave the pointer-generator results in Appendix C.1.

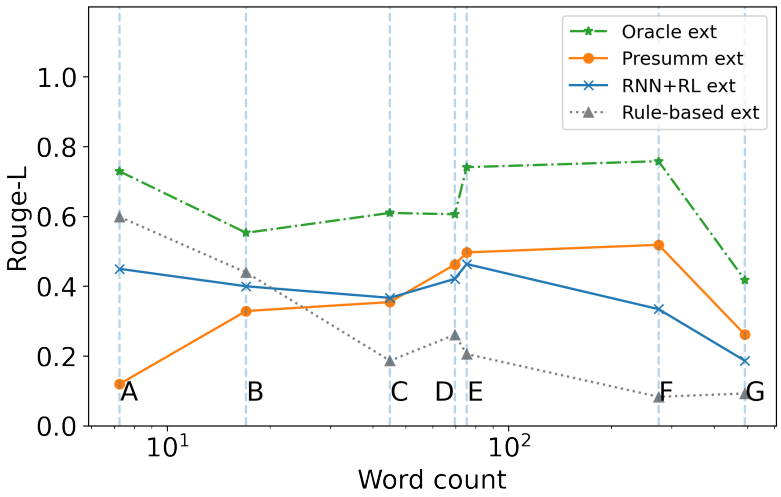

Figure 6.4: ROUGE-L of extractors vs. average word lengths (in log scale) of the discharge summary sections. Sections (dotted vertical lines) from short to long: (**A**) Chief complaint, (**B**) Family history, (**C**) Social history, (**D**) Medications on admission, (**E**) Past medical history, (**F**) History of present illness, and (**G**) Brief hospital course.

apply the same extraction method to the prior clinical documents.

Evaluating the extractor-abstractor pipeline. For the extractors, we report ROUGE scores as well as our proposed faithfulness-adjusted {precision/recall/F_3} scores across the seven discharge summary sections. These extractors include RNN+RL$_{\text{ext}}$, PRESUMM$_{\text{ext}}$, and PRESUMM$_{\text{ext-F}_3}$, as well as the two extraction baselines.

For the abstractors, we additionally measure *incorrect hallucination rate* as defined in Section 6.4. We measure the abstractive models in combination with the extractive models in our proposed pipeline. This implies measuring the performance of three models combinations: RNN+RL$_{\text{abs}}$ (uses RNN+RL$_{\text{ext}}$ as the extractor), RNN+RL$_{\text{ext}}$ + BART, and PRESUMM$_{\text{ext-F}_3}$ + BART.

6.9 Results and Discussion

Extractive summarization and abstractive summarization are often applied in different settings and should thus be compared separately. For results on ROUGE, see Table 6.3. For results on faithfulness-adjusted measures, see Table 6.4. Here we highlight the main findings.

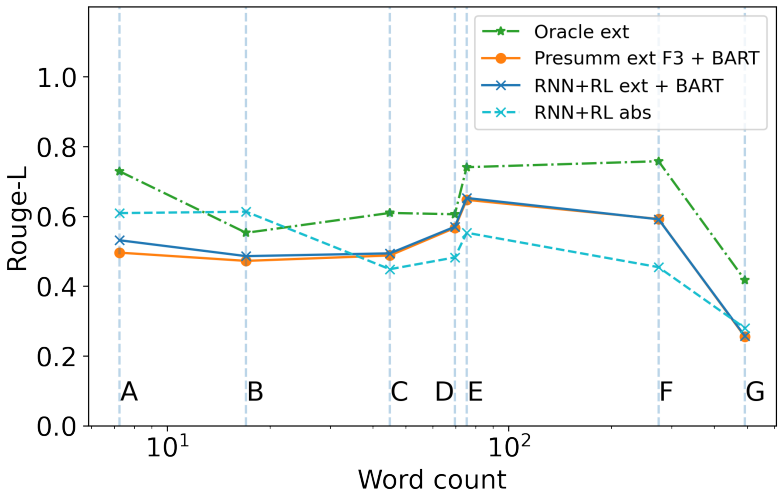

Figure 6.5: ROUGE-L of abstractors vs. average word lengths (in log scale) of the discharge summary sections. Sections order is the same as Figure 6.4. Note that PRESUMM$_{\text{ext-F}_3}$ + BART and RNN+RL$_{\text{ext}}$ + BART almost overlap completely.

	Chief Complaint	Family History	Social History	Medications on Admission	Past medical History	History of Present Illness	Brief Hospital Course
train / val / test	4,757/559/625	4,686/555/614	4,677/552/618	4,689/557/616	4,746/558/623	4,754/559/625	4,758/558/625
Output # words	7.25	17.03	44.90	69.58	75.36	274.88	491.97
Output # sents	2.04	2.63	4.93	4.67	5.99	16.62	35.39
ORACLE$_{\text{ext}}$	73.0/59.0/72.9	55.7/40.5/55.3	62.0/48.2/61.0	61.5/47.7/60.6	75.1/67.0/74.1	77.4/66.8/75.8	45.7/22.3/41.8
RULE-BASED$_{\text{ext}}$	**59.8/44.5/59.8**	**43.9/31.8/43.9**	18.6/12.1/18.6	26.1/22.2/26.1	20.6/16.3/20.6	08.3/07.3/08.3	09.2/08.5/09.2
RNN+RL$_{\text{ext}}$	45.1/33.1/45.0	40.2/28.6/40.0	**37.6/27.2/36.6**	43.4/35.6/42.1	47.9/40.2/46.3	34.8/28.3/33.4	21.3/6.7/18.6
PRESUMM$_{\text{ext}}$	12.3/06.9/11.9	33.2/24.0/32.9	**36.3/27.5/35.4**	**47.2/40.7/46.2**	**50.8/41.9/49.7**	**53.2/45.4/51.8**	**29.6/10.6/26.1**
PRESUMM$_{\text{ext-F}_3}$	11.7/06.2/11.3	32.4/23.6/32.1	28.4/20.4/27.3	38.2/32.0/37.2	48.6/40.3/47.4	48.2/40.6/46.7	26.9/8.9/23.4
RNN+RL$_{\text{ext}}$ + BART	53.5/37.5/53.1	48.9/38.6/48.6	**50.3/38.0/49.4**	**58.2/51.9/57.0**	**66.9/58.5/65.2**	**61.1/51.3/59.1**	28.2/10.6/25.7
PRESUMM$_{\text{ext-F}_3}$ + BART	49.9/33.0/49.6	47.4/37.5/47.2	**49.6/38.3/48.8**	**57.8/50.9/56.7**	**66.0/58.3/64.7**	**61.0/52.4/59.2**	28.0/**12.4**/25.5
RNN+RL$_{\text{abs}}$	**61.2/47.5/60.9**	**61.6/50.5/61.3**	45.9/33.7/44.8	49.9/42.2/48.2	57.5/47.9/55.3	47.6/38.4/45.4	**32.1**/10.4/**28.0**

Table 6.3: Dataset statistics and ROUGE-{1/2/L} scores. **Bold** score indicates statistical significance within the same extractor/abstractor group by paired t-test at $p \leq 0.05$.

The extractor's performance is modulated by the length of the output section. In Figure 6.4, we highlight the ROUGE-L scores (ROUGE-1 and ROUGE-2 have a similar pattern) of the two extractive summarization systems compared to the oracle and rule-based extractive summary. An interesting observation is the effect of length, defined as the average word count of the reference discharge summary section. RNN+RL$_{\text{ext}}$ outperforms PRESUMM$_{\text{ext}}$ on shorter sections, and vice-versa for the longer sections. This difference can be partially attributed to the way *cutoff* is being done at the extractors. For RNN+RL$_{\text{ext}}$, an RL agent is trained to decide when to stop extracting sentences. For the shorter sections, the RL agent learns to stop at just a few sentences (e.g., a typical chief complaint has two sentences, family history has on average 2.6 sentences). On longer sec-

tions, however, we find that the RL agent has difficulty stopping, causing over-extraction. In contrast, for PRESUMM$_{\text{ext}}$, a score cutoff threshold is tuned on the development set using the ROUGE-L score. This approach has a more balanced performance, but suffers at shorter sections. Another factor contributing to the lead of PRESUMM$_{\text{ext}}$ in the longer sections is our split-map-reduce framework, which enables the extractive model to conduct inference over all the clinical documents.

Interestingly, the baseline RULE-BASED$_{\text{ext}}$ performs surprisingly well on ROUGE-L for the two shortest sections. Upon inspection, most of the extraction is just the section's title, without any content. This observation is backed up by the lower faithfulness-adjusted recall of this baseline.

BART smooths out difference in extractors. We highlight the ROUGE-L of the three abstractors in Figure 6.5. Interestingly, after being abstracted by BART, both RNN+RL$_{\text{ext}}$ and PRESUMM$_{\text{ext-F}_3}$ converged to roughly the same ROUGE-L scores. This suggests that in our extractor-abstractor pipeline, BART is effective in taking different extracted content and *smoothing* them into the format and content expected for the discharge summary sections. On the other hand, RNN+RL$_{\text{abs}}$ outperforms BART at the shorter sections, and even ORACLE$_{\text{ext}}$ on the family history section. Note that ORACLE$_{\text{ext}}$ is not necessarily an upper-bound for the abstractive summarization models; abstractors allow rewriting content in prior notes into the format of discharge summary. Sentence segmentation (the basic unit of extraction) can also be noisy in clinical notes. On the other hand, the curve for RNN+RL$_{\text{abs}}$ is almost identical to RNN+RL$_{\text{ext}}$ in Figure 6.4, with a constant increase. This can largely be attributed to the sentence-level rewriting of the abstractor that allows RNN+RL$_{\text{abs}}$ to keep the benefit of its extractor counterpart, while rewriting the content to reduce over-extracted sentences.

RNN+RL$_{\text{ext,abs}}$ is more faithful and traceable Table 6.4 shows our faithfulness adjusted measures. For the extractors, RNN+RL$_{\text{ext}}$ outperforms almost all other extractors on faithfulness-adjusted F_3.[6] RNN+RL$_{\text{ext}}$ even outperforms ORACLE$_{\text{ext}}$ in the brief hos-

[6] A notable exception is PRESUMM$_{\text{ext-F}_3}$ on the two longest sections. Recall that PRESUMM$_{\text{ext-F}_3}$ directly tunes for ROUGE-L F_3.

	Chief Complaint	Family History	Social History	Medications on Admission	Past medical History	History of Present Illness	Brief Hospital Course
ORACLE$_{ext}$	71.1/85.2/83.6	52.8/75.4/72.3	63.4/73.3/72.2	69.7/66.5/66.8	74.2/80.8/80.1	76.6/83.9/83.1	44.7/51.5/50.7
RULE-BASED$_{ext}$	97.4/49.7/52.2	87.6/47.3/49.6	94.7/23.1/25.0	97.2/32.8/35.2	94.9/16.9/18.4	70.8/08.6/09.5	00.3/00.9/00.7
RNN+RL$_{ext}$	44.2/72.8/**68.4**	54.5/70.6/**68.6**	43.2/71.0/**66.7**	45.7/67.2/**64.2**	43.6/81.7/**75.1**	27.6/88.8/72.7	15.3/69.7/51.4
PRESUMM$_{ext}$	10.8/24.1/21.4	30.7/63.1/57.1	42.6/40.6/40.8	48.7/52.0/51.7	51.2/66.6/64.7	54.4/74.5/71.9	26.5/47.7/44.2
PRESUMM$_{ext\text{-}F_3}$	10.2/25.7/22.3	29.5/64.8/57.9	25.6/48.0/44.1	34.1/57.3/53.6	47.7/71.0/67.7	47.0/78.7/**73.7**	19.5/67.9/**54.4**
RNN+RL$_{ext}$ + BART	48.6/70.4/67.4	44.7/74.2/69.6	61.2/66.7/66.1	67.0/80.2/**78.7**	70.0/74.6/**74.2**	67.4/64.7/64.9	34.1/23.6/24.4
PRESUMM$_{ext\text{-}F_3}$ + BART	45.5/63.6/61.2	46.1/70.2/66.7	60.0/66.0/65.3	67.1/77.7/**76.5**	69.7/73.3/**72.9**	68.0/64.5/64.8	37.4/26.8/27.6
RNN+RL$_{abs}$	67.8/69.1/**69.0**	75.8/73.0/**73.3**	60.1/68.2/67.3	70.9/69.0/69.2	64.7/68.8/68.3	40.8/82.2/**74.6**	20.4/52.9/**45.6**

Table 6.4: Faithfulness-adjusted {Precision/Recall/**F_3**} scores based on medical mentions. **Bold F_3** score indicates statistical significance within the same extractor/abstractor group by paired t-test at $p \leq 0.05$.

