University of Nottingham
UK | CHINA | MALAYSIA

Deep Learning-based Natural Language Processing Techniques for Smart Healthcare

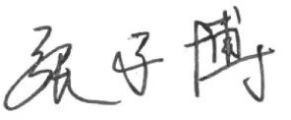
Thesis submitted to the University of Nottingham for the degree of **Doctor of Philosophy, Oct 2024.**

Zibo Zhang

20219263

Supervised by

Zheng Lu
Ruibin Bai
Tieyan Liu

Signature 

Date 2025/05/24

Abstract

Natural Language Processing (NLP) is one of the most essential technologies for smart healthcare. In recent years, deep learning-based NLP techniques have gathered significant attention. Despite promising results, existing deep learning techniques remain limited due to challenges including the variability and complexity of medical language, the difficulty of integrating external medical knowledge, and the gap between patient and healthcare provider's way of speaking. These challenges lead to issues in healthcare applications. This thesis aims to leverage deep learning-based NLP techniques towards smart healthcare by addressing these challenges. Specifically, a novel classification framework is first proposed to categorize chief complaints from patients' text, leveraging hierarchical clinical department label information to improve classification performance. Second, a medical dialogue generation framework is introduced, modeling patients and doctors separately and integrating external knowledge to generate contextually appropriate patient-doctor conversations. Third, a Rule-Enriched Attention-Based Deep Neural Network is devised to categorize physician responses into distinct social support types, supported by the development of the first dedicated social support lexicon for team-based teleconsultation, improving the quality of online consultations. Finally, a prompt-based Named Entity Recognition (NER) framework is developed to better capture medical entities in clinical text, overcoming challenges posed by

complex medical terminology and limited annotated data.

Acknowledgements

The doctoral journey has been rigorous yet deeply rewarding, greatly enhancing my knowledge and personal growth over these five years. This journey has cultivated in me a calm resilience to embrace imperfections, the courage to make deliberate moves, and the insight that the essence of swift learning lies in the quiet dedication of one's efforts. I am immensely thankful to all who have walked this path with me, offering their constant support and unwavering encouragement during times of challenge and triumph alike.

First and foremost, I extend my heartfelt appreciation to my principal supervisor, Dr. Zheng Lu, for generously sharing his extensive academic knowledge and playing a pivotal role in helping me establish myself within the academic community. Our meetings and conversations were vital in inspiring me to think outside the box, considering multiple perspectives to form a comprehensive and objective critique. I am equally grateful to my second supervisor, Prof. Ruibin Bai, for his invaluable insights and guidance that have enriched my research journey.

I am incredibly grateful for my lab mates at the University of Nottingham, including Chang Shu, Xinyu Gu, Jialu Zhang, Yiran Li, Menglin Cui, Jiandong Liu, Ruxin Ding, Shihe Wang, Chenglin Yao, Xingke Song, Wentao He, and many others. Their camaraderie and collaboration have made this experience all the more enjoyable and fulfilling.

Lastly, I would like to express my heartfelt gratitude to my parents for their unconditional love and unwavering support throughout my academic endeavors. They have always cared for my happiness and well-being, serving as exceptional role models of generosity and kindness.

Contents

List of Tables

List of Figures

Abbreviations

AI artificial intelligence.

BERT Bidirectional Encoder Representations from Transformers.

Bi-LSTM Bi-directional Long Short-Term Memory.

Bi-GRU Bidirectional Gated Recurrent Unit.

BOW Bag-of-Words.

CNER Clinical Named Entity Recognition.

CNN Convolutional Neural Network.

CRF Conditional Random Fields.

CVAE Conditional Variational Autoencoder.

CWAE Conditional Wasserstein Auto-Encoder.

DGAN Dictionary-guided Attention Network.

EMR Electronic Healthcare Records.

GNN Graph Neural Network.

HMM Hidden Markov Models.

HRNA Hierarchical Relational Network with Attention.

IDF Inverse Document Frequency.

LDA Latent Dirichlet Allocation.

LLM Large Language Model.

LSTM Long Short-Term Memory.

MLP Multi-layer Perceptron.

NER Named Entity Recognition.

NLP Natural Language Processing.

PIQN Parallel Instance Query Network.

R-ADNN Rules-enriched Attention-based Deep Neural Network.

RNN Recurrent Neural Network.

SIE Sequence Information Encoder.

SoTA State-of-the-Art.

SVM Support Vector Machine.

TF Term Frequency.

UMLS Unified Medical Language System.

WAE Wasserstein Autoencoder.

Chapter 1

Introduction

1.1 Background

The surge in digital communication has led to a significant increase in text data production, necessitating advanced methods for handling and analyzing this information. Over time, the field of Natural Language Processing (NLP) has developed various techniques to manage this growing volume of data. As a crucial aspect of artificial intelligence (AI), Natural Language Processing (NLP) aims to enable computers to understand, interpret, and generate human language. Originally rooted in computational linguistics, NLP has transitioned from early rule-based and statistical approaches to modern, deep learning-driven methods. Deep learning, a subset of machine learning, utilizes sophisticated neural networks to model complex patterns in data. The rapid evolution of deep learning techniques has revolutionized NLP, significantly enhancing the accuracy and efficiency of language processing tasks (Nagarhalli et al., 2021; Lauriola et al., 2022). This progress has profoundly impacted various sectors, especially in addressing challenges and improving outcomes in healthcare (Velupillai et al., 2018; Roy et al.,

2021).

Smart healthcare refers to the integration of advanced technologies, such as AI, big data analytics, the Internet of Things (IoT), and telemedicine, into the healthcare system to enhance patient care, improve outcomes, and streamline operations. Smart healthcare aims to create a more efficient, patient-centered system that leverages technology to enhance both clinical and operational processes. Traditional healthcare applications relied on in-person consultations and manual record-keeping, which often limited accessibility and delayed information retrieval. However, with the rapid growth of online digital tools, smart healthcare has emerged as an essential resource to improve patient care and optimizing operations. Such applications offer a range of functionalities, such as facilitating patient registration, managing appointments, and providing resources for clinical decision-making. For instance, patients can register online, access their medical records, and receive reminders for upcoming appointments, which significantly enhances the overall patient experience.

Despite these advancements, many processes within modern smart healthcare applications still require substantial manual effort, which can reduce efficiency and hinder patient care (Shamshirband et al., 2021). While NLP techniques have shown promises in optimizing these platforms, there remain challenges that need to be addressed. For example, although NLP can automate clinical documentation, the integration of these techniques is not yet seamless, and certain tasks may still necessitate human intervention. Additionally, NLP-driven chatbots, while effective in assisting with patient triage, may not fully capture the nuances of patient inquiries or provide comprehensive support. Therefore, automating these processes holds the promise of improving operational efficiency and enhancing patient outcomes, but significant advancements in NLP are still required to realize

this potential.

This thesis focuses on developing advanced NLP techniques for smart healthcare, ensuring that interactions between patients and healthcare providers are seamless, accurate, and timely. One of the primary objectives of this research is to facilitate the automation of patient interactions and the processing of clinical information, which are integral to smart healthcare. By leveraging deep learning-based NLP techniques, this thesis seeks to address several critical processes in healthcare, from classifying patient records to simulating doctor-patient conversations and extracting valuable information from clinical data.

1.2 Challenges

Despite the promising advancement in NLP, several challenges hinder its effective implementation in healthcare. One significant issue is the variability and complexity of medical language, which often includes numerous abbreviations, synonyms, and context-dependent phrases that can lead to misunderstandings. This linguistic variability complicates the training of NLP models, requiring them to accurately recognize and interpret a wide range of expressions. Additionally, effectively utilizing external medical knowledge, such as clinical guidelines and medical literature, is vital for understanding clinical text contextually; however, integrating this knowledge into NLP necessitates sophisticated mechanisms for real-time applicability. Another challenge lies in the differences in language use between patients and healthcare providers, where patients often communicate in informal language or layman's terms, while healthcare professionals typically employ more technical terminology. This disparity can create barriers in

communication and complicate the NLP model's ability to interpret and respond accurately. Addressing these challenges, particularly the integration of external knowledge and understanding the nuances of medical language, is essential for fully realizing the potential of NLP in enhancing patient care and operational efficiency within healthcare systems.

1.3 Aims and Motivations

The primary research problem addressed in this thesis is how to leverage the advanced deep learning-based NLP techniques towards smart healthcare while overcoming existing challenges. The thesis consists of four distinct but interrelated studies, including chief complaint text classification, medical dialogue generation, physician dialogue text classification, and medical named entity recognition. Each study tackles different challenges mentioned above. Figure 1.1 illustrates the relationships between the challenges and the four studies. Collectively, these studies aim to contribute to smart healthcare applications by streamlining processes such as clinical documentation and patient consultations. The motivations and objectives for each study are integral to the thesis, focusing on automating and advancing crucial stages in the modern healthcare system. Specifically, the

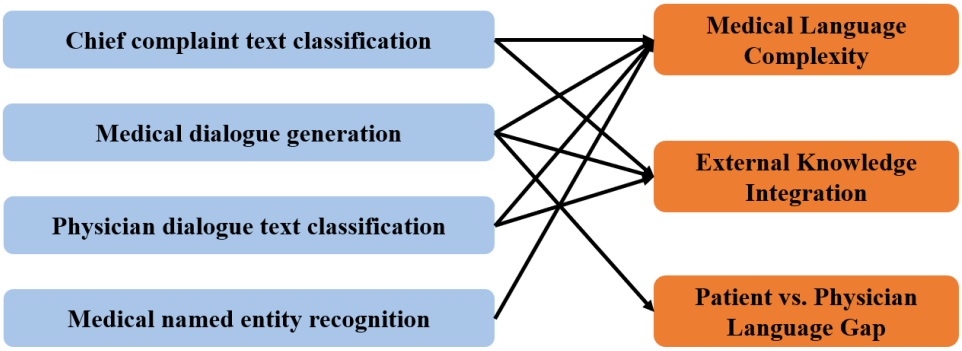


Figure 1.1: Relationships between the four studies and challenges.

main research objectives are formed as follows:

1. **To investigate a solution for accurately classifying chief complaint text into specific clinical departments by leveraging label information:** A prime aspect in online healthcare applications, this task involves interpreting patients' non-medical description of symptoms. With the growing use of online health portals, patients frequently express their symptoms using informal and ambiguous language, leading to challenges in classification. This text often lacks precision due to non-standard terms and may include redundant information or various expressions for similar symptoms. Developing a solution that can effectively categorize these unstructured descriptions is crucial. Moreover, utilizing label information—such as clinical departments' names and their descriptions—can significantly enhance classification accuracy, yet many existing approaches do not fully leverage this important information.

2. **To devise a solution that can generate contextually appropriate medical dialogues:** Another prime aspect of the online healthcare applications, this task aims to facilitate meaningful and contextually relevant interactions between patients and healthcare providers. The challenges include accurately understanding and generating medical terminology while remaining comprehensible to patients, as well as recognizing the nuances of patient-doctor conversations to produce contextually appropriate responses. Additionally, integrating external knowledge sources, such as clinical guidelines and medical databases, enhances the model's ability to generate informed and relevant dialogues.

3. **To investigate a solution that can classify physician text into corresponding communication forms:** The aim is to accurately classify physician responses into distinct categories, including direct

informational support, indirect informational support, and emotional support. Direct informational support involves providing professional advice on diagnosis and treatment, while indirect informational support refers to guiding patients to external resources, such as medical websites or referrals to other specialists. Emotional support reflects the physician's role in offering comfort, expressing care, or providing encouragement. By analyzing and categorizing these communication forms, the objective is to gain insights into how different strategies influence patient outcomes, thereby enhancing the quality and effectiveness of online consultations.

4. **To investigate a solution for extracting medical-related terms in clinical text:** Extracting key medical entities, such as diseases, medications, and procedures, is essential for structuring unstructured medical text in the healthcare applications. The aim is to accurately position and categorize these entities from medical data. With the growing volume of electronic health records, making this data more accessible and actionable is crucial for healthcare providers. More accurate medical entity recognition can support better clinical decision-making and ensure that relevant medical information is readily available to professionals.

1.4 Contributions

The main contributions of this thesis are outlined as follows:

The first contribution is casting the medical chief complaint classification problem as a multi-class classification task with a hierarchical structure of label descriptions. A novel deep learning-based framework has been devel-

oped, featuring a Sequence Information Encoder that utilizes a pre-trained BERT model for contextual embedding and a Bi-LSTM to effectively encode sequential information. Additionally, a Hierarchical Relational Network with an Attention module is proposed, capable of capturing complex relationships among the chief complaint text and hierarchical category descriptions. Experimental results demonstrate that this model outperforms state-of-the-art methods on real-world public medical datasets, illustrating the effectiveness of explicitly modeling and leveraging the hierarchical relationships among chief complaint label descriptions through the proposed Hierarchical Relational Network with Attention.

The second contribution is to devise a multi-turn dialogue generation framework consisting of two separate models representing patient and doctor roles, connected through a memory mechanism. A Knowledge-based Conditional Wasserstein Auto-Encoder is designed to effectively integrate dialogue history and external medical knowledge, ensuring the generated responses are both contextually accurate and medically relevant while enhancing the diversity of the outputs. The proposed framework is evaluated using two real-world medical dialogue datasets, demonstrating superior performance compared to existing baseline models.

The third contribution is to propose the R-ADNN, a hybrid network that combines rule-based technique with deep learning to classify medical text provided by physicians during teleconsultations to social support types. This framework incorporates a custom-built lexicon for social support in the teleconsultation setting, which to the best of our knowledge is the first of its kind. The R-ADNN categorizes physician responses into types such as direct informational support, indirect informational support, and emotional support. It utilizes a combination of 37 domain-specific rules and a deep-learning model that incorporates contextual information through

BERT embeddings, Bi-LSTM, and attention mechanisms. The R-ADNN demonstrates superior performance against state-of-the-art text classification models on real-world teleconsultation datasets, offering valuable insights into patient-physician communication.

The fourth contribution is to devsie a novel medical named entity recognition framework that integrates prompt learning into pre-trained deep learning models to tackle the complexities of medical entities and the challenges posed by limited annotated data. It introduces a prompt position predictor and a prompt type predictor with a relational network, designed to enhance the prediction of start and end indices as well as the types of recognized entities, by effectively capturing the relationships between prompts and medical text. The effectiveness of the proposed framework is validated through evaluations on a real-world medical dataset, demonstrating significant performance improvements over existing baseline models.

1.5 Thesis Outline

This thesis is structured as follows. Chapter 2 presents an in-depth review of the essential background information related to the research topics covered in this thesis. Chapter 3 details the development of a novel technique to classify chief complaints from patients' free-text description. In Chapter 4, the focus shifts to medical dialogue generation, exploring the challenges and proposing a solution to generate contextually aware medical conversations. Chapter 5 introduces a framework for social support type classification. Chapter 6 applies advanced NER techniques to clinical narratives, showcasing their effectiveness in extracting and categorizing critical medical entities. Finally, Chapter 7 concludes the thesis by pro-

viding a summary of the key findings. It also addresses the limitations encountered during the research and outlining potential future work.

Chapter 2

Literature Review

This chapter provides an overview of the essential background knowledge related to the thesis. It focuses on six key areas: general text classification, chief complaint text classification, general dialogue generation, medical dialogue generation, general named entity recognition and medical named entity recognition. For each area, various approaches and frameworks are introduced, laying the groundwork for a deeper understanding of the subsequent analyses and methodologies presented in the thesis.

2.1 General Text Classification

For the general text classification problem, pre-trained word embedding models and deep learning-based neural networks are widely used. Chalkidis et al. (2019) apply several novel classification models in the legal domain for a multi-label text classification task. Four main models are tried in their experiments, Bidirectional Gated Recurrent Unit (Bi-GRU) with an attention layer, Hierarchical Attention Network (Yang et al., 2016), Label-wise Attention Network with Bi-GRU encoder (Mullenbach et al., 2018), and

Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). The best result is produced by the Label-wise Attention Network with Bi-GRU encoder. Li et al. (2019c) improves the Hierarchical Recurrent Neural Network by adding a dual attention layer. The outputs of attention layer are then sent to a Bi-GRU layer and finally go through a Conditional Random Fields (CRF) layer (Chen et al., 2018) to calculate the final tag. Choi et al. (2019) use a Bi-GRU-based neural network to create a filter-generating network which can automatically generate filters for Convolutional Neural Network (CNN) layer used for classification. Huang et al. (2019) use Graph Neural Network (GNN) to classify a corpus. A text-level GNN is created to replace traditional corpus-level GNN, which performs with 2% improvement. Ohashi et al. (2020) propose a negative supervision method to solve the problem that pre-trained text representation models often incorrectly classifying sentences with different labels but having similar semantics. There are several works making use of label information to help classification. Kim et al. (2019) present a new study on text classification, where the authors not only use input text sequence but also metadata about labels to predict documents. Pappas and Henderson (2019) use two different encoders to separately encode input text and label, whose embeddings are concatenated as final features for classification. In terms of hierarchical labels, Banerjee et al. (2019) and Shimura et al. (2018) use transfer learning-based methods to tackle the task. The authors firstly train a neural network to classify input text into main categories and then transfer the model to classify text into sub categories. Zhang et al. (2022) propose a label-based attention neural network to capture attentional information among labels and text. Ma et al. (2022) propose a hybrid embedding method which use different ways to embed the original text, label name and the structure of hierarchical label. Wang et al. (2022) utilize contrastive learning to integrate hierarchical label information into

BERT encoding. Miyazaki et al. (2019) apply a hierarchical label embedding method to classify Twitter entries. In their work, Twitter text and label information are encoded before applying label-wise attention. While showing promising results, these methods do not fully exploit information embedded in hierarchical labels and relationships among different level of labels and text. For example, most works only embed those few words in label itself instead of longer descriptions about labels.

2.2 Chief Complaint Classification

Chief complaint classification methods can be briefly divided into two categories, rule-based and machine learning-based methods. Among rule-based methods, Lu et al. (2008) propose an ontology-based technique utilizing semantic relations in medical document to classify chief complaint into corresponding classes. Mikosz et al. (2004) create a set of keywords that is used to match a given chief complaint with its corresponding symptom. Similarly, Chapman et al. (2005c) also create a standard reference set to classify various chief complaints. Hsu et al. (2020) make use of Bag-of-Words approach to represent chief compliant with hand-crafted features created by professional physicians. Travers and Haas (2004) match chief complaint with Unified Medical Language System (UMLS) term to obtain its label after going through a series of preprocessing. Cui et al. (2019) propose a constructive heuristic method to generate regular expressions to classify chief complaint text. Similarly, Liu et al. (2020a) use genetic programming to automatically construct and evolve regular expressions for text classification.

Over the past decades, machine learning-based methods have become the

mainstream technique to tackle the chief complaint classification problem. Popular techniques include n-gram models (Brown et al., 2010), Support Vector Machine (SVM), Naive Bayes classifiers (Li et al., 2019a; Jernite et al., 2013; Chapman et al., 2005b), etc. Recently, deep learning-based models including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have played a major role in encoding chief complaint text before feeding to classifiers for better performance (Sulieman et al., 2017; Lee et al., 2019; Blanco et al., 2019; Li et al., 2021c). Among those deep learning-based techniques, the state-of-the-art deep learning-based language representation model, BERT (Devlin et al., 2018), is the most popular one used in NLP tasks including chief complaint classification (Chang et al., 2020; Valmianski et al., 2019; Schäfer et al., 2020). Although there are many methods tackling the chief complaint classification problem, most of those methods do not utilize the hierarchical structure of labels. In Chapter 3, the proposed method attempts to make best of use the hierarchical label descriptions from expert knowledge and shows performance improvement when the hierarchical label information is used.