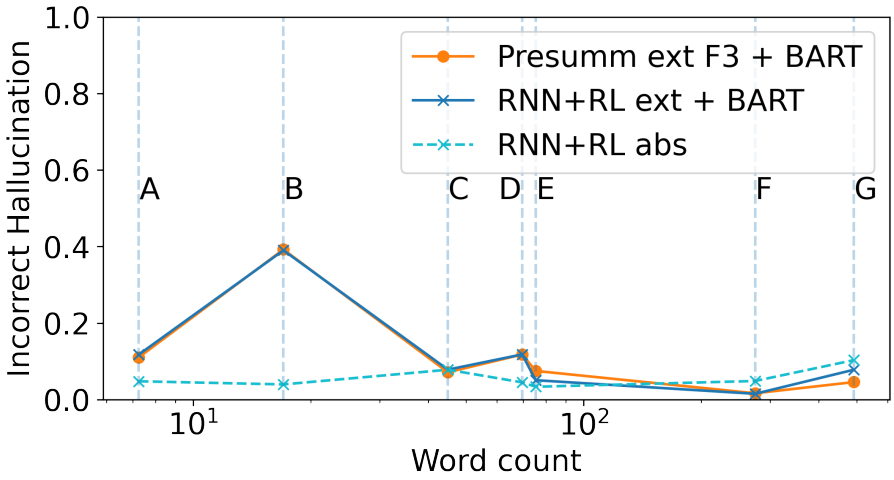

Figure 6.6: Medical mention-based incorrect hallucination rate of abstractive models vs. average word lengths (in log scale). Extractors do not hallucinate. Sections order is the same as Figure 6.4. Note that, again, PRESUMM$_{ext\text{-}F_3}$ + BART and RNN+RL$_{ext}$ + BART almost overlap completely.

pital course section. This is possible because ORACLE$_{ext}$ is selected using ROUGE-L, not faithfulness-adjusted F_3. For the abstractors, a similarly good performance is found for RNN+RL$_{abs}$, where its precision consistently increases compared to RNN+RL$_{ext}$. The good performance of RNN+RL$_{ext,abs}$ can largely be attributed to the high recall that has hurt their ROUGE-L performance in Figure 6.4 and Figure 6.5. Interestingly, the two BART models again perform roughly the same, with recall of RNN+RL$_{ext}$ + BART higher than PRESUMM$_{ext\text{-}F_3}$ + BART. For the longest section, generation for BART proves to be difficult, as indicated by the large drop of recall, whereas the sentence-wise rewriting strategy of RNN+RL$_{abs}$ has scaled better to longer sections.

Higher incorrect hallucination rate for shorter length sections. The overall incorrect hallucination rate shown in Figure 6.6 is relatively low, with the notable exception of the family history section. Inspection of the generated summaries shows that the most common hallucination of both BART systems is the phrase *"no family history"*. Interestingly, the ground truths corresponding to these hallucinations are mostly variations of the term *"non-contributory"*; there are often no family history information in the extracted summaries nor the source documents. Further inspection shows that the two phrases *"no family history"* and *"non-contributory"* are used interchangeably when no information about the family history is available in the source documents, which potentially explain the hallucination. That being said, there are still cases of hallucinations where *"no family history"* is followed by a condition (e.g., arrhythmia, cardiomyopathies) that is not mentioned in the source. Again, indicating the importance of summarization faithfulness and the need to involve healthcare professionals in the loop.

6.9.1 Qualitative Analysis

	Summary
Reference Summary	past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 # stable angina on long acting nitrate
PRESUMM$_{\text{ext-F}_3}$	# hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 nc occupation : changes to medical and family history :
RNN+RL$_{\text{ext}}$	# simvastatin 20 mg once a day # isosorbide mononitrate 40 mg once a day # furosemide 40 mg once a day # pantoprazole 40 mg once a day # diltiazem xr 180 mg once a day # tylenol for gum pain # proair hfa 90 mcg/actuation aerosol inhaler [hospital1] prn # prednisone per pt 's son 2 weeks ago # antibiotic for pneumonia per pt 's son 2 weeks ago past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 nc occupation : sinus rhythm .
PRESUMM$_{\text{ext-F}_3}$ + BART	past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3
RNN+RL$_{\text{ext}}$ + BART	past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3
RNN+RL$_{\text{abs}}$	past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.5 . .

Table 6.5: A randomly selected example showing summaries of the past medical history section. # is a symbol clinicians used to indicate an item in a list. Highlighted words are used to calculate faithfulness-adjusted measures. Blue indicates a medical mention in the reference summary that is also found in the source documents. Orange indicates a medical mention in the reference summary that is not found in the source documents. Green indicates a medical mention in the system summary that is found in the summary *and* found in the source. Red indicates a medical mention in the system summary that is not found in the summary *or* not found in the source. See Section 6.9.1 for the analysis.

	Summary
Reference Summary	family history : non-contributory . no family history of early mi , arrhythmia , cardiomyopathies , or sudden cardiac death ; otherwise non-contributory .
PRESUMM$_{\text{ext-F}_3}$	folate thiamine lisinopril carvedilol lipitor asa prevacid changes to medical and family history : changes to medical and family history : ssi folate thiamine mvi atorvastatin asa prevacid changes to medical and family history :
RNN+RL$_{\text{ext}}$	family history :
PRESUMM$_{\text{ext-F}_3}$ + BART	family history : no family history of premature coronary artery disease or sudden death .
RNN+RL$_{\text{ext}}$ + BART	family history : no family history of early mi , arrhythmia , cardiomyopathies , or sudden cardiac death .
RNN+RL$_{\text{abs}}$	family history :

Table 6.6: A randomly selected example showing summaries of the family history section. Same color coding as Table 6.5. See Section 6.9.1 for the analysis.

	Summary
Reference Summary	social history : non-smoker . denies etoh or drug use . patient is on disability . lives by himself in an apartment in [location (un) 86] .
PRESUMM$_{\text{ext-F}_3}$	adhesive tape / ibuprofen social history : denies tobacco , etoh abuse . lives by himself in an apartment in [location (un) 168] . 98 , 101/75 , 149 , 19 , 100 % 2lnc gen : denies lives in a special apartment for disabled elderly . resp were unlabored , no accessory muscle use .
RNN+RL$_{\text{ext}}$	social history : denies lives in a special apartment for disabled elderly . lives by himself in an apartment in [location (un) 168] . denies tobacco , etoh abuse . denies -illicit drugs : adhesive tape / ibuprofen social history : -tobacco history : denies -etoh :
PRESUMM$_{\text{ext-F}_3}$ + BART	social history : denies tobacco , etoh abuse . lives by himself in an apartment in [location (un) 86] .
RNN+RL$_{\text{ext}}$ + BART	social history : lives in a special apartment for disabled elderly . -tobacco history : denies -etoh : denies -illicit drugs : denies
RNN+RL$_{\text{abs}}$	social history : denies . lives by himself in an apartment in [location (un) 86] . denies tobacco , etoh abuse . denies -illicit drugs : -tobacco history : denies -etoh :

Table 6.7: A randomly selected example showing summaries of the social history section. Same color coding as Table 6.5. See Section 6.9.1 for the analysis.

Table 6.5 shows a randomly chosen summary of a past medical history section. In this case, RNN+RL$_{\text{ext}}$ over-extracted content from the previous sections (indicated by the red medical mentions). However, after passing through BART, BART successfully smooths out the noise and generates the same output as PRESUMM$_{\text{ext-F}_3}$ + BART. In this case, RNN+RL$_{\text{abs}}$ happens to be hallucinating (mapping cr 1.3 to cr 1.5), although our medication mentions do not capture that. All summarization systems missed *"# stable angina on long acting nitrate"*; mention of "stable angina" is actually not present in the source documents. Thus, we do not count "stable angina" as missing in our faithfulness-

adjusted measures.

Table 6.6 shows an example of the family history section. Note that many medical mentions in the reference summary cannot be found in the source documents (indicated by orange). Most of these mentions are variations of the word "non-contributory". This example helps illustrate our discussion from the previous section, where we mention that a significant source of incorrect hallucinations was caused by the interchangeable usage between "non-contributory" and "no family history". However, most of the mentions here, such as "sudden cardiac death", were hallucinated, as they are not present in the extracted summaries. RNN+RL$_{\text{ext}}$ + BART, for example, produces results that are almost verbatim to the reference summary (i.e., factual), while the extractor it relies on, RNN+RL$_{\text{ext}}$, contains only a single word – family history. This showcases the importance of a faithfulness-adjusted measure. In a setting where being faithful is more important than being factual, we should not encourage summarization systems to generate information that is only found in the reference summary but not in the source documents.

While all summaries capture correctly that the patient denies alcohol (e.g., etoh), tobacco, and drug use, we can see that the medical mention model we used struggles to capture the terms correctly. Specifically, the medical mention model is not aware that "non-smoker" is the same as "denies tobacco" and "denies etoh" is the same as "denies ... etoh abuse". Notably, RNN+RL$_{\text{ext}}$ + BART hallucinated, saying "that the patient lives in a special apartment for disabled elderly", which the patient specifically denied. RNN+RL$_{\text{abs}}$, on the other hand, contains repetitive information. The short sentence "denies ." also runs the risk of potentially confusing healthcare professionals. Our medical mention-based measure fails to capture these mistakes. This is a limitation to our approach of using mention overlap as a proxy to a specific aspect of information overlap. These issues can potentially be addressed by modeling negation and performing entity linking (Wu et al., 2014; Aronson, 2001; Bhatia et al., 2019).

6.10 Towards A Faithful and Traceable Clinical Summary

By extracting and composing the discharge summary from the vast number of clinical notes into a format healthcare professionals are already required to produce, our work has the potential to reduce the time healthcare professionals spend on writing the discharge summary, allowing them to allocate more time to the patients.

Clinical applications have the genuine potential to affect people's lives. As we have emphasized in Section 6.1, this chapter is not about a discussion for deployment, but rather a first step in understanding how summarization models perform as a starting point for further development. Importantly, we need to understand the failure modes of these systems and how to address these failures.

Our emphasis on faithfulness and traceability of summarization reflects those priorities. Our results show that the design of an extractive-abstractive summarization pipeline is a promising framework to address the challenges. RNN+RL$_{\text{ext,abs}}$ in particular demonstrates consistent performance on the faithfulness-adjusted measures. The nature of sentence-level rewriting in RNN+RL$_{\text{abs}}$ also gives it the advantage to produce a traceable summary on a per-sentence level, as each abstracted sentence has a direct mapping to an extracted sentence.

One limitation of this chapter is our choice to evaluate our system without explicitly considering human behavior. While we design our extract-then-abstract framework to allow healthcare professionals to review and modify (by deleting irrelevant content or adding new content) both the extracted content and the abstracted summary, there can still be error modes introduced by a human in the loop. The trade-off between the time saved and potential errors need to be further studied. For example, in machine translation, users place trust more based on fluency rather than the faithfulness of the translation (Martindale et al., 2019). More research is needed to judge if the fluent output of the abstractors can potentially mislead healthcare professionals to overlook faithfulness. Conversely, when hallucinations do happen, how does that affect the healthcare professionals' trust in using these systems? Another limitation to our approach is that if there is novel information available only when writing the discharge summary, there will be no way of

summarizing it. It is also important to note that since we are using MIMIC III for training and evaluation, the results shown are biased toward the dataset. MIMIC III is an English-language collection from the ICU of a single hospital, not necessarily applicable to other clinical settings.

However, the three challenges we identify and the extract-then-abstract framework can serve as the first of many future steps to alleviate the documentation burden of clinicians and ultimately result in better quality of care for patients.

Chapter 7: Conclusions and Future Work

Information exchange between healthcare professionals and patients, and between different healthcare professionals, is critical to better patient care (Weiner and Biondich, 2006). With the wide adoption of EHR and growing health-related information, however, handling both patients and their data has proven to be laborious, as healthcare professionals simply do not have enough time. NLP holds the potential to make this process in healthcare more efficient. By modeling the extensive unstructured data in patients' records, NLP can assist assessment, allocating more time for healthcare professionals to take care of patients (Demner-Fushman et al., 2009).