2.3 General Dialogue Generation

The encoder-decoder framework is a commonly employed approach in dialogue generation. In their work, Shang et al. (2015) present a Neural Responding Machine, utilizing a Recurrent Neural Network as both encoder and decoder for short-text conversations. Gu et al. (2016) introduce CopyNet, incorporating a copying mechanism into the RNN-based encoder-decoder framework. To enhance response diversity, Li et al. (2016) apply diversity promotion objectives. Serban et al. (2016) propose a hierarchical recurrent encoder-decoder framework, named HRED, using an utterance

encoder to encode the input utterances and then using a context encoder to capture the sequential relationship between context turns. In order to further capture the intricate relationship between the input context sequences, Serban et al. (2017) propose a VHRED model that incorporates the Conditional Variational Autoencoder (CVAE) to the HRED model. Building on the HRED model, VHRED adds a latent variable into the decoder and transforms the decoding process into a two-stage generation process. The first stage involves sampling a latent variable, and the second stage involves generating the response based on this variable. The VHRED model increases the quality and diversity of the generated responses compared with the HRED model. One of the primary obstacles faced by CVAE-based models is the "posterior collapse" problem. Zhao et al. (2017) introduce a supplementary "bag-of-words" loss to the decoder to alleviate this issue. Shen et al. (2018) also propose a collaborative CVAE model. The latent variable in this model is sampled by transforming random Gaussian noise using multi-layer perceptrons. Park et al. (2018) add a hierarchical latent variable structure to the VHRED model. Gu et al. (2019) propose the DialogWAE model which takes a different approach from VAEs in modeling the distribution of data. It trains a Wasserstein-GAN to minimize the Wasserstein distance between the prior and posterior distributions. Their model performs well on the DailyDialog dataset and Switchboard dataset compared with the above VAE-based dialogue models. Zhang et al. (2019) combine the self-attention mechanism with the encoder-decoder framework, which can better capture the utterances' distant dependencies in the context. Zhang et al. (2020a) propose a co-attention mechanism to capture the relationships between context and response and utilize it when generating the latent variable. Liu et al. (2021b) design a CVAE-based model with affective information to generate affective responses. Li et al. (2022b) combine the VAE model with a hierarchical contrastive learning mechanism

that can capture different levels of semantic meaning in the input context.

For the pre-trained language models, Zhang et al. (2020b) propose DialogGPT, employing GPT2 (Radford et al., 2019) as the foundational structure for generating dialogues. DialogBERT, introduced by Gu et al. (2021), adopts a hierarchical transformer architecture based on BERT (Devlin et al., 2018). Although these pre-trained language model show impressive results in dialogue generation task (Yang et al., 2019; Chen et al., 2022; Thoppilan et al., 2022), their generated responses in the medical field can often be ambiguous and contextually unclear, potentially leading to incorrect diagnoses. Furthermore, due to their extensive training on massive datasets, these large language models exhibit high sensitivity to the specific requests submitted (Caruccio et al., 2024). In chapter 4, a dialogue generation model is introduced to emphasize traditional machine learning methods, avoiding reliance on large language model-based architectures. Similar to the work of Gu et al. (2019), the model incorporates Wasserstein autoencoders. However, the existing approach does not fully utilize role information, as it only appends a binary role ID to the end of the context embedding, which proves insufficient. To better model the patient and doctor roles in medical dialogue generation, two separate Wasserstein autoencoder-based dialogue models are employed to represent each role in this thesis. These models are interconnected by a memory mechanism and trained recurrently until a complete dialogue is generated. Additionally, a medical guidance book is used as supplementary information to enhance the generation of medically relevant sentences.

2.4 Medical Dialogue Generation

Machine learning-based methods are widely adopted in medical dialogue generation. Liu et al. (2016a) propose a Long Short-Term Memory (LSTM)-based framework to identify the patients' input symptoms. Then the model generates the symptom-related questions. Wei et al. (2018) use a reinforcement learning method to create an automatic diagnosis system. Xu et al. (2019b) propose an end-to-end dialogue system that uses medical knowledge graphs to enhance the topic transition when generating the response. Zeng et al. (2020) prepare a large medical dialogue dataset which contains millions of dialogues. The authsos use the large medical corpus to pre-train a large language model and then tested the model on a COVID-19 dialogue dataset (Zhou et al., 2021). Li et al. (2021a) develop a system that firstly predicts the entities in a dialogue history and then utilizes the entities to help solve dialogue generation tasks. Li et al. (2021b) develope a system that predicts entities present in a conversation's history, and then uses these entities to aid in generating medical dialogues. Yan et al. (2022) apply a contrastive learning method on several pre-trained language models to evaluate their proposed new ReMeDi dataset. Most Medical dialogue generation models generally relied on additional labeled information, such as the patient's condition, the type of entity involved, and the actions taken by the doctor. This information necessitates a large number of annotations by humans. The model introduced in chapter 4 addresses this by utilizing Wasserstein autoencoders (Wasserstein Autoencoder (WAE)) to capture additional information within narrative dialogue text without the need for extensive human annotations. Two language models with the same architecture are used to represent the distinct roles of patients and doctors, enabling the model to better capture role-specific information in the conversation.

2.5 General Named Entity Recognition

Early Named Entity Recognition (NER methods primarily relied on rule-based systems, which depended heavily on manually crafted linguistic rules and domain-specific expertise. Although effective in controlled environments, these systems struggled with scalability and adaptability across different languages and domains (Farmakiotou et al., 2000; Abdallah et al., 2012; Petasis et al., 2001). The field then transitioned to machine learning approaches, where models like Hidden Markov Models (HMM) and Conditional Random Fields (CRF) improve performance by better managing sequential data and integrating contextual information. However, these approaches still require significant feature engineering (Zhou and Su, 2002; Fu and Luke, 2005; Lafferty et al., 2001; Liu et al., 2011).

The advent of deep learning revolutionized NER. RNN, particularly Long Short-Term Memory (LSTM) networks, significantly enhance the ability to capture long-range dependencies in text, leading to more accurate entity recognition. For instance, Graves et al. (2013) first propose a Bi-directional LSTM for NER tasks, which was later improved upon by Huang et al. (2015) through the integration of a CRF module. Furthermore, Dyer et al. (2015) introduce the Stack-LSTM, modifying the traditional LSTM architecture to derive hidden states from a stack, thus capturing information from various positions within the stack.

Recent advancements have seen the introduction of hybrid frameworks such as AMFF, proposed by Yang et al. (2020), which captures multi-level features using a combination of Bi-LSTM, CNN, and attention mechanisms, processing both local and global features before final sequence labeling with a BiLSTM-CRF network. Transformer-based models, exemplified by Vaswani et al. (2017), have further advanced NER by leveraging atten-

tion mechanisms to capture global dependencies, setting new benchmarks across multiple languages and domains. For example, Schweter and Baiter (2019) utilize BERT as a character representation model, while Wang et al. (2021) introduce a hierarchical tagging approach with BERT embeddings to enhance sub-optimal path identification. Shen et al. (2022) propose the Parallel Instance Query Network (PIQN), framing NER as a machine reading comprehension task using BERT to create learnable global queries. Additionally, Li et al. (2022a) introduce the W^{2}NER framework, which employs word-word relation classification to address multiple NER tasks simultaneously. Finally, Yan et al. (2023) present a model incorporating CNNs with BERT to capture spatial relations for nested NER, demonstrating superior performance on various datasets. Shen et al. (2023) explore prompt learning in NER, using a dual-slot prompt template with BERT to achieve strong performance, highlighting an emerging approach in the field.

2.6 Medical Named Entity Recognition

Medical Named Entity Recognition has evolved from its early reliance on rule-based methods, which used domain-specific resources and structured vocabularies to identify and classify medical entities such as diseases, symptoms, treatments, and medications. Early systems, often based on the Unified Medical Language System (UMLS) and other ontologies, provided high precision but struggled with scalability and variability in medical language Wang et al. (2008); Kang et al. (2013); Quimbaya et al. (2016); Eftimov et al. (2017). These limitations restricted their applicability in diverse and unstructured clinical texts Kundeti et al. (2016).

The transition to deep learning has significantly advanced medical NER. For instance, Luo et al. (2018) introduce an attention mechanism to improve token consistency and capture global document-level information in biomedical texts. Similarly, Pomares-Quimbaya et al. (2018) utilize a BiLSTM-CRF model for extracting key concepts from medical records. Wang et al. (2019) develope a Bi-LSTM-based neural network that integrates dictionary information to address medical NER tasks effectively. Jin et al. (2019) combine a knowledge graph with a BiLSTM-CRF model to perform NER in traditional Chinese medicine contexts. Furthermore, Li et al. (2020) leverage BERT with a BiLSTM-CRF layer for the Chinese Clinical Named Entity Recognition (CNER) task, enhancing performance by pre-training on Chinese clinical records. An et al. (2022) introduce a deep neural network model combining Bi-LSTM with a multi-head self-attention mechanism for CNER tasks. Most recently, Zhu et al. (2023) propose the Dictionary-guided Attention Network (DGAN), which enhances semantic understanding by aligning text with a biomedical dictionary and using optimized attention to focus on key medical concepts, while employing semi-supervised learning to manage unseen entities.

Chapter 3

Medical Chief Complaint Classification with Hierarchical Structure of Label Descriptions

After laying the theoretical groundwork in chapter 2 with a comprehensive review of the essential background and state-of-the-art techniques in NLP and healthcare automation, chapter 3 delves into the first practical task of this thesis: chief complaint text classification. As an integral component of online healthcare platforms, accurately categorizing patients' free-text symptom descriptions is crucial for streamlining patient care. This task poses unique challenges due to the variability and ambiguity of patient language, often lacking formal medical terms. A novel deep learning-based approach is presented that leverages a hierarchical chief complaint label structure and sequence information encoding to effectively address these challenges, marking the initial step in automating key processes within

modern healthcare systems.

3.1 Introduction

Text classification is an important task in the field of NLP. It has been widely used in healthcare domain such as helping doctors diagnose whether a patient has certain condition based on their Electronic Healthcare Records (EMR) (Garla et al., 2013; Avci and Turkoglu, 2009), providing medical assistant to online users through question and answering (Gupta et al., 2021), structuring narrative clinical notes by identifying relationships between two medical terms (Luo, 2017), classifying medical documents to pre-defined topic sets (Hughes et al., 2017; Saibene et al., 2021), extracting relevant words with specific biomedical information from unstructured clinical records (Zhu et al., 2023; Y. Mahajan and Rana, 2023; AlMahmoud and Hammo, 2024), etc.

Applying general text classification techniques to healthcare is particularly challenging due to various reasons. For example, medical text is often unstructured or semi-structured, and contains information that requires vast domain knowledge for proper understanding (Friedman and Johnson, 2006). There are also very diversified types of medical text, including narrative text from physicians that often contains acronyms and abbreviations for both common words and professional medical terms, narrative text from patients that is often informal and ambiguous, semi-structured text records that are generated by computer systems with large amount of numerical values and symbols. This chapter focuses on tackling the classification problem of *Chief Complaint*, a specific type of medical text provided by patients that contains narrative sentences describing symptoms, condi-

tions, previous diagnoses, questions, and more. Benefit from development of Internet Technology, many online healthcare systems are available to both physicians and patients including Internet hospitals, EMR systems, online healthcare community software, etc. Automatic chief complaint classification plays an important role in such systems as it could provide smart features such as triage or recommending doctors that are specialized in the medical category in which patients have potential conditions.

The task of chief complaint text classification has been widely studied in areas such as early disease detection and public health surveillance (Chapman et al., 2005a; Clifford et al., 2021). Unlike general medical text provided by physicians or generated by computer systems, chief complaint is usually written by patients in spoken language. This poses additional challenges to the classification task. On one hand, such text is less precise due to reasons including using informal or ambiguous words instead of medical terms, or giving out information without context. On the other hand, the text is less concise containing redundant information or using various forms of expressions. For instance, chief complaints under medical category "Gastroenterology" include, "I vomited and got a diarrhea last night and today I have a serious stomachache.", "I have a poor appetite and my stomach feels bloated.", "Feel pain around my belly button and there is flatulence in my upper abdomen.", etc. It can be observed that different words and various ways of expressions are used to describe the same symptom.

Many existing works use rule-based methods to classify chief complaint such as creating keyword sets or more complicated rules that heavily rely on domain knowledge from human experts. In order to alleviate this problem, machine learning techniques are introduced including n-gram model, Support Vector Machine (SVM), deep neural networks (Brown et al., 2010; Lee et al., 2019; Chang et al., 2020), etc. While producing promising re-

sults, most of these methods ignore a simple but very important fact, unlike other classification tasks such as object recognition in computer vision, labels for medical categories corresponding to important domain information that are strongly related to the words in the input text. Moreover, the labels of chief complaints have inherently hierarchical structure since their categories usually correspond to departments or specialized medical areas in healthcare systems. For instance, an upper level category (main category) "Internal medicine" contains several lower level categories (sub category) such as "Neurology", "Gastroenterology", "Respiratory medicine", "Endocrine", "Cardiovascular medicine", etc. Effectively utilizing this hierarchical structure of labels could further improve the performance of chief complaint classification.

This chapter proposes a novel text classification framework for chief complaint by embedding both the input chief complaint text and the hierarchical structure of label descriptions based on deep neural networks. There are three branches in the proposed framework, chief complaint branch, sub-category branch, and main-category branch. In the chief complaint branch, chief complaint text is firstly embedded by a Sequence Information Encoder (SIE) consists of pre-trained word embedding model, Bidirectional Encoder Representations from Transformers (BERT), to capture input semantics with contextual information and Bi-directional Long Short-Term Memory (Bi-LSTM) to further encode sequential information. Then a Hierarchical Relational Network with Attention (HRNA) module is devised to reason the complex relationships among chief complaint and hierarchical label descriptions focusing on informative words. The representations of such relationships are then fed to a Multi-layer Perceptron (MLP) for final classification. The proposed framework takes full advantage of the hierarchical structure of labels by capturing relationships among label descriptions ex-

tracted from expert knowledge and input chief complaint with attentional scores. Compared with conventional single-branch neural networks and the State-of-the-Art (SoTA) hierarchical structure label methods, the proposed method demonstrates significant performance improvement on two real-world public datasets.

The main contributions presented in this chapter can be summarized as follows:

1. The medical chief complaint classification problem is formulated as a multi-class classification problem with a hierarchical structure of chief complaint label descriptions, accompanied by a novel deep learning-based medical text classification framework with three branches.

2. A Sequence Information Encoder is introduced to effectively encode sequential information from input text by incorporating a pre-trained model, BERT, to embed input text with contextual information, along with a Bi-LSTM to further encode the sequential information.

3. A novel Hierarchical Relational Network with an Attention module is proposed, capable of capturing complex relationships among input chief complaint text and hierarchical chief complaint label descriptions, focusing on informative words.

4. The proposed model demonstrates superior performance compared to state-of-the-art (SoTA) models on two real-world public medical datasets, utilizing hierarchical chief complaint label descriptions extracted from medical books and websites to illustrate its capability and effectiveness in leveraging such information.

3.2 Methodology

3.2.1 Problem Definition

Formally the problem concerned by this chapter can be defined as follows: firstly, Firstly, a set of chief complaint text is defined as C, a main-category set as $M = \{m_1, m_2, ..., m_n\}$, and a sub-category set as $S = \{s_1^1, s_1^2, ..., s_1^{l_1}, ..., s_n^1, ..., s_n^{l_n}\}$, where n equals to the number of pre-defined main categories and l_n equals to the number of sub-categories belong to main-category m_n. The sub-category set belonging to main-category m_i is defined as S_i, where $i \in [1, n]$. Given each chief complaint text $c \in C$, main-category descriptions d_m for all $m \in M$, sub-category descriptions d_s for all $s \in S$, the task is to classify c to one of the sub categories $s \in S$. Note that in this problem each sub category has a corresponding main category and each main category has more than one sub categories. Inherently, the classification of chief complaint text falls in this hierarchical structure of labels. The proposed framework utilizes this structure to perform multi-class classification task.

3.2.2 Overview

Figure 3.1 illustrates the overall architecture of the proposed framework. There are three branches in the model, i.e. chief complaint branch, main-category branch, and sub-category branch. The input of the chief complaint branch is chief complaint text c to be classified and the outputs are the feature vectors containing the hidden representations of each word in c. The input of the main-category branch is one of the main-category descriptions d_m. The input of the sub-category branch is one of the sub-category

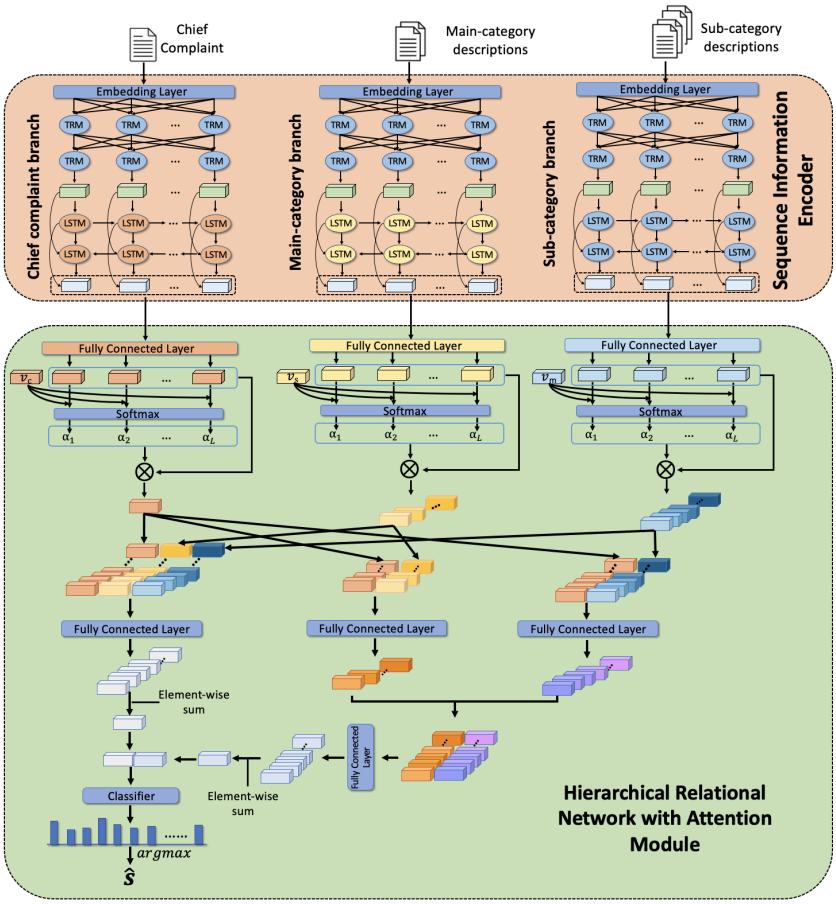


Figure 3.1: Overall architecture of the proposed framework.

descriptions d_s. Each branch contains a SIE module that is used to obtain the hidden representations of input. In a SIE module, the pre-trained BERT model is firstly incorporated to embed input text with contextual information and then a Bi-LSTM is used to further encode sequential information in the input. The HRNA module after these branches is used to capture the complex relationships among input chief complaint text and hierarchical category descriptions before the MLP for final classification. The attention mechanism in the HRNA module enables the module to pay more attention to those informative words in input sequences. The details of the SIE and HRNA modules are elaborated in Section 3.2.3 and Section 3.2.4.

The classification procedure for a given chief complaint text c works as

follows. Firstly, c is fed to the chief complaint branch to obtain its hidden representation $\tilde{h_c}$. Then for each possible sub category $s \in S$ with its corresponding main category $m \in M$, the corresponding descriptions (d_s and d_m respectively) are fed to the sub-category and main-category branches to obtain their hidden representations, denoted as $\tilde{h_{d_s}}$ and $\tilde{h_{d_m}}$. In this way, x pairs of category hidden representations ($x = |S|$) are obtained. After that, the hidden representation vector of chief complaint $\tilde{h_c}$ and each $\tilde{h_{d_s}}$, $\tilde{h_{d_m}}$ are fed into the HRNA to encode the relationships among the chief complaint, main-category, and sub-category information by utilizing their hierarchical structure. Finally the encoded relationships are forwarded to an MLP to predict the final category label.