In this dissertation, we examine NLP's role in assisting healthcare professionals using three examples – computer-assisted coding (Chapter 3), suicidality risk assessment using social media postings (Chapters 4 and 5), and discharge summary composition from prior clinical notes (Chapter 6). Throughout these examples, we ask (1) how to better model the extensive patient data and (2) how to design systems to surface relevant information to support healthcare professionals.

Modeling the Patient as a Set of Documents. Modeling patient data is challenging as structured and unstructured data about the patient are extensive and complex. In the lens of NLP for healthcare applications, we show that modeling the patient as *a set of documents*, the unstructured information about the patient, is a useful abstraction. This abstraction allows us to represent information at the patient or encounter level while connecting to the lower document level. For example, in our clinical coding work, we represent information at the patient's encounter level as a set of clinical notes after identifying an important label mismatch problem: clinical codes are assigned to the patient's encounter, but most prior work focuses on code prediction for a single document. In our

discharge summary composition work, we again represent information at the patient's encounter level as a set of clinical notes. Instead of inferring clinical codes, we extract and rewrite the information in the encounter into discharge summary sections. In suicidality risk assessment, we represent the individual-level information as a sequence of social media postings. The risk of suicide is a property of an individual, but the language evidence we intend to study is on a subset of the documents they posted.

Modeling the *entire* document set, however, is a challenge to NLP in healthcare settings. Information relevant to the task is often buried in the extensive document collection. For example, in suicidality assessment, high signal postings are often diluted by a large number of postings not directly related to suicidal intent. Similarly, in clinical coding, not all documents contain evidence supporting the medical codes. In Chapters 3 and 5, we introduce document-level attention and show how it provides a mechanism for the model to jointly learn which documents are important for prediction and what the model should predict. We show that by introducing a mechanism to narrow the model's focus on the subset of documents with information relevant to the prediction, the model's predictive performance consistently improves over the baselines that use averaging.

On the practical side, the extensiveness of patient data also presents a challenge to current deep learning methods. This is especially true in Chapter 6, where the focus is to summarize the encounter into a discharge summary. Averaging 36K words and 64 documents, the entire clinical encounter simply does not fit into memory. To address this challenge, we observe that, similar to Chapters 3 and 5, not all content in the encounter is relevant to the summary. The key, then, is to extract these relevant snippets in the documents and then merge them at the encounter level. We thus apply an extractive summarization system to individual documents separately to extract relevant snippets. The extracted snippets collected from all documents are then merged either using a duplication removal process or an abstractive summarization system.

Prioritizing Relevant Information for Assessment. The extensive data of a patient are a challenge to the models *and* a challenge to the healthcare professionals. While healthcare professionals already lack time, many tasks in NLP for healthcare applications

cannot and should not be fully automated without healthcare professionals' involvement. Thus, the NLP system must be evaluated not just by its ability to make inferences correctly but also by its ability to provide evidence for that inference.

The time saved by surfacing evidence in the assessment can potentially give healthcare professionals a chance to address more patients. That being said, models are often not evaluated by their ability to save time. In Chapter 5, we reframe suicidality assessment from a classification problem to a hierarchical ranking problem: ranking both the individuals and their postings. To measure the potential time we can save for healthcare professionals, we introduce a theoretically grounded measure, hTBG. Using that together with the expert-annotated UMD Reddit Suicidality dataset, we demonstrate that we can potentially reduce human assessment time by using document attention to surface documents with high suicidal signals. In Chapter 6, our focus on traceability again reflects a similar emphasis on surfacing evidence. For clinical summarization systems that aim to help healthcare professionals write discharge summaries, there should also be a mechanism that allows the summary to be traced back to the source, providing a form of evidence to the healthcare professionals. Our extractive-abstractive pipeline design allows healthcare professionals to trace the source of the abstractive summary to the extractive summary and the source of the extractive summary to the source documents.

Throughout the three examples in this dissertation, we have demonstrated how extensive patient data could challenge both models and healthcare professionals. We address the modeling challenge by allowing the system to have a mechanism to attend to a smaller subset of the patient data. Interestingly, the same mechanism can assist the healthcare professionals by surfacing relevant information for their tasks. This duality between what is useful for modeling patients and what is useful for healthcare professionals is exemplified throughout this dissertation. Chapter 3 shows that document attention leads to improved clinical code prediction, and the attention learned aligns with professional coders' expectations. In Chapter 5, we again show that document attention leads to better suicidality risk prediction, and the attention can be used to rank document, potentially saving healthcare professionals' time. Finally, in Chapter 6, our extractive-abstractive pipeline design helps us scale abstractive summarization systems to the extensive patient's data while still

maintaining the ability to be traced back to the extractive content and the source documents. This duality between humans and machines sheds light on a promising direction that will hopefully lead to building more human-centric NLP systems for healthcare.

7.1 Limitations

Robustness of the Results. Availability and access to test collections for NLP for healthcare and mental health research are often limited and sometimes nonexistent. As a result, testing the robustness of our conclusions can be difficult. In this dissertation, we aim for a robust result whenever possible. Effectiveness of document attention is tested in Chapter 3 using 3M Health Information Systems datasets and again in Chapter 5 using the UMD Reddit Suicidality dataset we collected in Chapter 4. For clinical coding, we test the document attention on a held-out test set where the patients are from different hospitals. We tested our extractive and abstractive models across seven different discharge summary sections of varying lengths and properties for the robustness of discharge summary composition.

However, further studies are needed to see if the results can transfer across different domains. We demonstrate the potential of using document attention to surface documents on the UMD Reddit Suicidality dataset. The same dataset was also used for a shared task in CLPsych 2019 (Zirikly et al., 2019) and has since been shared with more than 35 teams internationally. However, whether the same results will hold needs to be tested for social media settings other than Reddit; in particular, evidence suggests that users show different behaviors when posting anonymously, with both positive and negative implications (Christopherson, 2007; De Choudhury and De, 2014). It is also important to note that, similar to many mental health-related datasets (Harrigian et al., 2021), the annotations are proxy diagnostic (in our case, expert judgments), not clinical ground truth. Thus the external clinical validity needs to be further tested (Ernala et al., 2019). We derive our data for discharge summary composition from MIMIC III, a database collected from a single hospital ICU unit. While it is freely available, it is limited in size and domain. Like the many other studies that rely on MIMIC III, the transferability of our

results to other healthcare settings needs further study.

Transferability across different times and cultures can also be an issue. Clinical codes are updated annually to account for new diagnoses and new procedures. Furthermore, the ICD code set goes through significant structural changes roughly every ten years. Major health events can also lead to significant changes. For example, six new diagnosis codes were added to ICD-10-CM due to COVID-19. Clinical notes in the ICU and social media postings indicating the risk of suicide can also look very different before and after COVID-19. Our studies are primarily US-based and involve the English language. In a different culture, the conclusions and even framing of the problem may need to be adjusted. For example, suicidality assessment in a community where suicide is stigmatized may need to be approached differently. Clinical coders in different countries, besides potentially speaking different languages, face different challenges (McKenzie et al., 2004). Canada, for example, has different coding standards for medical procedures for each province and territory (Welch et al., 1993; Hu, 2021).

Effects on Key Stakeholders. Healthcare applications have the genuine potential to affect people's lives. Therefore, understanding how our system can affect key stakeholders – in this case, the patients, the healthcare professionals, and the sometimes overlooked researchers – is important. In this dissertation, we focus on providing a mechanism to surface relevant information to healthcare professionals. However, advances in NLP for healthcare may affect patients. Potential bias to different patient demographics in the problem formulation and dataset needs to be further studied. Ethical and privacy concerns in collecting patients' data are also critical discussions that need to happen.

Benton et al. (2017) and Chancellor et al. (2019) discuss some of these issues in-depth in the context of health and mental health research using social media. There is a tension between the potential benefits and risk of harm. Potential benefits may include early detection of depression (Losada et al., 2019), alternative source of evidence for suicidality assessment (Resnik et al., 2020, also Chapter 5), and design of new intervention techniques for schizophrenia and suicide (Mitchell et al., 2015; de Andrade et al., 2018). On the other hand, incorrect and hard-to-decipher prediction of the systems

may cause harm to the individuals involved. Malintentioned actors can take advantage of these systems to harass and stalk individuals at risk even when the systems themselves were built with good intention (Barrie, 2014). Intentional and unintentional bias in how and where the dataset is collected can amplify the already uneven distribution of health resources (Olteanu et al., 2019; Blodgett et al., 2020).

Addressing this tension is an ongoing discussion that should involve healthcare professionals, researchers, and individuals who are the object of these predictions (Chancellor et al., 2019). Among these participants, opinions from the individuals were probably the most neglected. Mikal et al. (2016) and Fiesler and Proferes (2018) conducted focus group and survey studies with Twitter users that reveal a range of opinions on social media research. For example, while users, in general, understand that Twitter data is public, most users are not aware that their data can be used for large-scale health research. Obtaining informed consent, or "opt-in", is thus a practice researchers should follow whenever possible. Sensitive information should also be protected with appropriate measures and de-identified when it is not needed for analysis (Benton et al., 2017). Another conversation, perhaps equally important, is the mechanism and design of these NLP systems for healthcare (Abebe and Goldner, 2018; Green, 2019). Understanding where, how, and even *if* these NLP systems should be deployed is critical if we want to realize the potential benefits and not end up harming the individuals these systems are designed to help.

In Chapter 5, we reframe suicidality assessment as a prioritization problem. By ranking the at-risk individuals jointly with their documents, we show that our system has the potential to support a more efficient assessment process, allowing healthcare professionals to assess more at-risk individuals in a given time budget. However, similar to other ranking applications, ranking can lead to potential biases, amplifying the un-fairness between groups and individuals (Yang and Stoyanovich, 2017; Singh and Joachims, 2018; Biega et al., 2018). In the setting of suicidality assessment, a careful study of the fairness of ranking is needed. Additionally, our desire to maximize the number of at-risk individuals that can be assessed in a given time budget leads to the *speed-biased* criterion (see Section 5.3). That is, for equally at-risk individuals, the measure rewards ranking the individual who can be assessed more quickly closer to the top. This implies that our eval-

uation measure has a bias for individuals who write shorter and more explicit documents. While this is a direct consequence of our criterion, one can argue that this behavior is not ideal. New criteria that address these concerns might be needed. However, these criteria are fundamentally normative statements about the role of NLP in the suicidality assessment process. Thus, the discussion should involve the research community and the groups and individuals who will be affected the most – patients and healthcare professionals.

Modeling Assumptions and Need for User Studies. George Box said, "All models are wrong, but some are useful." (Box, 1979) Throughout this dissertation, we made many assumptions and approximations to our evaluation measure and our task formulation. To better understand whether the conclusions hold, user studies are needed. Here we list our key assumptions.

A contribution of this dissertation is modeling the patient or the individual as a set of documents. However, the data of these patients and individuals are often much more complex. A patient's clinical encounter, for example, contains structured and unstructured data. In Chapter 3 and Chapter 6, we do not take structured data, such as the laboratory results, medications, and transfer data from different hospital units, into account. Metadata about the documents can also be important. For example, the timestamp of the document and the time between consecutive documents can be a valuable signal for suicidality assessment. Data about the patient can also be incomplete. In Chapter 6, MIMIC III can have missing clinical notes due to the method of construction, which we have to mitigate by limiting the encounters to those with admission notes. Ground truth can also be challenging to obtain. In Chapter 5, we annotate the risk of suicidality using experts' and crowdsourcers' consensus. However, it is important to know that this is just a proxy to the actual risk and not a clinical ground truth.

Another contribution of this dissertation is surfacing relevant information for assessment. What information should be surfaced and how it should be surfaced depends on how the task is formulated. Our evaluation measures that evaluate how well this information is surfaced, as we mentioned, is a model of the task. For example, hTBG, the evaluation measure used in our suicidality work, has many assumptions about how healthcare

professionals use the system. One of them is the document independence assumption, where we assume that a single isolated document that contains signals of suicide risk is enough to identify an at-risk individual. Our faithfulness-adjusted measures in Chapter 6 use the overlap of the UMLS medical mentions as a proxy to measure information overlap. However, it is well understood that errors in generation can come in many forms. The modifiers (e.g., negation, degree) used on the UMLS medical mentions can carry important information. Context can also matter. For example, whether the information is meant to describe the patient or their family member can have very different implications. Thus, our approach of using the UMLS medical mentions only measures a specific aspect of the information overlap. How well these assumptions hold in the real world needs to be tested with user studies.

7.2 Future Work

Many limitations in the previous section have the potential to be pursued in future work. Here we focus on two directions.

Patient as a set of *distributed* documents. Our work demonstrates how a patient can be modeled as a set of documents. In some cases, these patient data contain an extensive set of documents, creating a challenge to fit the documents into the model. One possibility to address this challenge is explored in Chapter 6, where we apply extractive models to a subset of the patient records individually and then combine them to form the final summary. This framework is similar to distributed computing, specifically, the MapReduce framework (Dean and Ghemawat, 2008; Wickham et al., 2011). Under this framework, we first *map* a document-level model to individual documents, possibly in a distributed setting. Another *reduce* model can then be used to aggregate the results.

An interesting possibility on top of this is to explore how the gradient may flow through this framework. Using 3HAN as an example, we can apply 2HAN to individual documents to obtain document-level representation. The document-level representations collected from all individual documents can then be aggregated by document attention before making the final inference. To enable distribution of the models and allow the

gradient to flow through the process, we can build on existing work such as model parallelism (Kim et al., 2020) and asynchronous models training (Guu et al., 2020).

Traceability of abstractive summarization. The ability to trace back to the source of the summary is important in a healthcare setting. In extractive summarization, a clear default is to trace back to where the summary snippets are extracted. In abstractive summarization, however, there are many potential candidate approaches. One possible approach is to use heuristics. These include lexical-based approaches that greedily select the source snippets by maximizing ROUGE scores between the generated summary and the potential source snippets, or attention-based approaches that use attention (if available) in the abstractive model to locate possible snippets. Abstractive models can also encourage traceability by design. Chen and Bansal (2018), mentioned in Chapter 6, use an extractor to extract relevant sentences. Each extracted sentence is then independently mapped to an abstractive sentence. It thus provides a natural way to trace each summary sentence back to the source sentence.

The discussion of traceability extends beyond the healthcare setting, as it strongly relates to the faithfulness of a summary. We can view traceability as a complementary approach to summarization faithfulness. A traceable summary snippet indicates that it is faithful; a non-traceable summary snippet may alert a potential hallucination. In contrast to faithfulness measures, which only generate a single score indicating whether a summary is faithful to the source, traceability aims to provide evidence of the summary snippets by pointing back to the source documents.

It is unclear, however, which of these methods are the most effective in terms of traceability. We thus break down this future direction into two sub-questions. (1) How do we design a user interface that supports the traceability of the summary? (2) Under this user interface, how do we evaluate the effectiveness of the traceability of different approaches?

7.3 Implications

In Plato's *Phaedrus*, Pharaoh Thamus said about writing: "If men learn this, it will implant forgetfulness in their souls: They will cease to exercise memory because they rely on that which is written." (Plato, 370 BCE, trans. 1972) This is far from just an argument against writing; it is pointing out that the implications and social impacts of new technologies are always difficult to foresee. This dissertation focuses on building NLP systems for healthcare applications and making them center around humans – the patients and the healthcare professionals. It is difficult to foresee the potential positive impact and negative impact of these NLP systems. There are still many limitations to our work. Many more iterations of research about ethics, bias, and privacy also need to happen before discussions on deployment can take place. This work can be viewed as the first step of, hopefully, many future steps that will lead to technologies that can positively impact society. We hope that by designing NLP systems around healthcare professionals' needs and by placing patients at the center of modeling, our work can contribute to a more efficient workflow for healthcare professionals and ultimately better care for patients.

Appendix A: Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings

A.1 Dataset Availability and Ethical Considerations

The research we report was approved by the University of Maryland's Institutional Review Board (IRB). As Benton et al. (2017) discuss, human subjects research using previously existing data falls into a category exempted from the requirement of full IRB review as long as the data are either from publicly available sources or they do not provide a way to recover the identity of the subjects. In our case, the data are publicly available *and* from a site where users are anonymous. As an extra precaution we replace Reddit usernames with numeric identifiers.

Benton et al. (2017) point out that even exempt research needs to be reviewed by an IRB to make an exemption determination. In addition, they discuss the importance of taking particular care with sensitive data. In order to share the data with other researchers while ensuring appropriate standards are met, the dataset has been made available through a collaborative process with the American Association of Suicidology (AAS), an organization whose mission is to promote the understanding and prevention of suicide and support those who have been affected by it.[1] AAS helped develop, and participates in, governance in which researchers submit requests for access, with panel review ensuring, for example, that proper IRB procedures have been followed, that the researchers will provide appropriate protections for sensitive data, and that there will be no linkage of the dataset to other sites that could jeopardize user anonymity. As of this writing the dataset has been shared with more than 35 research teams internationally.

As discussed by Chancellor et al. (2019), there is another ethical point to be con-

[1] http://www.suicidology.org/about-aas/mission

sidered: protection not of subjects but of annotators and researchers. Reading postings like the ones that appear on SuicideWatch is hard, and it is difficult to know how the experience might affect non-experts and even experts. With that in mind, our instructions to annotators warned explicitly that the materials might be upsetting, and we encouraged people to err on the side of caution and stop doing the annotation task if it was affecting them in a personal way. We also provided contact information for the National Suicide Prevention Hotline, Crisis Text Line, National Suicide Prevention Lifeline, and a link to the SuicideWatch hotlines page.

A.2 Annotation Instructions

The following is the *long* instruction used for the annotation of the UMD Reddit Suicidality dataset. We include the instruction here for reference and completeness. Some examples of this instruction contain snippets from real Reddit posts. While they were fine to include for the annotators, we obfuscate them here for privacy. The credit of this instruction goes to Philip Resnik and the co-authors of Shing et al. (2018).[2]

Identifying Risk Of Suicide In Social Media Posts

This task will help with a project where the ultimate goal is finding new ways to help prevent suicides. But before you go any further, please recognize that some of the things you will see here are from people in real distress, and they can be difficult or upsetting to read. If you believe that you might be affected personally in a negative way by doing this task, *please err on the side of caution and stop here; do NOT do the task*. If you start the task and you find that it's upsetting, *please stop*. If you're feeling like you (or someone you know) could use some support or assistance, *please take advantage of one of the following resources*:

- National Suicide Prevention Lifeline: 1-800-273-8255 (TALK).
  - Veterans please press 1 to reach specialized support.

[2]Contact Philip Resnik for the full instruction.

- Spanish: 1-800-SUICIDA
- Crisis Text Line: Text "START" to 741-741
- Online chat: http://www.suicidepreventionlifeline.org/gethelp/lifelinechat.aspx
- https://www.reddit.com/r/SuicideWatch/wiki/hotlines - This page provides information about phone and chat hotlines and online resources in the U.S. and worldwide.

Please note that all the posts you're looking at are anonymous – not even the researchers know who these people are, and the posts were made over a period of years. Although it's tragic that there is no direct way for us to help the people who have written these posts who may be at risk of suicide, you are contributing to an effort aimed at better understanding the factors connected with suicide attempts, using that information to do a better job assessing risk, and hopefully contributing to more effective ways of getting people help.

Instructions

You'll be looking at posts written by users in an online discussion forum. We are trying to answer this question: *What is the risk of this person attempting suicide?*

Identifying risk of suicide accurately can't be done perfectly, but here are factors that are often taken into account when judging risk. Note that some of them may not be obvious – for example, research shows that someone who is showing signs of agitation can be at higher risk than someone who just seems down or depressed. Here are other factors that could bump the assessment of risk from a lower level to a higher level, grouped roughly into *thoughts*, *feelings*, *logistics*, and *context*.

- Thoughts
  - Thinking about suicide, having suicide on their mind
  - Having told friends or family they are thinking about suicide
  - Feeling that they are a burden to others

  - Endorsement of suicidal beliefs, even without the word suicide (e.g., I deserve to die, I can never be forgiven for the mistakes I made)

- Feelings
  - Expressing lack of hope for things to get better
  - A sense of agitation, not being able to "stand still" physically or mentally
  - Indications of being impulsive; risky behavior (e.g. reckless driving, promiscuity)

- Logistics
  - Talking about plans that involve suicide
  - Talking about methods of attempting suicide, even if not planning
  - Preparation, actually taking actions to prepare for an attempt
  - Having access to lethal means (a way to take their own life), especially firearms
  - Having enough privacy or isolation to make an attempt

- Context
  - Previous attempts
  - An event or life change that is leading them think about suicide
  - Isolation from friends and family

Here are the ways you can label a user:

- **No Risk**:
  - **I don't see evidence that this person is at risk for suicide. (If this person were my friend, the idea of them possibly making a suicide attempt might not even occur to me.)**
  - Example: *[obfuscated for privacy]*

    - This person is talking about someone else who may be at risk. There is no evidence they are themselves at risk.
  - Example: *[obfuscated for privacy]*
    - This person may be having some feelings of isolation, but there is no evidence that they are at risk for suicide.

- **Low Risk**:
  - **There may be some factors here that could suggest risk, but I don't really think this person is at much of a risk of suicide. (If this person were my friend, the possibility of them making a suicide attempt is not something I would feel worried about.)**
  - Example: *[obfuscated for privacy]*
    - This person suffers from depression, but otherwise there are no suicidal thoughts, feelings (e.g. lack of hope, impulsivity), logistics/planning, or context that would suggest the possibility of a suicide attempt.
  - Example: *[obfuscated for privacy]*
    - This person is talking mainly about someone else, which is similar to the first "no risk" example above. However, the context is a tragic life event that just took place, and the person is also talking about their own feelings of confusion and loss. Although it's a borderline example, those factors are enough to err on the side of saying "low risk" according to the description above, rather than "no risk".

- **Moderate Risk**:
  - **I see indications that there could be a genuine risk of this person making a suicide attempt. (If this person were my friend, the possibility of a suicide attempt is something I would be feeling worried about.)**
  - Example: *[obfuscated for privacy]*

    - This person has expressed suicidal beliefs or wishes ("make me wish I wasn't around"). However, what they are saying does not yet indicate that these thoughts and feelings are moving in the direction of action.
  - Example: *[obfuscated for privacy]*
    - This person has expressed suicidal thinking ("is there even a point to keep trying at life") and hopelessness, and described a specific event or change (knowing that they are going to flunk out). However, what they are saying does not yet provide an indication that these thoughts are moving in the direction of action.

- **Severe Risk**:
  - **I believe this person is at high risk of attempting suicide in the near future. (If this person were my friend, I would be feeling really urgently worried.)**
  - Example: *[obfuscated for privacy]*
    - *[obfuscated for privacy]* [overdose] (previous attempt, discussing methods).
    - *[obfuscated for privacy]* (suicidal thinking)
    - *[obfuscated for privacy]* (farewell, feeling like a burden, sense of a plan to take action)
    - *[obfuscated for privacy]* (discussing methods)
    - *[obfuscated for privacy]* (concerns about lack of control or impulsivity)
    - *[obfuscated for privacy]* (overt reference to suicidal thoughts)
    - *[obfuscated for privacy]* ("game over" thinking)
    - *[obfuscated for privacy]* (suicidal wishes)
  - Example: *[obfuscated for privacy]*
    - *[obfuscated for privacy]* (farewell, hopelessness)
    - *[obfuscated for privacy]* (specific plan/method)
    - *[obfuscated for privacy]* (feeling like a burden, clear intent and determination)

    - *[obfuscated for privacy]* (clear intent and determination)
    - *[obfuscated for privacy]* ("game over", farewell)

As another example, consider the following user:

User 8579374590

- **Post 1.** *[obfuscated for privacy]*
- **Post 2.** *[obfuscated for privacy]*
- **Post 3.** *[obfuscated for privacy]*
- **Post 4.** *[obfuscated for privacy]*
- **Post 5.** *[obfuscated for privacy]*

To what extent would you judge this person as being at risk of making a suicide attempt in the near future?

- a [ ] I don't see evidence that this person is at risk for suicide. (If this person were my friend, the idea of them possibly making a suicide attempt might not even occur to me.)
- b [ ] Low risk: There may be some factors here that could suggest risk, but I don't really think this person is at much of a risk of suicide. (If this person were my friend, the possibility of them making a suicide attempt is not something I would feel worried about.)
- c [x] Moderate risk: I see indications that there could be a genuine risk of this person making a suicide attempt. (If this person were my friend, the possibility of a suicide attempt is something I would be feeling worried about.)
- d [ ] I believe this person is at high risk of attempting suicide in the near future. (If this person were my friend, I would be feeling really urgently worried.)