3.2.3 Sequence Information Encoder

The objective of SIE is to encode input text to a feature vector representing its semantics. The popular BERT (Devlin et al., 2018) model is chosen as the main embedding mechanism. BERT is one of the state-of-the-art pre-trained language representation models that is able to capture contextual information and utilize them to represent words from input. BERT has been widely used in many downstream NLP tasks (Radford et al., 2018) while requiring minimal modification of its architecture, which can be briefly considered as a stack of transformer encoders (Vaswani et al., 2017) containing self-attention layer, normalization layer and feed-forward neural networks.

The transformer encoder firstly tokenizes input text into a sequence of tokens based on a large pre-defined dictionary, before being sent to an embedding layer which creates a large lookup matrix through learning. Each row of the lookup matrix represents the embedding vector of each

token. Similarly, BERT also embeds positional information of input tokens and segments. The final input sequence representation, the summation of word embedding, position embedding, and segment embedding, is then forwarded to a self-attention layer to emphasize on more informative words by looking at other words in the input sentence. The output of the attention layer is then normalized and fed to a feed forward neural network.

In this framework, the input sequence for the BERT model is constructed by adding a classifier token [CLS] at the beginning of each input text and a sentence separator token [SEP] at the end. For example, the hidden representation E_c for a given chief complaint text c is obtained as follows:

$$E_c = BERT([CLS], c_1, c_2, ..., c_t, ..., [SEP]), \tag{3.1}$$

where c_t denotes different tokens in the chief complaint text. The input sequence is truncated if longer than a given length, or padded using the padding token [PAD] otherwise. The BERT output E_{d_s} from the sub-category branch for sub category s and the BERT output E_{d_m} from the main-category branch for main category m are obtained in the same way as in Equation (3.1). Note that the same BERT model is used in all the three branches.

Although BERT model is designed to capture the sequential information of input text, Wang et al. (2020) and Liu et al. (2020c) argue that transformer-based models perform poorly on capturing relative distant information among tokens. In order to alleviate this problem, a specific RNN model, Bi-LSTM, is used to better capture the sequential relationship of the input. While capable of encoding sequential relationship, conventional RNN however usually suffers from the vanishing gradient problem and hence can hardly capture the information from long input sequences. LSTM model

(Hochreiter and Schmidhuber, 1997) is introduced to solve this problem by proposing a memory cell and three gates to keep history information selectively when processing the sequence. For example, in the chief complaint branch, LSTM model is defined as follows:

$$f_c^t = \sigma\left(W_f E_c^t + U_f h_c^{t-1} + b_f\right) \tag{3.2}$$

$$q_c^t = \sigma\left(W_i E_c^t + U_i h_c^{t-1} + b_i\right) \tag{3.3}$$

$$m_c^t = f_c^t \circ m_c^{t-1} + q_c^t \circ \tanh\left(W_a E_c^t + U_a h_c^{t-1} + b_a\right) \tag{3.4}$$

$$o_c^t = \sigma\left(W_o E_c^t + U_o h_c^{t-1} + b_o\right) \tag{3.5}$$

$$h_c^t = o_c^t \circ \tanh\left(m_c^t\right), \tag{3.6}$$

where m_c^t denotes the memory cell state, W and U for each gate denote the learnable weight matrix, b for each gate denotes the learnable bias vector, σ denotes the sigmoid function, $\circ$ denotes the element-wise multiplication, and h_c^t denotes the hidden state output at current timestep. At each time step t, after receiving E_c^t from BERT, forget gate f_c^t from LSTM block firstly selects whether to keep or forget the history. Then input gate q_c^t determines whether there is some useful information in the input token at current timestep E_c^t needed to update the memory cell state. Finally output gate o_c^t decides what information to output. Note that instead of encoding the input from the beginning to the end, Bi-LSTM encodes the same input from both directions and concatenates into the final representation $\tilde{h}_c^t$ for each timestep t. It is also noted that the same Bi-LSTM model is used for both the sub-category branch and the main-category branch, which differs from the Bi-LSTM model used for the chief complaint branch.

3.2.4 Hierarchical Relational Network with Attention

The idea of simple relational network is proposed by Raposo et al. (2017) for relational reasoning. Santoro et al. (2017) applies the network to solve the Visual Question Answering (VQA) task by finding the relationship among objects in the picture and the question text. It is shown that relational network is able to effectively learn the relationships among objects at the same time while ignoring their orders. Given an object set $o = (o_1, o_2, ..., o_n)$, the simplest relational network is defined as follows:

$$r = f_\theta(\sum_{i,j} g_\phi(o_i, o_j)), \tag{3.7}$$

where f_θ is an MLP capturing the relationships between each object pair and g_ϕ is also an MLP capturing the overall relational information.

In this problem, strong semantic relationships exist among the chief complaint text and the descriptions of its hierarchical labels. Furthermore, relationships also exist between the two types of descriptions themselves: sub-category and main-category descriptions. Capturing these relationships is critical for effectively tackling the problem. Inspired by the work of Santoro et al. (2017), the HRNA module is devised to capture these complex relationships using several relational network-like structures with a hierarchical structure and an attention mechanism.

After getting the hidden representations for each branch, attention mechanism is firstly applied to the Bi-LSTM outputs which aids in capturing the pivotal words that contribute more to the complete semantics of a sentence. For example, attention mechanism is applied to the output of the

chief complaint branch as follows:

$$u_c^t = \tanh(W_h \tilde{h}_c^t + b_h) \tag{3.8}$$

$$\alpha_{h_c^t} = \frac{\exp\left({u_c^t}^\top v_c\right)}{\sum_t \exp\left({u_c^t}^\top v_c\right)} \tag{3.9}$$

$$p_c = \sum_t \alpha_{h_c^t} \tilde{h}_c^t. \tag{3.10}$$

Each hidden state $\tilde{h}_c^t$ at timestep t is firstly fed into a single fully connected layer with *tanh* as the activation function to obtain hidden representation u_c^t. Then, a learnable vector v_c is used to obtain the similarity score by multiplying with each hidden representation u_c^t. Function $softmax$ is then applied to calculate the weight for each hidden state. The final representation of the input text, p_c, is the weighted sum to all the hidden states based on their weights (see the following equations for details). The weights calculated this way represent the importance level of each word in the input text. The attention mechanism is applied similarly to the sub-category branch and the main-category branch to calculate p_{d_s} and p_{d_m}. Note that separate fully connected layers and learnable vectors are utilized for each branch.

A hierarchy of several relational network-like structures is used to capture the complex relationships among the encoded chief complaint text and its hierarchical label descriptions after the attention mechanism. Specifically, this is achieved by capturing the following four types of relationships: 1) the relationship between chief complaint and main-category descriptions; 2) the relationship between chief complaint and sub-category descriptions; 3) the hierarchical relationship between the first and second relationships; and 4) the direct relationship among chief complaint, main-category descriptions, and sub-category descriptions. The design aims to decompose

the complex hierarchical relationship into multiple straightforward and direct relationships that are easily comprehensible and captured by these structures. This arrangement enables the module to naturally grasp the intrinsic attributes of hierarchical relational reasoning.

To capture the relationship between the chief complaint text c and each main category m_i, the hidden representation of the chief complaint text p_c is concatenated with the hidden representation of each main-category description $p_{d_{m_i}}$, and then fed into a two-layer MLP g_{ϕ_1}, as follows:

$$r_{cm_i} = g_{\phi_1}(p_c, p_{d_{m_i}}). \tag{3.11}$$

Note that for each sub category SIE encodes its main category separately leading to multiple hidden representations for the same main category. In the HRNA, the average of these hidden representations is used, denoted as $p_{d_{m_i}}$. The relationship between chief complaint text c and each sub-category description is captured similarly as follows:

$$r_{cs_i^j} = g_{\phi_2}(p_c, p_{d_{s_i^j}}). \tag{3.12}$$

For each sub category s_i^j, the resultant $r_{cs_i^j}$ and r_{cm_i} are concatenated and fed to another two-layer MLP g_{ϕ_3} to capture the hierarchical relationships between the first two relationships, which are then aggregated using element-wise sum as follows:

$$r_{ms} = \sum_{i \in |M|} \sum_{j \in |S_i|} g_{\phi_3}(r_{cm_i}, r_{cs_i^j}), \tag{3.13}$$

where the aggregated vector r_{m_s} represents the relationship between chief complaint and the hierarchical labels. Similarly, the direct relationship among the chief complaint, main-category descriptions, and sub-category

descriptions is captured as follows:

$$r_{cms} = \sum_{i \in |M|} \sum_{j \in |S_i|} g_{\phi_4}(p_c, p_{d_{s_i^j}}, p_{d_{m_i}}), \tag{3.14}$$

where g_{ϕ_4} is a two-layer MLP. Finally, the resultant r_{m_s} and r_{cms} are concatenated and fed into the another two-layer MLP f_θ whose output nodes correspond to each class label (sub category) as follows:

$$\hat{s} = f_\theta(r_{ms}, r_{cms}). \tag{3.15}$$

The label corresponding to the highest score is the final prediction result.

3.2.5 Training Details

In last few years, research works leverage pre-trained models to alleviate the problem of limited data for downstream tasks. This is done by initializing network modules with parameters that are trained with more general and vast amount of data and then fine-tuning the whole model with downstream data. While fine-tuning improves performance compared with using pre-trained model only, the training process usually takes much longer time to properly update the pre-trained model for downstream tasks. Furthermore it is hard to decide how many epochs that should be used to fine-tune different models.

Ideally, fine-tuning the pre-trained BERT models of the proposed framework together should provide the best performance. However, this approach requires excessive GPU memory and time during the training process due to the complexity of the framework. Instead, three fine-tuning methods are proposed in this chapter: 1) randomly fine-tuning one of

the three branches while freezing the other two branches at each iteration (denoted as Random); 2) fine-tuning only the chief complaint branch while freezing the other two (denoted as CC); and 3) fine-tuning a separate BERT model (denoted as Separate) instead of the more complicated proposed model, then using the fine-tuned model in the proposed framework without further fine-tuning. This separate model simply uses BERT with a MLP classifier, making it much less resource-demanding for fine-tuning.