If you chose (b), (c), or (d), which post most strongly supports your conclusion?

[ ]1 [ ]2 [ ]3 [x]4 [ ]5

In this case, Post 1 suggests that the person is down and feeling isolated, with a long-term history of negative thinking. In this case it would make sense to upgrade your choice from "no risk" (a) to "low risk" (b), since these are factors that could suggest risk, although depression by itself does not necessarily trigger concern about suicidality.

Post 2 uses the phrase "kill myself", but the person is likely to be joking or using the phrase in an informal way, so it doesn't indicate risk, even in the context of the previous post.

Post 3 says explicitly that this person suffers from depression, and subjectively one might also say it contains a hint of despair. However, it does not suggest a cause for worry that would lead to upgrading risk beyond (b).

Post 4 has a clear indication that this person has suicide on their mind, even if they wrote "I would never kill myself". There are clear references to thinking about suicide. Based on this post, you would upgrade your response to (c), "moderate risk".

Post 5 mentions a negative event but does not provide any additional sense of urgency or suggest that a suicide attempt is more likely in the near future, which is the distinction between (c) and (d).

You *should never downgrade your choice*. If an earlier post suggests a person is at severe risk ("I'm going to kill myself"), and you read a later post suggesting the risk has decreased ("I've decided not to kill myself"), please stick with the higher risk in your answer, and list the severe-risk post as the basis for your judgment.

Again, please note that if you're having trouble with the difficult content in some of these postings, you can decline to do the task or you can stop at any time, and there are resources listed at the top of these instructions if you feel like you or someone you know could use some support or assistance. Thank you for your help.

Rater info

Before we get started, please fill in the following:

Name _____

Email _____

Brief description of training/experience. Please provide a few sentences describing any relevant training and/or background you have pertaining to assessment of suicidality.

Thanks!

Appendix B: A Prioritization Model for Suicidality Risk Assessment

B.1 Appendix: Ethical Considerations

Our research involving the University of Maryland Reddit Suicide Dataset has undergone review by the University of Maryland Institutional Review Board with a determination of Category 4 Exempt status under U.S. federal regulations. For this dataset, (a) the original data are publicly available, and (b) the originating site (Reddit) is intended for anonymous posting. In addition, since Reddit is officially anonymous, but that is not enforced on the site, the dataset has undergone automatic de-identification using named entity recognition aggressively to identify and mask out potential personally identifiable information such as personal names and organizations, in order to create an additional layer of protection (Zirikly et al., 2019). In an assessment of de-identification quality, we manually reviewed a sample of 200 randomly selected posts (100 from the SuicideWatch subreddit and 100 from other subreddits), revealing zero instances of personally identifiable information.

Following Benton et al. (2017), we treat the data (even though de-identified) as sensitive and restrict access to it, we use obfuscated and minimal examples in the dissertation and presentations, and we do not engage in linkage with other datasets.

The dataset is available to other researchers via an application process put in place with the American Association of Suicidology that requires IRB or equivalent ethical review, a commitment to appropriate data management, and, since ethical research practice is not just a matter of publicly available data or even IRB approval (Zimmer, 2010; Benton et al., 2017; Chancellor et al., 2019), a commitment to following additional ethical guidelines. Interested researchers can find information at http://umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html.

B.2 Appendix: Proofs for TBG and hTBG

B.2.1 Time-Biased Gain

In order to prove that TBG statisfies the *speed-biased* criterion, consider two individuals ranked at consecutive positions k and $k+1$; if we swap the two individual, the change in TBG score is:

$$\begin{aligned}\Delta\text{TBG} = &(g_{k+1} - g_k)D(T(k)) \\ &+ g_k D\left(T(k) + t(k+1)\right) \\ &- g_{k+1} D\left(T(k) + t(k)\right)\end{aligned} \tag{B.1}$$

This leads to Lemma B.2.1-B.2.3:

Lemma B.2.1. *Swapping a not-at-risk individual ranked at k with an at-risk individual ranked at $k+1$ always increases TBG.*

Proof. Let $g_k = 0$ and $g_{k+1} > 0$. Equation B.1 simplifies to

$$\Delta\text{TBG} = g_{k+1}\left(D(T(k)) - D(T(k) + t(k))\right) \tag{B.2}$$

which is always positive because the decay function monotonically decreases, and each assessment of an individual requires at least T_s seconds. □

Lemma B.2.2 (Risk-based Criterion)**.** *The optimal value of TBG under binary relevance is obtained only if all not-at-risk individuals are ranked below all at-risk individuals.*

Proof. Let π be a ranking of individuals that yields the optimal value of TBG. Assume that in π there exist not-at-risk individuals ranked before at-risk individuals. Let the k-position be the lowest ranked not-at-risk individual that is at least in front of one at-risk individual, we can then apply Lemma B.2.1 to increase TBG. This leads to a contradiction. □

Lemma B.2.3. *Swapping an at-risk individual of longer assessment time ranked at k of with an at-risk individual of shorter assessment time ranked at $k+n$, where $k+n$ is the closest at-risk individual ranked lower than k, always increases TBG.*

Proof. Let $g_k = g_{k+n} > 0$, and $\forall i \in \{i | k < i < k+n\}, g_i = 0$. We have

$$\begin{aligned} \Delta\text{TBG} = g_k(&D(T(k+n) + t(k+n) - t(k)) \\ &- D(T(k+n))) \end{aligned} \tag{B.3}$$

which is always positive because the decay function monotonically decreases, and $t(k+n) < t(k)$ from the assumption that the individual at $k+n$ has shorter assessment time. □

Lemma B.2.3 naturally leads to a proof for the speed-biased property of TBG:

Proof for Theorem 5.3.1. Applying Lemma B.2.3, we know that swapping k and $k+r$ leads to a positive gain between the two. Now, consider all at-risk individuals ranked between k and $k+r$: $\forall u$, s.t. $k < u < k+r$, the difference is:

$$g_u(D(T(u) + t(k+r) - t(k)) - D(T(u))) \tag{B.4}$$

which is always greater than or equal to zero due to the fact that the decay function monotonically decrease, and $t(k+r) < t(k)$. Thus, the net difference is always larger than zero, thus satisfying the *speed-biased* criterion. □

Finally, combing previous results, we can easily show:

Proof for Theorem 5.3.2. A direct consequence of Theorem 5.3.1 is that if the at-risk individuals are sorted by assessment time in ascending order, no swapping between any two individuals can increase TBG. This, combined with Lemma B.2.2, that all at-risk individuals are on top of not-at-risk individuals, leads to the necessary condition. Because any swapping within the not-at-risk individuals does not change TBG when no at-risk individuals are ranked lower, this implies that ranking according to Theorem 5.3.2 gives

us a unique and optimal value, which satisfies the sufficient condition of Theorem 5.3.2.

□

B.2.2 Hierarchical Time-Biased Gain

The assessment time of an individual ranked at k, $t(k)$, is monotonic with E_i, thus showing minimal value of E_i suffices. Recall that E_i is calculated as:

$$E_i = T_\alpha \sum_{l=1}^{L} \left(W_{i,l} \prod_{m=1}^{l-1} (1 - R_{i,m}) \right) + T_\beta \tag{B.5}$$

Consider, again, swapping a document at rank l with a document at rank $l+1$ belonging to the same individual i. The change in E_i is:

$$\Delta E_i = \kappa_{i,l} \left(W_{i,l+1} R_{i,l} - W_{i,l} R_{i,l+1} \right) \tag{B.6}$$

where $\kappa_{i,l} = T_\alpha \prod_{j=1}^{l-1} (1 - R_{i,j}) \geq 0$ is a fixed term that is not affected by the swap.

Equation B.6 also points to an important observation:

Lemma B.2.4. *If $W_{i,l+1} R_{i,l} - W_{i,l} R_{i,l+1} < 0$ and $R_{i,j} < 1$ for all $j < l$, then swapping document l with document $l+1$ will decrease E_i.*

Proof. This follows directly from Equation B.6. □

Lemma B.2.5. *If $R_{i,j} < 1$ for all j, then minimum individual assessment time is obtained if and only if the documents are sorted in descending order by*

$$\frac{R_{i,l}}{W_{i,l}}. \tag{B.7}$$

Proof. Let τ be a document ranking that yields the minimum individual assessment time, and for the sake of contradiction, not a ranking that can be obtained by ranking according to $\frac{R_{i,l}}{W_{i,l}}$. We can, thus, find two neighboring documents, without loss of generality, l and $l+1$, such that:

$$\frac{R_{i,l}}{W_{i,l}} < \frac{R_{i,l+1}}{W_{i,l+1}} \tag{B.8}$$

this leads to:

$$R_{i,l}W_{i,l+1} - R_{i,l+1}W_{i,l} < 0 \tag{B.9}$$

since all $W > 0$. Lemma B.2.4 together with the prerequisite that $R_{i,j} < 1$ for all j then suggest that swapping the two leads to a decrease of E_i. This contradicts with the assumption that τ is an optimal ranking. This proves that to achieve minimum individual assessment time, it is necessary to sort by $\frac{R_{i,l}}{W_{i,l}}$. The sufficient condition follows by the fact that swapping tied documents does not lead to change in E_i, as shown in Equation B.6 □

Proof for Theorem 5.3.3. Let τ be a document ranking according to $\frac{R_{i,l}}{W_{i,l}}$. Let m be the document such that $R_{i,m} = 1$ and is ranked closer to the top then any other document with $R_{i,:} = 1$ (i.e. with the shortest $W_{i,:}$). Now, consider using m to cut the documents into two partitions: the first partition of documents are ones ranked before m. Applying Lemma B.2.5, this partition of documents are already in optimal sorted order, since there's no $R_{i,:} = 1$. The second partition, documents ranked lower than m, the ranking simply does not matter, as Equation B.5 shows, the $(1 - R_{i,m})$ term will make everything zero afterwards.

Now, let's consider moving a document from the second partition to the first partition. Since any documents in the second partition has a $\frac{R_{i,j}}{W_{i,j}}$ that is smaller than any documents in the first partition, after you move the document, the optimal ranking for the first partition will put the document at the bottom, right next to m. And since $\frac{R_{i,m}}{W_{i,m}} \geq \frac{R_{i,j}}{W_{i,j}}$ due to the original ordering, we can apply Lemma B.2.4, which can swap the document back below m. Next, consider moving the lowest ranked document of the first partition (the one ranked at $m - 1$) to the second partition. This will always increase E_i, as shown from Lemma B.2.4. Moving any other document in the first partition will also increase E_i as least as much as before, since the process is equivalent to swapping with (and thus potentially increase E_i) any intermediate documents in between.

Combine these two together, we show that E_i is at a minimum value when sorted in descending order according to $\frac{R_{i,l}}{W_{i,l}}$. □

B.2.3 Relationship between ERR and hTBG

Here we show the derivation from the cascading user model in ERR to the individual assessment time estimation (E_i) in hTBG. ERR assumes a stopping probability (written in hTBG terms):

$$P(\text{stop at l}) = R_{i,l} \prod_{j=1}^{l-1} (1 - R_{i,j}) \tag{B.10}$$

The expected words read, can then be calculated as:

$$\begin{aligned} &\sum_{l=1}^{L} \left(P(\text{stop at } l) \sum_{d=1}^{l} W_{i,l} \right) \\ =&\sum_{l=1}^{L} \left(R_{i,l} \prod_{j=1}^{l-1} (1 - R_{i,j}) \left(\sum_{d=1}^{l} W_{i,l} \right) \right) \end{aligned} \tag{B.11}$$

This can be rearranged to the formula we used in hTBG:

$$\sum_{l=1}^{L} \left(W_{i,l} \prod_{m=1}^{l-1} (1 - R_{i,m}) \right) \tag{B.12}$$

by letting $R_{i,L} = 1$ (the user has to stop reading at the last document). To show this, observe that $W_{i,1}$ appears in all L terms of the summation, thus the coefficient for $W_{i,1}$ is simply $\sum_{l=1}^{L} (R_{i,l} \prod_{j=1}^{l-1} (1 - R_{i,j})) = 1$, from both simple manipulation and the fact that we are summing over probability. Similarly, $W_{i,2}$ appears in all L terms except with $l = 1$, thus $(1 - R_{i,1})$. For $W_{i,3}$ it is $(1 - R_{i,1}) - R_{i,2}(1 - R_{i,1}) = \prod_{j=1}^{2} (1 - R_{i,j})$. The rest follows.

B.3 Appendix: Training Details

All models are built using AllenNLP (Gardner et al., 2018). Tokenization and sentence splitting are done using spaCy (Honnibal and Johnson, 2015).

The CROWDSOURCE dataset is split into a training set (80%) and a validation set (20%) during model development. We did not test on the EXPERT dataset until all param-

eters of the models were fixed. Cross validation on the training set is used for hyperparameter tuning. For 3HAN, we used ADAM with learning rate 0.003, trained for 100 epochs with early stopping on the validation dataset, with patience set to 30. For 3HAN_Av, the same hyperparameters are used. For LR, we used SGD with learning rate 0.003, trained for 100 epochs with early stopping on the validation dataset, with patience set to 30.

Both 3HAN and 3HAN_Av's Seq2Vec layers use bi-directional GRU with attention. The word-to-sentence layer has input dimension of 200, hidden dimension of 50, and output dimension of 100, since the GRU is bi-directional. The sentence-to-document and document-to-individual layer, similarly, has input dimension of 100, hidden dimension of 50, and output dimension of 100. Hyperparameters were selected using cross validation on the training set split of the CROWDSOURCE dataset.

Appendix C: Learning to Compose Discharge Summaries from Prior Notes

C.1 Appendix: Full Results

See Table C.1 and Table C.2 for the full scores for all models on all seven sections.

	Chief Complaint	Family History	Social History	Medications on Admission	Past medical History	History of Present Illness	Brief Hospital Course
$\text{Oracle}_{\text{ext}}$	73.0/59.0/72.9	55.7/40.5/55.3	62.0/48.2/61.0	61.5/47.7/60.6	75.1/67.0/74.1	77.4/66.8/75.8	45.7/22.3/41.8
$\text{Rule-based}_{\text{ext}}$	**59.8/44.5/59.8**	**43.9/31.8/43.9**	18.6/12.1/18.6	26.1/22.2/26.1	20.6/16.3/20.6	8.3/7.3/8.3	9.2/8.5/9.2
$\text{RNN+RL}_{\text{ext}}$	45.1/33.1/45.0	40.2/28.6/40.0	**37.6/27.2/36.6**	43.4/35.6/42.1	47.9/40.2/46.3	34.8/28.3/33.4	21.3/6.7/18.6
$\text{Presumm}_{\text{ext}}$	12.3/6.9/11.9	33.2/24.0/32.9	**36.3/27.5/35.4**	**47.2/40.7/46.2**	**50.8/41.9/49.7**	**53.2/45.4/51.8**	**29.6/10.6/26.1**
$\text{Presumm}_{\text{ext-F}_3}$	11.7/6.2/11.3	32.4/23.6/32.1	28.4/20.4/27.3	38.2/32.0/37.2	48.6/40.3/47.4	48.2/40.6/46.7	26.9/8.9/23.4
$\text{RNN+RL}_{\text{ext}}$ + PointGen	21.2/13.2/21.1	29.8/22.0/29.5	36.7/26.3/36.2	49.2/41.7/48.1	46.3/38.6/45.0	38.8/28.3/37.4	20.6/8.6/19.2
$\text{Presumm}_{\text{ext-F}_3}$ + PointGen	19.8/11.6/19.7	30.6/23.5/30.5	42.5/31.1/41.4	50.0/43.0/49.0	52.4/45.0/51.2	43.0/35.2/41.6	20.9/9.6/19.4
$\text{RNN+RL}_{\text{ext}}$ + BART	53.5/37.5/53.1	48.9/38.6/48.6	**50.3/38.0/49.4**	**58.2/51.9/57.0**	**66.9/58.5/65.2**	**61.1/51.3/59.1**	28.2/10.6/25.7
$\text{Presumm}_{\text{ext-F}_3}$ + BART	49.9/33.0/49.6	47.4/37.5/47.2	**49.6/38.3/48.8**	**57.8/50.9/56.7**	**66.0/58.3/64.7**	**61.0/52.4/59.2**	28.0/**12.4**/25.5
$\text{RNN+RL}_{\text{abs}}$	**61.2/47.5/60.9**	**61.6/50.5/61.3**	45.9/33.7/44.8	49.9/42.2/48.2	57.5/47.9/55.3	47.6/38.4/45.4	**32.1**/10.4/**28.0**
# words	7.25037	17.026	44.9034	69.5803	75.3616	274.881	491.971
# sents	2.04183	2.63082	4.92901	4.67285	5.99115	16.6193	35.389

Table C.1: ROUGE-{1/2/L} scores, across different models and sections

	Chief Complaint	Family History	Social History	Medications on Admission	Past medical History	History of Present Illness	Brief Hospital Course
$\text{ORACLE}_{\text{ext}}$	71.1/85.2/83.6	52.8/75.4/72.3	63.4/73.3/72.2	69.7/66.5/66.8	74.2/80.8/80.1	76.6/83.9/83.1	44.7/51.5/50.7
$\text{RULE-BASED}_{\text{ext}}$	97.4/49.7/52.2	87.6/47.3/49.6	94.7/23.1/25.0	97.2/32.8/35.2	94.9/16.9/18.4	70.8/08.6/09.5	00.3/00.9/00.7
$\text{PRESUMM}_{\text{ext}}$	10.8/24.1/21.4	30.7/63.1/57.1	42.6/40.6/40.8	48.7/52.0/51.7	51.2/66.6/64.7	54.4/74.5/71.9	26.5/47.7/44.2
$\text{PRESUMM}_{\text{ext-F}_3}$	10.2/25.7/22.3	29.5/64.8/57.9	25.6/48.0/44.1	34.1/57.3/53.6	47.7/71.0/67.7	47.0/78.7/**73.7**	19.5/67.9/**54.4**
$\text{RNN+RL}_{\text{ext}}$	44.2/72.8/**68.4**	54.5/70.6/**68.6**	43.2/71.0/**66.7**	45.7/67.2/**64.2**	43.6/81.7/**75.1**	27.6/88.8/72.7	15.3/69.7/51.4
$\text{PRESUMM}_{\text{ext-F}_3}$ + POINTGEN	31.3/62.6/56.9	37.0/72.3/66.0	54.7/61.9/61.1	65.1/73.7/72.8	64.0/62.6/62.7	69.8/42.4/44.1	42.2/17.9/19.0
$\text{RNN+RL}_{\text{ext}}$ + POINTGEN	40.6/70.2/65.4	38.2/73.9/67.6	59.9/58.7/58.8	66.4/72.7/72.0	65.6/59.0/59.6	69.1/37.1/38.9	39.8/15.2/16.2
$\text{PRESUMM}_{\text{ext-F}_3}$ + BART	45.5/63.6/61.2	46.1/70.2/66.7	60.0/66.0/65.3	67.1/77.7/**76.5**	69.7/73.3/**72.9**	68.0/64.5/64.8	37.4/26.8/27.6
$\text{RNN+RL}_{\text{ext}}$ + BART	48.6/70.4/67.4	44.7/74.2/69.6	61.2/66.7/66.1	67.0/80.2/**78.7**	70.0/74.6/**74.2**	67.4/64.7/64.9	34.1/23.6/24.4
$\text{RNN+RL}_{\text{abs}}$	67.8/69.1/**69.0**	75.8/73.0/**73.3**	60.1/68.2/67.3	70.9/69.0/69.2	64.7/68.8/68.3	40.8/82.2/**74.6**	20.4/52.9/**45.6**

Table C.2: Faithfulness-adjusted $\{Precision/Recall/F_3\}$ scores based on UMLS medical mentions.

C.2 Appendix: Reproducibility

Here we describe the training details of the models for reproducibility.

RNN+NL$_{ext}$ and RNN+NL$_{abs}$. Both models are trained following the original recipe from Chen and Bansal (2018). The training setup involves the following steps: (1) use gensim to train a word2vec embedding from scratch from the training set of the source documents, (2) construct pseudo pairs of sentences (source sentence, summary sentence): for each summary sentence, greedily finds the one-best source sentence using ROUGE-L recall, (3) use the pseudo pairs to train an RNN extractor, (4) use the pseudo pairs to train a pointer-generator that rewrites the sentences, and (5) train an RL agent that fine-tunes the RNN extractor with the sentence-rewriting pointer-generator. Model is trained on one V100 GPU, with an Adam optimizer of learning rate 1e-3. Here we use the same set of hyperparameters as Chen and Bansal (2018). For more details, please refer to the original paper.

For each of the seven medical sections, we follow the training recipe, and repeat it five times. The reported models are chosen based on the validation set. We found that the RL fine-tuning step can potentially be very unstable. For longer sections (e.g., brief hospital course and history of present illness), the RL fine-tuning can even fail to converge.

PRESUMM$_{ext}$. We use the original implementation released with PRESUMM (Liu and Lapata, 2019b). Learning rate is set to 2e-3 and extractor dropout rate is set to 0.1, following the original paper. `bert-base-uncased` is used as the pretrained BERT model. We made three important changes: (1) increase the maximum tokens the encoder can consume to 1024 tokens, (2) in the data preprocessing step, we construct pseudo pairs of sentences that will be later used to train the extractor: for each summary sentence, greedily finds the one-best source sentence using ROUGE-L recall, and (3) before the training begin, we split the source documents and their labels into segments smaller than 1024 tokens. After inference finishes, we concatenate the segments (together with a extraction score for each sentence) back together in the original order.

For each of the seven medical sections, we train the model on 4 V100 GPUs, with 150,000 training steps and model checkpointing every 2,000 steps. We report the model with the lowest model loss on the validation set. Since the model only assigns scores to

sentences, we sweep the threshold of score cutoff on the validation set using ROUGE-L score, and apply that cutoff on the test set.

POINTGEN. We use an open implementation of pointer-generator (See et al., 2017), implemented with PyTorch and AllenNlp.[1] Our model follows the original paper and has 256-dimensional hidden states and 128-dimensional word embeddings. The vocabulary size is set to 50k words for both source and target. The model is optimized using Adagrad with learning rate 0.15 and an initial accumulator value of 0.1, and trained on one v100 GPU for 50 epochs with early stopping on the validation set.

BART. We use the `Fairseq` (Ott et al., 2019) implementation of BART-large (Lewis et al., 2019) as it is shown to achieve the state-of-the-art ROUGE scores for abstractive summarization. We fine-tune the BART-large model with the standard learning rate of 3×10^{-5}. We utilize a machine with 8 GPUs and batch size of 2048 input tokens per GPU. We train for a maximum of 10 epochs with early stopping to select the checkpoint with the smallest loss on the validation set. During decoding, we use beam search with beam size of 6. We restrict the generation length to be between 10 to 300 tokens.

[1] https://github.com/kukrishna/pointer-generator-pytorch-allennlp

Bibliography

Rediet Abebe and Kira Goldner. Mechanism design for social good. *AI Matters*, 4(3): 27–34, October 2018. doi: 10.1145/3284751.3284761. URL https://doi.org/10.1145/3284751.3284761.

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, page 5–14, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605583907. doi: 10.1145/1498759.1498766. URL https://doi.org/10.1145/1498759.1498766.

Hans Åhlfeldt, Lars Borin, Natalia Grabar, Catalina Hallett, David Hardcastle, Dimitrios Kokkinakis, Clara Mancini, Kornél Markó, Magnus Merkel, Christian Pietsch, et al. Literature review on patient-friendly documentation systems. 2006.

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 561–568. MIT Press, 2002. URL http://papers.nips.cc/paper/2232-support-vector-machines-for-multiple-instance-learning.

Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

Kristjan Arumae and Fei Liu. Guiding extractive summarization with question-answering rewards. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2566–2577, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Krisztian Balog. *Entity-Oriented Search*, volume 39 of *The Information Retrieval Series*. Springer, 2018. ISBN 978-3-319-93933-9. doi: 10.1007/978-3-319-93935-3. URL https://doi.org/10.1007/978-3-319-93935-3.

Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2–3):127–256, 2012. ISSN 1554-0669. doi: 10.1561/1500000024. URL http://dx.doi.org/10.1561/1500000024.