Additionally, the popular Cross-Entropy loss function is used to train the model. Cross-Entropy loss has proven to be very effective when dealing with multi-class classification problems in both computer vision and NLP. See Equation 3.16 for the details of the loss function:

$$loss = \frac{1}{K}\sum_{k=1}^{K} -\sum_{s\in S} \log \frac{\exp\left(z_{k,s}\right)}{\exp\left(\sum_{i=1}^{|S|} z_{k,i}\right)} y_{k,s}, \tag{3.16}$$

where K denotes the batch size, z denotes the predicted probability score for each sub category, and y denotes the one-hot representation for the ground truth sub-category label.

3.3 Chief Complaint Data and Label Description Extraction

3.3.1 Dataset and Pre-processing

The proposed model is evaluated on two public medical datasets. The first one is the cMedQA v1.0 dataset provided by Zhang et al. (2017). The cMedQA dataset, collected from an online Chinese medical Question and Answering forum, contains questions posed by online patients and their

answers from certified physicians. Each question is assigned with its main category, e.g., "Internal Medicine", and sub category, e.g., "Neurology" by patients or physicians, which are considered as ground truth labels. There are around 55,000 questions in the dataset and 16 main categories and totally 217 sub categories. Table 3.1 shows some of the representative examples from the dataset. Note that the original text is translated from Chinese to English in the table.

Chief Complaint	Main Category	Sub Category
I have stomach ache and acid reflux. I also have nausea and diarrhea.	Internal Medicine	Gastroenterology
My knees are swollen and cannot be bent. They will tremble when I straighten my legs.	Surgery	Orthopedics
My right breast is swollen and painful. There is a capsule-sized lump under my right armpit and sometimes I feel pain around it.	Surgery	Breast Surgery
I am 41 years old. I have a severe dysmenorrhea with blood clots. There are discolored spots on my face. Are these spots caused by dysmenorrhea? How to treat it?	Obstetrics	Dysmenorrhea
My right ankle is broken. It has been in a plaster cast for seven days. I felt pain when I stood up these days.	Surgery	Orthopedics
My face is often allergic with red bumps and swelling.	Dermatology	Allergy
I feel pain in my temples when I catch a cold.	Internal Medicine	Respiratory

Table 3.1: Examples of chief complaints and their hierarchical labels from the cMedQA dataset.

Although the dataset serves multiple purposes, such as question answering, dialogue generation, and text classification, this chapter focuses solely on using questions as chief complaints without incorporating answers. Note that all the questions in the dataset are medical related and usually contain problems, symptoms, conditions, previous diagnosis, etc., which are essentially the chief complaints of those online patients. In addition, only samples from the top five main categories and 38 sub categories are used in the experiments. This is because other categories have significantly less number of samples. Subcategories containing fewer than 200 samples are excluded from consideration. Other pre-processing includes removals of all

punctuation, emojis, and system generated text, and replacement of all Chinese characters for numbers with Arabic numbers. Note that stemming is not required as Chinese language does not have any inflection of words as English does, e.g. "take vs taken", or "inject vs injection". In total, there are 19,686 questions in the training set and 10,858 questions in the testing set.

The second dataset is the kaMed dataset collected by Li et al. (2021b). The data are also from Chinese QA forums that have similar architecture of the cMedQA dataset, containing 63,754 dialogues. For every dialog, the first utterance from the patient is taken as the chief complaint text, and the main-category and sub-category information are stored in the "disease_-grad" tag. There are 13 main categories and 40 sub categories. The non-applicable sub category "All Department" is not used, leaving 54,333 chief complaints. The data is randomly divided into 70% for training and 30% for testing. In total, there 38,033 dialogues in the training set and 16,300 dialogues in the testing set.

Although the two datasets share similarities in format and source—both derived from Chinese online medical QAs—they differ in the number of categories and the granularity of annotation. Specifically, the kaMed dataset contains more main categories (13 v.s. 5 used in cMedQA) and offers a slightly broader distribution across departments. By evaluating the model on both datasets, this thesis aims to demonstrate its robustness and generalization capability under different category configurations. Experimental results show that the "Orthopedics" department achieves the best classification performance in both datasets, most likely due to its clearer symptom descriptions and relatively consistent terminology used by patients.

3.3.2 Label Descriptions

The cMedQA dataset and kaMed dataset only contain category names or labels, and there are no detailed descriptions for those categories. While category names carry some information, such information is insufficient for the classification task. More information for each category is collected from existing data, such as medical textbooks and online sources like Wikipedia, to form comprehensive label descriptions. This section outlines the process of preparing descriptions for each main category and sub-category.

Main Category Chinese medical book series, namely "Medical Guidance Books for Clinical Doctor Qualification Examination", are used to extract main-category descriptions. The book series are selected because they are the official guide books for the National Examination for Physicians License in China, which cover almost every aspect of medical knowledge. The books are firstly converted to electronic text using OCR technique and then divided into paragraphs. Manual verification of the OCR output is conducted for any necessary error corrections. Each paragraph is annotated with metadata that contains names of chapter, section, and subsections it belongs to. The sections in the book series have a structure of four level subsections. Table 3.2 illustrates example paragraphs extracted from the book series with their chapter and section names at different levels.

Ideally, all the paragraphs in the chapter corresponding to each main category in the experiments could serve as category descriptions. However, each chapter contains thousands of words, which are too lengthy to be input into the BERT model. Furthermore, not all of these words convey significant information about the topic in the chapter. Instead, the TF-IDF technique is utilized to extract the most informative words for the main-category descriptions. Specifically, for each chapter, the "Jieba" (Sun, 2020) Chi-

Chapter	Level-1 section	Level-2 section	Level-3 section	Level-4 section	Content
Anatomy	Locomotor system	Osteology	Classification of bones	Flat bone	Flat bones are plate-shaped and participate in the formation of cranial, thoracic, and pelvic walls. It can also protect organs, such as skull and ribs.
Anatomy	Digestive system	Pancreas	Distribution of pancreas	Cauda pancreatis	Cauda pancreatis is thinner, running from the upper left to the left quarter rib area, and is in contact with the visceral surface of the spleen below the splenic hilum.
Internal Medicine	Heart valve disease	Aortic valve stenosis	Complication	Body circulation embolism	Rare, more common in calcified aortic stenosis.
Surgery	Intestinal obstruction and appendicitis	Acute appendicitis	Anatomy and physiology of appendix	Appendix vein	Inflammation of the appendix can be through the appendix vein-colon vein-superior mesenteric vein-portal vein-liver. Therefore, inflammation of the appendix can cause portal phlebitis and liver abscess.
Pathology	Local blood circulation disorder	Thrombosis	Thrombus outcome	(No_Name)	The newly formed thrombus can be softened, dissolved and absorbed.

Table 3.2: Example paragraphs from the reference books with chapter and section names at different levels.

nese text segmentation module is used to tokenize text into Chinese words. Note that unlike English language there is no space or other symbols that are used in Chinese to delimit different words and a single Chinese word may contain one or multiple characters. The Term Frequency (TF) for each word is calculated according to Equation 3.17 and Inverse Document Frequency (IDF) according to Equation 3.18 as follows:

$$tf(w_i, d) = \frac{f_d(w_i)}{\sum_k f_d(w_k)} \tag{3.17}$$

$$idf(w_i, D) = \log \frac{N}{1 + |\{d \in D : w_i \in d\}|} \tag{3.18}$$

$$tf\text{-}idf(w_i, d, D) = tf(w_i, d) \times idf(w_i, D). \tag{3.19}$$

Finally, the top 50 words with highest TF-IDF scores are selected according to Equation 3.19, and these words are concatenated in order to form the description of the corresponding main category. Table 3.3 illustrates the top 10 words for two example main categories.

Ranking	Obstetric		Surgery	
	word	TF-IDF	word	TF-IDF
1	contractions	0.0505	fracture	0.0346
2	reveal	0.0496	choledochus	0.0338
3	fetal head	0.0485	reduction	0.0334
4	stage	0.0476	injury	0.0330
5	uterine orifice	0.0446	spermatic cord	0.0330
6	fetal membrane	0.0442	femur	0.0325
7	fibroids	0.0432	maneuver	0.0323
8	contraception	0.0429	calculus	0.0321
9	hydatidiform mole	0.0424	shank	0.0313
10	parturition	0.0413	closed	0.0313

Table 3.3: Top 10 words with highest TF-IDF scores for main category "Obstetric" and "Surgery". Note that some entries in this table consist of multiple English words because some of the original single Chinese words are translated into multiple English words.

Sub Category Unlike main-category descriptions, there is no easy way to extract sub-category descriptions from the guide book series because there is no one-to-one mapping between sub categories and book sections. Instead, entries corresponding to sub-categories from Wikipedia and Baidu Baike (a Chinese online encyclopedia similar to Wikipedia) serve as sources for sub-category descriptions. Although not as dedicated as the guide book series, Wikipedia and Baidu Baike are widely accepted as quasi-professional references in many domain works. For each sub-category, the summary paragraphs (usually the first paragraph) of the corresponding entries are used as the description. The summary paragraph is preferred over the whole document because summaries for sub-categories typically cover sufficient information, while summaries for main categories are generally more abstract. Furthermore, it is not necessary to select top words based on TF-IDF scores for sub-category descriptions as the summaries are much shorter.

Table 3.4 shows example descriptions for main category "Surgery" and sub category "Neurosurgery". Note that if there is no corresponding chapter for a main category, the same approach is used to enrich the main-category description.

Surgery	Neurosurgery
fracture, choledochus, reduction, injury, spermatic cord, femur, maneuver, calculus ,shank closed, shoulder joint, periosteum, fire burn, caput femoris, fibula, luxation, gallbladder, lung cancer, conservative therapy, abduction, gypsum, urinary, cramp, bonetumor, kidney cancer, humerus, wound, dehydration, pancreatitis, colorectal cancer, cut, postoperative, diseased limb, traction, the wounded, germ, injury, pancreatic duct, articulation of knee, intestinal obstruction, radial bone, abducent nerve, tourniquet, boundege, tendon, appendicitis, nephrophthisis, gastric cancer, hydronephrosis, extensor muscle	Neurosurgery or neurological surgery, known in common parlance as brain surgery, whose main treatment method is based on operation. It applies unique neurosurgery research methods to study nervous system including brain, spinal cord, central and peripheral nervous system, and cerebrovascular system. The related diseases or injuries will also be concerned by neurosurgery such as skull traumas, meningitis, brain tumors, malformation, certain genetic metabolic disorders or dysfunction diseases, e.g. epilepsy, Parkinson's disease, neuralgia.