Joshua Barrie. People are freaking out over this new anti-suicide twitter app, Nov 2014. URL https://www.businessinsider.com/people-freaking-out-over-samaritans-twitter-app-2014-11.

Philip J Batterham, Maria Ftanou, Jane Pirkis, Jacqueline L Brewer, Andrew J Mackinnon, Annette Beautrais, A Kate Fairweather-Schmidt, and Helen Christensen. A systematic review and evaluation of measures for suicidal ideation and behaviors in population-based research. *Psychological assessment*, 27(2):501, 2015.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. Multi-label classification of patient notes: Case study on ICD code assignment. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 409–416. AAAI Press, 2018. URL https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16881.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In Dirk Hovy, Shannon L. Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna M. Wallach, editors, *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL*, pages 94–102. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-1612. URL https://doi.org/10.18653/v1/w17-1612.

Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. Comprehend medical: a named entity recognition and relationship extraction web service. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851. IEEE, 2019.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414, 2018.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

David E. Bloom, Elizabeth Cafiero, Eva Jané-Llopis, Shafika Abrahams-Gessel, Lakshmi Reddy Bloom, Sana Fathima, Andrea B. Feigl, Tom Gaziano, Ali Hamandi, Mona

Mowafi, Danny O'Farrell, and Emre. The Global Economic Burden of Noncommunicable Diseases. PGDA Working Papers 8712, Program on the Global Demography of Aging, January 2012. URL https://ideas.repec.org/p/gdm/wpaper/8712.html.

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

George EP Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

Bureau of Health Workforce. Designated health professional shortage areas: Statistics, first quarter of fiscal year 2021 designated hpsa quarterly summary, December 2020. Health Resources and Services Administration (HRSA) U.S. Department of Health & Human Services, https://data.hrsa.gov/Default/GenerateHPSAQuarterlyReport.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685, 2017.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. *arXiv preprint arXiv:1711.04434*, 2017.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.*, 77:329–353, 2018. doi: 10.1016/j.patcog.2017.10.009. URL https://doi.org/10.1016/j.patcog.2017.10.009.

Ben Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*, pages 903–912, 2011.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. doi: 10.1145/3287560.3287587.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009*, pages 621–630. ACM, 2009. doi: 10.1145/1645953.1646033. URL https://doi.org/10.1145/1645953.1646033.

Yen-Chun Chen and Mohit Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*, pages 675–686, 2018.

Munmun De Choudhury. Role of social media in tackling challenges in mental health. In Pablo César, Matthew Cooper, David A. Shamma, and Doug Williams, editors, *Proceedings of the 2nd international workshop on Socially-aware multimedia, SAM@ACM*

Multimedia 2013, pages 49–52. ACM, 2013. doi: 10.1145/2509916.2509921. URL https://doi.org/10.1145/2509916.2509921.

Kimberly M Christopherson. The positive and negative implications of anonymity in internet social interactions:"on the internet, nobody knows you're a dog". *Computers in Human Behavior*, 23(6):3038–3056, 2007.

Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, 1:343–359, 2007. doi: 10.4324/9780203837702.

Charles LA Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 75–84, 2011.

Arman Cohan and Nazli Goharian. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, 2016.

Mike Conway and Daniel O'Connor. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82, 2016.

G. Coppersmith, C. Hilland, O. Frieder, and R. Leary. Scalable mental health analysis in the clinical whitespace via natural language processing. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 393–396, 2017.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*, 2015.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117, 2016.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018.

Darcy J Corbitt-Hall, Jami M Gauthier, Margaret T Davis, and Tracy K Witte. College students' responses to suicidal content on social networking sites: An examination using a simulated facebook newsfeed. *Suicide and Life-Threatening Behavior*, 46(5): 609–624, 2016.

Zheng Dai, Siru Liu, Jinfa Wu, Mengdie Li, Jialin Liu, and Ke Li. Analysis of adult disease characteristics and mortality on mimic-iii. *PLOS ONE*, 15(4):1–12, 04 2020. doi: 10.1371/journal.pone.0232176. URL https://doi.org/10.1371/journal.pone.0232176.

Bharath Dandala, Venkata Joopudi, and Murthy Devarakonda. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. *Drug safety*, 42(1):135–146, 2019.

Gary C David, Angela Cora Garcia, Anne Warfield Rawls, and Donald Chand. Listening to what is said–transcribing what is heard: the impact of speech recognition technology (srt) on the practice of medical transcription (mt). *Sociology of Health & Illness*, 31 (6):924–938, 2009.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. Ethics and artificial intelligence: suicide prevention on facebook. *Philosophy & Technology*, 31(4):669–684, 2018.

Munmun De Choudhury and Sushovan De. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *ICWSM*, 2014.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM, 2016.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Dina Demner-Fushman and Jimmy Lin. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 841–848, 2006.

Dina Demner-Fushman, Charlotte Seckman, Cheryl Fisher, Susan E Hauser, Jennifer Clayton, and George R Thoma. A prototype system to support evidence-based practice. In *AMIA Annual Symposium Proceedings*, volume 2008, page 151. American Medical Informatics Association, 2008.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2009.08.007. URL https://www.sciencedirect.com/science/article/pii/S1532046409001087. Biomedical Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*

(Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3. URL https://doi.org/10.1016/S0004-3702(96)00034-3.

HP Dinwoodie and RW Howell. Automatic disease coding: the'fruit-machine'method in general practice. *British journal of preventive & social medicine*, 27(1):59, 1973.

Keith J Dreyer, Mannudeep K Kalra, Michael M Maher, Autumn M Hurier, Benjamin A Asfaw, Thomas Schultz, Elkan F Halpern, and James H Thrall. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology*, 234(2):323–329, 2005.

Richard Duszak Jr, Michael Nossal, Lyle Schofield, and Daniel Picus. Physician documentation deficiencies in abdominal ultrasound reports: frequency, characteristics, and financial impact. *Journal of the American College of Radiology*, 9(6):403–408, 2012.

Tracy Edinger, Aaron M Cohen, Steven Bedrick, Kyle Ambert, and William Hersh. Barriers to retrieving patient information from electronic health record data: failure analysis from the trec medical records track. In *AMIA annual symposium proceedings*, volume 2012, page 180. American Medical Informatics Association, 2012.

Katie G Egan, Rosalind N Koff, and Megan A Moreno. College students' responses to mental health status updates on Facebook. *Issues in mental health nursing*, 34(1): 46–51, 2013.

Noemie Elhadad and Komal Sutaria. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56, 2007.

Elinore F. McCance-Katz, SAMHSA. *The National Survey on Drug Use and Health: 2019*. Center for Behavioral Health Statistics and Quality, September 2020. URL https://www.samhsa.gov/data/sites/default/files/reports/rpt29392/Assistant-Secretary-nsduh2019_presentation/Assistant-Secretary-nsduh2019_presentation.pdf. p. 46.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

Richárd Farkas and György Szarvas. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, page S10. BioMed Central, 2008.

Oladimeji Farri, David S Pieckiewicz, Ahmed S Rahman, Terrence J Adam, Serguei V Pakhomov, and Genevieve B Melton. A qualitative analysis of ehr clinical document synthesis by clinicians. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1211. American Medical Informatics Association, 2012.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding topic signals in large-scale text. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade, editors, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016. doi: 10.1145/2858036.2858535. URL https://doi.org/10.1145/2858036.2858535.

Joshua C. Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F. Sittig. Summarization of clinical information: A conceptual model. *Journal of Biomedical Informatics*, 44(4):688 – 699, 2011. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2011.03.008. URL http://www.sciencedirect.com/science/article/pii/S1532046411000591.

Ji Feng and Zhi-Hua Zhou. Deep MIML network. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1884–1890. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14747.

Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366, 2018.

Auguste H Fortin, Francesca C Dwamena, Richard M Frankel, and Robert C Smith. *Smith's patient centered interviewing: an evidence-based method*. McGraw Hill Professional, 2012.

Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187–232, 2017. ISSN 1939-1455. doi: 10.1037/bul0000084. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/bul0000084.

Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

Devin Gaffney and J. Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PLOS ONE*, 13(7):1–13, 07 2018. doi: 10.1371/journal.pone.0200162. URL https://doi.org/10.1371/journal.pone.0200162.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. Association for Computational Linguistics, July 2018. doi: 10.18653/v1/W18-2501. URL https://www.aclweb.org/anthology/W18-2501.

Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik, editors. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Online, June 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.clpsych-1.0.

Susan Dorr Goold and Mack Lipkin Jr. The doctor–patient relationship: challenges, opportunities, and strategies. *Journal of general internal medicine*, 14(Suppl 1):S26, 1999.

Ben Green. Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, 2019.

Kathleen M Griffiths, Louise Farrer, and Helen Christensen. The efficacy of internet interventions for depression and anxiety disorders: a review of randomised controlled trials. *Medical Journal of Australia*, 192(11):S4, 2010.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

David Hardcastle and Catalina Hallett. Exploring the use of nlp in the disclosure of electronic patient records. In *Biological, translational, and clinical language processing*, pages 161–162, 2007.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.2. URL https://www.aclweb.org/anthology/2021.clpsych-1.2.

Brian Hazlehurst, H Robert Frost, Dean F Sittig, and Victor J Stevens. Mediclass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association*, 12(5):517–529, 2005.

Holly Hedegaard, Sally C Curtin, and Margaret Warner. Suicide rates in the United States continue to increase. *National Center for Health Statistics*, 2018. URL https://www.cdc.gov/nchs/data/databriefs/db309.pdf.

Paul M Heider, Jihad S Obeid, and Stéphane M Meystre. A comparative analysis of speed and accuracy for three off-the-shelf de-identification tools. *AMIA Summits on Translational Science Proceedings*, 2020:241, 2020.

Robert E. Hirschtick. Copy-and-Paste. *JAMA*, 295(20):2335–2336, 05 2006. ISSN 0098-7484. doi: 10.1001/jama.295.20.2335. URL https://doi.org/10.1001/jama.295.20.2335.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, 2014.

Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1373–1378. The Association for Computational Linguistics, 2015. doi: 10.18653/v1/d15-1162. URL https://doi.org/10.18653/v1/d15-1162.

George Hripcsak, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of internal medicine*, 122(9):681–688, 1995.

Jia Hu. Coding our way to a more agile health system. *Policy Options*, Mar 2021. URL https://policyoptions.irpp.org/magazines/march-2021/coding-our-way-to-a-more-agile-health-system/.

Dereck L Hunt, R Brian Haynes, Steven E Hanna, and Kristina Smith. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *Jama*, 280(15):1339–1346, 1998.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2132–2141. PMLR, 2018. URL http://proceedings.mlr.press/v80/ilse18a.html.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digital Medicine*, 3(1):1–9, dec 2020. doi: 10.1038/s41746-020-0267-x.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1357. URL https://doi.org/10.18653/v1/n19-1357.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL http://doi.acm.org/10.1145/582415.582418.

Hongyan Jing and Kathleen R McKeown. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136, 1999.

Alistair Johnson and Chaitanya Shivade. Notes and data not in mimic-iii · issue 771 · mit-lcp/mimic-code, Jul 2020. URL https://github.com/MIT-LCP/mimic-code/issues/771.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Thomas E Joiner, Jr, Rheeda L Walker, Jeremy W Pettit, Marisol Perez, and Kelly C Cukrowicz. Evidence-based assessment of depression in adults. *Psychological Assessment*, 17(3):267, 2005.

Thomas E Joiner Jr, Rheeda L Walker, M David Rudd, and David A Jobes. Scientizing and routinizing the assessment of suicidality in outpatient practice. *Professional psychology: Research and practice*, 30(5):447, 1999.

Chiheon Kim, Heungsub Lee, Myungryong Jeong, Woonhyuk Baek, Boogeon Yoon, Ildoo Kim, Sungbin Lim, and Sungwoong Kim. torchgpipe: On-the-fly pipeline parallelism for training giant models, 2020.

Sunil Kripalani, Frank LeFevre, Christopher O Phillips, Mark V Williams, Preetha Basaviah, and David W Baker. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *JAMA*, 297(8):831–841, 2007.

Klaus Krippendorff. Reliability in content analysis. *Human communication research*, 30 (3):411–433, 2004.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, 2020.

Scott H. Lee. Natural language generation for electronic health records. *npj Digital Medicine*, 1(1):1–7, dec 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0070-0.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. A novel system for extractive clinical note summarization using ehr data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, 2019.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157, 2003.