Table 3.4: Example descriptions of main category "Surgery" and sub category "Neurosurgery".

3.4 Experimental Results

3.4.1 Experimental Settings

The PyTorch library (Paszke et al., 2019) is used to implement the framework, utilizing BERT-Base-Chinese as the pre-trained model, which consists of 12 transformer encoders with 12 heads for multi-head attention. The hidden dimension of the transformers is 768 and the maximum input sequence length allowed is 512. Single layer is used for Bi-LSTM with hidden size set to 256. The hidden size for MLP is set to 2048. Adam op-

timizer (Kingma and Ba, 2014) is used to train the model with automatic mixed precision method to accelerate the training process. The batch size for training is set to 10. Learning rate warm-up proportion is set to 0.1 and weight decay coefficient is set to 0.01. Drop out rate is set to 0.1. The number of epochs for training is set to 10 for time efficiency.

The popular F-measure for NLP tasks is employed to compare the framework with other models. Since the task is a multi-class classification problem, both macro F1 and micro F1 are utilized. Additionally, macro precision and macro recall are provided for further insights. To evaluate the effectiveness of the proposed framework, the performance is compared with the following baselines.

BERT This baseline model only contains a basic vanilla BERT model followed by a single-layer feed-forward neural network serving as the classifier. This is the most straight forward way of using Large Language Model (LLM). The model only takes chief complaint text as input without using hierarchical label information. The hyperparameters used in this model are the same as ours.

P-tuning Introduced by (Liu et al., 2023b), the method exploits prompt engineering by integrating learnable variables into the embedded input to create continuous prompts. The model is proven to be effective to improve the performance of pre-trained language models in downstream tasks. Eight trainable prompts are integrated into BERT, positioned before the original input. All the other settings are the same as the BERT baseline.

BERT+LSTM This baseline model consists of BERT and Bi-LSTM. This is the simplest model combining BERT with Bi-LSTM, which is often used in many NLP tasks. The hidden states at different time step from Bi-LSTM are concatenated before being fed to a single layer feed forward network for

final classification. The hyperparameters used in this model are the same as ours.

BERT+GRU GRU is another widely used RNN module that has similar gate structure as in LSTM and often produces comparable results (Cho et al., 2014). This baseline consists of base BERT module and Bi-GRU, similar to the baseline BERT+LSTM. The hidden states at different time step from Bi-GRU are concatenated before being fed to a single layer feed forward network for final classification.

BERT+GRU+Attention This baseline adds attention mechanism to model BERT+GRU. Attention (Vaswani et al., 2017) mechanism becomes much more popular since introduced and proven to be very effective when using with base encoding methods such as BERT.

ChatGPT ChatGPT serves as the cutting-edge conversational AI developed by OpenAI. As LLM, ChatGPT generates responses that closely resemble human conversation on a wide array of subjects. The default gpt-3.5-turbo model (Brown et al., 2020) with manually designed prompt is used as another baseline model. The prompt is structured as follows: "Here is the category list: [Respiratory Medicine, Gynecology, ...]. Please select the category that best matches the following chief complaint: *chief complaint text*. Kindly respond with only the category name." This prompt enables ChatGPT to identify the most suitable class label based on the given chief complaint text.

LEHS The technique proposed by Miyazaki et al. (2019) shares a similar idea, utilizing the hierarchical structure of labels as additional information to enhance classification performance. The same hyperparameters provided by the authors are employed, except for class weights. The authors suggest that using class weights for the training can improve the performance.

However, experiments indicate that omitting class weights yields better results, likely due to the more balanced nature of the data compared to that used in their study. Therefore, only results without class weights are presented in this chapter.

HE-HMTC This technique proposed by Ma et al. (2022) uses a combined text representation incorporating a Bi-GRU-based text representation, along with a graph embedding based on the hierarchical category structure, and a word embedding based on category names. In this chapter, the category names are enriched by their corresponding label descriptions. Jieba (Sun, 2020) tokenization tool and Chinese word2vec (Qiu et al., 2018) are also used to represent category names and input text.

LA-HCN The technique proposed by Zhang et al. (2022) incorporates a unique label-based attention module that hierarchically extracts significant information from the input text, leveraging the hierarchical label structure across different levels. In contrast, the focus of this chapter is on the second-level label classification results (sub-category in this application), as it is argued that this level is the most useful in real-world scenarios.

3.4.2 Ablation Studies

Ablation studies are conducted to evaluate the effectiveness of the two components in the proposed framework: SIE and HRNA. The model is assessed using two configurations, SIE-single and SIE-multiple. SIE-single consists only of the chief complaint branch, with the sub-category and main-category branches removed from the SIE module. The last hidden state output from each direction of the Bi-LSTM is concatenated and then fed into the final MLP for classification. In contrast, SIE-multiple retains all

three branches but omits the HRNA module. The last hidden state outputs of the Bi-LSTM for each branch are concatenated and sent to the final MLP for classification, allowing an evaluation of the utility of hierarchical label information. The SIE-single model serves as a base model, similar to conventional text classification models that lack additional information, such as expert knowledge. The final MLP structure remains consistent across the other models.

Model	Micro F1	Macro F1	Macro Precision	Macro Recall
SIE–single	0.563	0.448	0.508	0.456
SIE–multiple	0.585	0.506	0.519	0.523
SIE+HRNA	**0.644**	**0.618**	**0.608**	**0.641**

Table 3.5: Ablation studies evaluating different components of the proposed framework on the cMedQA dataset.

Model	Micro F1	Macro F1	Macro Precision	Macro Recall
SIE–single	0.741	0.678	0.700	0.681
SIE–multiple	0.759	0.717	0.731	0.714
SIE+HRNA	**0.796**	**0.768**	**0.768**	**0.769**

Table 3.6: Ablation studies evaluating different components of the proposed framework on the kaMed dataset.

The ablation study results are shown in Table 3.5 and 3.6. On the cMedQA dataset, the proposed framework achieves 0.644 for micro F1 and 0.618 for macro F1 which is the highest among all the models compared. On the kaMed dataset, this framework also achieves the highest micro F1 score and macro F1 score, which are 0.796 and 0.768. Both micro F1 and macro F1 scores drop significantly when the HRNA module is removed. The same performance improvement can be seen for macro precision and macro recall as well. The ablation results clearly show that using HRNA to capture the relationships from hierarchical structure of label descriptions can significantly improve the overall performance of chief complaint classification. From the tables, it is evident that the additional label information is beneficial for classifying chief complaint text, even when simply concatenated.

3.4.3 Comparisons to Baselines

The comparison of this framework with the state-of-the-art models on the two datasets is summarized in Table 3.7 and Table 3.8. As shown in these tables, the model outperforms all other baselines on both datasets for almost every evaluation metric. This is due to the effective combination of the SIE and HRNA module, resulting in significant performance improvement over LEHS and LA-HCN. Specifically, compared to the LEHS model, the model improves the micro F1 score by more than 7% on the cMedQA dataset and 5% on the kaMed dataset. In addition, this model improves the macro F1 score by over 12% on the cMedQA dataset and 5% on the kaMed dataset compared to the LEHS model. Compared to the LA-HCN model, the model still performs the best on all four metrics in both datasets. This model improves the micro F1 score by 7% on the cMedQA dataset and 2% on the kaMed dataset. Furthermore, the model improves the macro F1 score by nearly 5% on the cMedQA dataset and 2% on the kaMed dataset compared to the LA-HCN model. Compared to the HE-HMTC model, the model still performs competitively. This model improves the macro F1 score for more than 4% on both dataset. As for the micro F1 score, the model performs better on the kaMed dataset, while the HE-HMTC model performs better on the cMedQA dataset. One potential explanation to this is that the HE-HMTC model splits the classification task into multiple tasks, using separate models for each level of label classification. The results of the current level are then used in the classification task of the next level. And there are only five main categories in the cMedQA dataset, which are much less than seven main categories in the kaMed dataset. It is observed that the baselines relying solely on LLMs such as BERT and ChatGPT perform poorly in this classification task. One potential explanation is that these LLMs are not trained to effectively utilize hierarchical label

information, which hinders their ability to accurately comprehend and classify chief complaint text. The P-tuning approach does improve the BERT model in the experiments but still has a significant difference compared to ours. Compared to the other BERT based models such as BERT+LSTM, BERT+GRU, and BERT+GRU+Attention, this model still provides significant improvement. The performance gap between the models without a hierarchical structure (BERT, P-tuning, BERT+LSTM, BERT+GRU, BERT+GRU+Attention, ChatGPT) and those with a hierarchical structure (LEHS, LA-HCN, HE-HMTC, and the proposed model) is evident. This emphasizes the critical role of incorporating the hierarchical structure of label information into the model for improved performance.

Model	Micro F1	Macro F1	Macro Precision	Macro Recall
BERT	0.472	0.416	0.412	0.433
P-tuning	0.552	0.505	0.515	0.519
BERT+LSTM	0.548	0.488	0.492	0.501
BERT+GRU	0.526	0.458	0.452	0.475
BERT+GRU+Attention	0.575	0.405	0.493	0.414
ChatGPT	0.466	0.393	0.444	0.443
LEHS	0.566	0.495	0.502	0.510
LA-HCN	0.574	0.569	0.541	0.613
HE-HMTC	**0.668**	0.575	0.591	0.596
Ours	0.644	**0.618**	**0.608**	**0.641**

Table 3.7: Evaluation results for different baselines on the cMedQA dataset.