Fabienne Lind, Maria Gruber, and Hajo G Boomgaarden. Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures*, 11(3):191–209, 2017.

Kathryn P. Linthicum, Katherine Musacchio Schafer, and Jessica D. Ribeiro. Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law*, 37(3):214–222, may 2019. ISSN 0735-3936. doi: 10.1002/bsl.2392. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bsl.2392.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Tong Liu, Qijin Cheng, Christopher M Homan, and Vincent Silenzio. Learning from various labeling strategies for suicide-related messages on social media: An experimental study. *ACM International Conference on Web Search and Data Mining Workshop on Mining Online Health Reports*, February 2017.

Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, 2019a.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of EMNLP*, pages 3721–3731, 2019b.

David E Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer, 2019.

David E. Losada, Fabio Crestani, and Javier Parapar. eRisk 2020: Self-harm and depression challenges. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells,

Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 557–563. Springer, 2020. doi: 10.1007/978-3-030-45442-5_72. URL https://doi.org/10.1007/978-3-030-45442-5_72.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. Ontology-Aware Clinical Abstractive Summarization. In *Proceedings of SIGIR*, pages 1013–1016. Association for Computing Machinery, Inc, may 2019. URL http://arxiv.org/abs/1905.05818.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.7. URL https://www.aclweb.org/anthology/2021.clpsych-1.7.

Gregory Makoul, Patrick Brunett, Thomas Campbell, Kathleen Cole-Kelly, Deborah Danoff, Robert Frymier, Michael Goldstein, Geoffrey Gordon, Daniel Klass, Suzanne Kurtz, Jack Laidlaw, Forrest Lang, Anne-Marie MacLellan, Steven Miller, Dennis Novack, Elizabeth Rider, Frank Simon, David Sluyter, Susan Swing, and Gerald Whelan. Essential elements of communication in medical encounters: The kalamazoo consensus statement. *Academic Medicine*, 76:390–393, 04 2001. doi: 10.1097/00001888-200104000-00021.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems 10, [NIPS Conference, Denver, Colorado, USA, 1997]*, pages 570–576. The MIT Press, 1997. URL http://papers.nips.cc/paper/1346-a-framework-for-multiple-instance-learning.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243, 2019.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://www.aclweb.org/anthology/2020.acl-main.173.

Kirsten McKenzie, Susan Walker, Claire Dixon-Lee, Gareth Dear, and Judy Moran-Fuke. Clinical coding internationally: a comparison of the coding workforce in australia, america, canada and england. In *The 14th International Federation of Health Records Organizations (IFHRO) Congress and the 76th AHIMA National Convention Proceedings*, pages 52–64. American Health Information Management Association, 2004.

Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics*, 17 (1):1–11, 2016.

David N. Milne. Triaging content in online peer-support: an overview of the 2017 CLPsych shared task, 2017. Available online at http://clpsych.org/shared-task-2017.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. CLPsych 2016 shared task: Triaging content in online peer-support forums. In Kristy Hollingshead and Lyle H. Ungar, editors, *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016*, pages 118–127. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/w16-0312. URL https://doi.org/10.18653/v1/w16-0312.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20, 2015.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25 – 37, 2016. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2016.01.003. URL http://www.sciencedirect.com/science/article/pii/S0933365716000051.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 1101–1111. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1100. URL https://doi.org/10.18653/v1/n18-1100.

Harvey J Murff, Vimla L Patel, George Hripcsak, and David W Bates. Detecting adverse events for patient safety research: a review of current methodologies. *Journal of biomedical informatics*, 36(1-2):131–143, 2003.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 97–109, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1010. URL https://www.aclweb.org/anthology/P18-1010.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 2727–2733, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.235.

Ani Nenkova and Rebecca J Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152, 2004.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL https://www.aclweb.org/anthology/W19-5034.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Kevin A Padrez, Lyle Ungar, Hansen Andrew Schwartz, Robert J Smith, Shawndra Hill, Tadas Antanavicius, Dana M Brown, Patrick Crutchley, David A Asch, and Raina M Merchant. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Qual Saf*, pages bmjqs–2015, 2015.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118693.1118704. URL https://doi.org/10.3115/1118693.1118704.

Rebecca J Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, 2014.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Umashanthi Pavalanathan and Munmun De Choudhury. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 315–321. ACM, 2015.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543. ACL, 2014. doi: 10.3115/v1/d14-1162. URL https://doi.org/10.3115/v1/d14-1162.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, 2013.

John P. Pestian, Chris Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *BioNLP@ACL*, 2007.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Pedro H. O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1713–1721. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298780. URL https://doi.org/10.1109/CVPR.2015.7298780.

Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

Plato. *Phaedrus 275a*. translated by Reginald Hackforth, Cambridge University Press, 370 BCE, trans. 1972.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343, 2016.

Dina Popovic, Eduard Vieta, Jean-Michel Azorin, Jules Angst, Charles L Bowden, Sergey Mosolov, Allan H Young, and Giulio Perugi. Suicide attempts in major depressive episode: evidence from the bridge-ii-mix study. *Bipolar disorders*, 17(7):795–803, 2015.

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

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29:3567, 2016.

Reddit. Reddit privacy policy, March 2018. Downloaded March 22, 2018, https://www.reddit.com/help/privacypolicy/.

Philip Resnik. The (in)ability to triangulate in data driven healthcare research. In Holly G. Rhodes, editor, *Changing Sociocultural Dynamics and Implications for National Security: Proceedings of the SBS Decadal Survey Workshop on Culture, Language, and Behavior*, Washington, DC, 2018. The National Academies Press, National Academies of Sciences, Engineering, and Medicine. ISBN 978-0-309-47377-4. doi: 10.17226/25056.

Philip Resnik, Michael Niv, Michael Nossal, Gregory Schnitzer, Jean Stoner, Andrew Kapit, and Richard Toren. Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding. In *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings*, pages 2006–2006, 2006.

Philip Resnik, Rebecca Resnik, and Margaret Mitchell, editors. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014. URL http://www.aclweb.org/anthology/W/W14/W14-32.

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 2020. doi: 10.1111/sltb.12674. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sltb.12674.

Alexander M Rush, SEAS Harvard, Sumit Chopra, and Jason Weston. A neural attention model for sentence summarization. In *ACLWeb. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2017.

Naomi Sager, Carol Friedman, and Margaret S Lyman. *Medical language processing: computer management of narrative data*. Addison-Wesley Longman Publishing Co., Inc., 1987.

Tetsuya Sakai. Graded relevance assessments and graded relevance measures of NTCIR: A survey of the first twenty years. *CoRR*, abs/1903.11272, 2019. URL http://arxiv.org/abs/1903.11272.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

Natalie Schluter. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, 2017.

Allison Schuck, Raffaella Calati, Shira Barzilay, Sarah Bloch-Elkouby, and Igor Galynker. Suicide Crisis Syndrome: A review of supporting evidence for a new

suicide-specific diagnosis. *Behavioral sciences & the law*, 37(3):223–239, 2019. doi: 10.1002/bsl.2397.

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL*, pages 1073–1083, 2017.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1289–1296. Curran Associates, Inc., 2007. URL http://papers.nips.cc/paper/3252-multiple-instance-active-learning.

Tait D Shanafelt, Lotte N Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P West. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. In *Mayo Clinic Proceedings*, volume 91, pages 836–848. Elsevier, 2016.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*, 2017.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik, and Philip Resnik, editors, *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych@NAACL-HTL*, pages 25–36. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-0603. URL https://doi.org/10.18653/v1/w18-0603.

Han-Chin Shing, Guoli Wang, and Philip Resnik. Assigning medical codes at the encounter level by paying attention to documents. In *ML4H, Machine Learning for Health Workshop at NeurIPS*, 2019. URL https://arxiv.org/abs/1911.06848.

Han-Chin Shing, Philip Resnik, and Douglas W Oard. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137, 2020.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *arXiv preprint arXiv:2104.13498*, 2021.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, 2014.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016. doi: 10.7326/M16-0961. URL https://www.acpjournals.org/doi/abs/10.7326/M16-0961. PMID: 27595430.

Mark D. Smucker and Charles L. A. Clarke. Time-based calibration of effectiveness measures. In William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12*, pages 95–104. ACM, 2012. doi: 10.1145/2348283.2348300. URL https://doi.org/10.1145/2348283.2348300.

Irena Spasic and Goran Nenadic. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984, 2020.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651, 2010.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1139–III–1147. JMLR.org, 2013. URL http://dl.acm.org/citation.cfm?id=3042817.3043064.

Daniel S Tawfik, Jochen Profit, Timothy I Morgenthaler, Daniel V Satele, Christine A Sinsky, Liselotte N Dyrbye, Michael A Tutty, Colin P West, and Tait D Shanafelt. Physician burnout, well-being, and work unit safety grades in relationship to reported medical errors. In *Mayo Clinic Proceedings*, volume 93, pages 1571–1580. Elsevier, 2018.

Cornelis J Van Rijsbergen. Information retrieval. (2nd ed.). *University of Glasgow*, pages 133–134, 1979.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

M Johnson Vioulès, Bilel Moulahi, Jérôme Azé, and Sandra Bringay. Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development*, 62(1):7–1, 2018.

Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001. doi: 10.1007/3-540-45691-0_34. URL https://doi.org/10.1007/3-540-45691-0_34.

Byron C. Wallace. Thoughts on "attention is not not explanation". https://medium.com/@byron.wallace/thoughts-on-attention-is-not-not-explanation-b7799c4c3b24, August 2019. Medium, Accessed: December, 2019.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://www.aclweb.org/anthology/2020.acl-main.450.

Sida Wang and Christopher D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 90–94, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390665.2390688.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34 – 49, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.11.011. URL http://www.sciencedirect.com/science/article/pii/S1532046417302563.

Yanshan Wang, Ahmad Tafti, Sunghwan Sohn, and Rui Zhang. Applications of natural language processing in clinical research and practice. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 22–25, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5006. URL https://www.aclweb.org/anthology/N19-5006.

Michael Weiner and Paul Biondich. The influence of information technology on patient-physician relationships. *Journal of general internal medicine*, 21(1):35–39, 2006.

W Pete Welch, Steven J Katz, and Stephen Zuckerman. Physician fee levels: Medicare versus canada. *Health Care Financing Review*, 14(3):41, 1993.

C. P. West, L. N. Dyrbye, and T. D. Shanafelt. Physician burnout: contributors, consequences and solutions. *Journal of Internal Medicine*, 283(6):516–529, 2018. doi: 10.1111/joim.12752. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/joim.12752.

Hadley Wickham et al. The split-apply-combine strategy for data analysis. *Journal of statistical software*, 40(1):1–29, 2011.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002. URL https://doi.org/10.18653/v1/D19-1002.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774, 2014.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, 2020.

Pengtao Xie and Eric Xing. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, 2018.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. Deep learning architecture for patient data de-identification in clinical records. In *Proceedings of the clinical natural language processing workshop (ClinicalNLP)*, pages 32–41, 2016.

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6, 2017.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1174. URL https://doi.org/10.18653/v1/n16-1174.

Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, sep 2017. doi: 10.18653/v1/D17-1322. URL http://arxiv.org/abs/1709.01848http://aclweb.org/anthology/D17-1322.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2020.

Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1–9, 2006.

Danchen Zhang, Daqing He, Sanqiang Zhao, and Lei Li. Enhancing automatic icd-9-cm code assignment for medical texts with pubmed. In *BioNLP*, 2017.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. Learning to Summarize Radiology Findings. In *Proceedings of the Workshop on Health Text Mining and Information Analysis (EMNLP-LOUHI)*, pages 204–213. Association for Computational Linguistics (ACL), 2018. URL http://arxiv.org/abs/1809.04698.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.458. URL https://www.aclweb.org/anthology/2020.acl-main.458.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In Maxime Descoteaux, Lena Maier-Hein, Alfred M. Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III*, volume 10435 of *Lecture Notes in Computer Science*, pages 603–611. Springer, 2017. doi: 10.1007/978-3-319-66179-7_69. URL https://doi.org/10.1007/978-3-319-66179-7_69.

Michael Zimmer. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325, dec 2010. ISSN 1388-1957. doi: 10.1007/s10676-010-9227-5. URL http://link.springer.com/10.1007/s10676-010-9227-5.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/W19-3003. URL https://www.aclweb.org/anthology/W19-3003.