Model	Micro F1	Macro F1	Macro Precision	Macro Recall
BERT	0.704	0.671	0.676	0.669
P-tuning	0.743	0.708	0.714	0.709
BERT+LSTM	0.745	0.711	0.718	0.709
BERT+GRU	0.738	0.701	0.706	0.700
BERT+GRU+Attention	0.764	0.701	0.725	0.705
ChatGPT	0.583	0.506	0.559	0.511
LEHS	0.755	0.711	0.723	0.709
LA-HCN	0.774	0.747	0.735	0.761
HE-HMTC	0.778	0.727	0.759	0.738
Ours	**0.796**	**0.768**	**0.768**	**0.769**

Table 3.8: Evaluation results for different baselines on the kaMed dataset.

3.4.4 Evaluation of Fine-tuning Methods

To ensure fair comparison, previous experiments are carried out in a way that the base BERT model is not updated during training (denoted as Frozen) while the rest of the models are updated with the training data. In this section, proposed different ways (see Section 4.2.6) are evaluated on the cMedQA dataset to fine-tune the proposed model aiming to improve the overall performance with limit computational resources such as GPU memory and time.

Fine-tuning	Micro F1	Macro F1	Macro Precision	Macro Recall
Frozen	0.644	0.618	0.608	0.641
Random	0.656	0.646	0.643	0.667
CC	**0.679**	0.670	**0.667**	0.695
Separate	0.678	**0.672**	0.650	**0.705**

Table 3.9: Results on different fine-tuning methods on the cMedQA dataset.

Table 3.9 shows that fine-tuned models always outperform the pre-trained model. Furthermore, only fine-tuning the chief complaint branch (CC) performs much better than randomly updating different branches (Random). The result of fine-tuning a separate model before loading to the proposed framework (Separate) is similar to fine-tuning the chief complaint branch. Possible explanation for this could be the fact that chief complaints cover a large variety of words while words for category descriptions are limited. Hence focusing fine-tuning with chief complaint text converges to better results. Note that both CC and Separate methods use chief complaint text only for fine-tuning.

Main category	Sub category	Chief Complaint
Internal Medicine	Respiratory	I feel pain in my temples when I catch a cold .
Internal Medicine	Gastroenterology	I have stomachache and acid reflux . I also have nausea and diarrhea .
Dermatology	Allergy	My face is often allergic with red bumps and swelling .
Surgical	Orthopedics	My right ankle is broken . It has been in a plaster cast for seven days . I felt pain when I stood up these days .
Surgical	Breast Surgery	My right breast is swollen and painful . There is a capsule-sized lump under my right armpit and sometimes I feel pain around it .
Surgical	Orthopedics	My knees are swollen and cannot be bent . They will tremble when I straighten my legs .
Obstetrics	Dysmenorrhea	I am 41 years old . I have a servere dysmenorrhea with blood clots . There are discolored spots on my face . Are these spots caused by dysmenorrhea ? How to treat it ?

Figure 3.2: Words from chief complaints with attention values. Note that words with higher attention values are highlighted with darker red.

3.4.5 Effectiveness of Attention Mechanism

As introduced in Section 3.2.4, attention mechanism is able to assign higher weights to the informative words from the input. This section visualizes the learned attentional scores for multiple examples of chief complaints to verify the effectiveness of the attention mechanism used in the HRNA module.

Sub category	Description
Orthopedics	Orthopedics or Orthopedic surgery , is a branch of surgery . Orthpedic surgeons use both surgical and nonsurgical means to treat musculoskeletal trauma , spine diseases , sports injuries , degenerative diseases , infections , tumors and congenital disorders .
Cardiovascular Medicine	The cardiovascular diseases treated by cardiovascular medicine department include angina , hypertension , arrhythmia , heart failure , premature beat , myocardial infarction , cardiomyopathy , myocarditis , acute myocardial infarction .

Figure 3.3: Words from sub-category descriptions with attentional scores. Note that words with higher attentional scores are highlighted with darker red.

Figure 3.2 highlights the important words according to their attentional scores from the chief complaint branch of the proposed framework. It is

Ranking	TF-IDF words	Attention words
1	contractions	cervical
2	reveal	ovary
3	fetal head	amenorrhea
4	stage	uterus
5	uterine orifice	fetal heart
6	fetal membrane	villus
7	fibroids	contraception
8	contraception	early pregnancy
9	hydatidiform mole	gravida
10	parturition	bleeding

Table 3.10: Top 10 words with highest TF-IDF scores and attentional scores for main category "Obstetric". Note that some entries in this table consists of multiple English words because some of the original single Chinese words are translated into multiple English words.

clear that the attention mechanism effectively captures significant words from the chief complaint that are closely related to the corresponding main category and subcategory. For instance, given the chief complaint under sub category "Gastroenterology", "I have stomachache and acid reflux. I also have nausea and diarrhea.", words that are common symptoms for digestive diseases, such as "stomachache", "acid reflux", "nausea", and "diarrhea", have higher attentional scores (highlighted in different shades of red). Similarly, Figure 3.3 highlights the important words according to their attentional scores from the sub-category branch of the proposed framework.

Table 3.10 shows the comparison of the top 10 words with highest TF-IDF scores and with highest attentional scores respectively from an example main-category description, "Obstetric". The attention mechanism re-ranks the description words based on their relevance to chief complaints, placing terms that are more commonly used by patients at the top. For example, words such as "ovary", "gravida", "bleeding" are more commonly used by patients than those with higher TF-IDF scores.

3.5 Conclusion

In this chapter, a novel deep-learning-based framework is proposed that utilizes the hierarchical structure of labels with external expert knowledge to classify a specific type of medical text, chief complaint, into different categories that usually correspond to departments or specialized areas in the context of online healthcare systems. The proposed framework consists of three branches: chief complaint branch, main-category branch, and sub-category branch, along with two modules: Sequence Information Encoder and Hierarchical Relational Network with Attention module. This framework effectively embeds text from both chief complaint and hierarchical structure of label descriptions extracted from professional medical resources, leveraging the hierarchical structure of labels by capturing complex relationships among label descriptions and input text with attentional scores. Experimental results on the cMedQA and kaMed datasets demonstrate the capability of the proposed framework, outperforming the baseline models by a significant margin on almost all metrics used. Future work will focus on enhancing the model within the medical domain by extending it to other types of medical narratives such as diagnostic notes and discharge summaries. In addition, I plan to investigate more sophisticated methods for incorporating expert knowledge, such as using knowledge graphs or ontology-based representations, for further improvement.

Chapter 4

Multi-turn Medical Dialogue Generation Using Alternating Recurrent Wasserstein Autoencoders

Having established a robust framework for classifying chief complaint texts, chapter 4 transitions to the next critical aspect of patient-provider interactions: medical dialogue generation. While accurate classification of chief complaints lays the groundwork for effective triage and referral processes, the ability to generate contextually appropriate medical dialogues is essential for facilitating meaningful communication between patients and healthcare providers. In this chapter, a novel framework is introduced to simulate patient-doctor conversations, consisting of two identical models with different parameters to effectively represent the distinct roles of doctors and patients. By integrating narrative medical knowledge from clinical guidelines and employing advanced techniques such as a Wasserstein auto-encoder,

the framework generates coherent and contextually relevant dialogues that reflect both medical terminology and patients' informal expressions.

4.1 Introduction

Dialogue generation is one of the popular research topics in the area of Natural Language Processing (NLP). Multi-turn dialogue frameworks, unlike single-turn interactions, require explicit modeling of dialogue context—the sequential dependency between utterances—to maintain coherence and relevance across turns. This is especially critical in medical dialogues, where context (e.g., symptom progression, patient history) directly influences diagnostic and communicative outcomes. In healthcare, adopting automatic dialogue generation techniques has significant benefits, such as aiding doctors in gathering patient information, potentially reducing labor costs (Tang et al., 2016), assisting doctors in making clinical decisions (Liu et al., 2017; Xia et al., 2020), providing patients with access to healthcare services when face-to-face appointments with doctors are restricted (Varshney et al., 2023), etc.

The application of general dialogue generation techniques to the healthcare domain has significant challenges due to the inherent nature of medical text. Often unstructured or semi-structured medical text contain information that demands extensive domain-specific knowledge for proper comprehension. Moreover, conversations between physicians and patients vary greatly: 1) physician narratives often employ acronyms and abbreviations for both common and professional medical terminology; 2) patient narratives usually are less accurate because of the usage of informal and vague words instead of medical jargon, or providing information without proper

context. Dialogue generation approaches need to be able to accommodate these differences to generate proper and meaningful responses according to the current speaking role.

In previous works, in order to understand the medical-related knowledge in the dialogue utterances, most medical dialogue generation methods require large human annotations which describe the corresponding utterances' state, action and related entities, etc. (Liu et al., 2022; Varshney et al., 2023). These human-annotated data are normally in a well-structured form such as a knowledge graph. Creating these well-structured annotated data requires vast human efforts. Furthermore, traditional sequence-to-sequence dialogue generation models are more likely to produce uninformative and repetitive responses (Sato et al., 2017) which is not suitable for medical scenarios. To mitigate this issue, some researchers have proposed Conditional Variational Autoencoder (CVAE) to generate diverse responses by utilizing a latent variable to capture the underlying information in the given context (Sohn et al., 2015; Zhao et al., 2017; Shen et al., 2018). However, CVAE-based dialogue generation models often face the challenge of "posterior collapse". To address this challenge, the dialogue models based on the Wasserstein Auto-Encoder are proposed, modeling the prior and posterior distributions by training a Wasserstein GAN (Gu et al., 2019; Zhang et al., 2020a). Despite producing promising output, dialogue models presented by previous works usually ignore the different behaviors between the two roles, physician and patient, by simply directly concatenating the utterances from both speaking roles as dialogue history (Gu et al., 2019; Zhang et al., 2019).

This chapter aims to devise a multi-turn dialogue generation framework that incorporates medical knowledge in a flexible way based on WAE. Narrative medical data is utilized to supplement the generation of medical-

related dialogues. Unlike well-structured human-annotated data, narrative data is much easier to use, as it does not require complicated or costly processing. To properly distinguish between the roles of patients and physicians in medical conversations, two separate network models with the same architecture are employed to independently capture these different speaking behaviors. A memory mechanism is used to enable interaction between the two models. Specifically, a novel dialogue generation framework for patient-doctor dialogue generation is proposed by introducing two WAE-based language models to model the roles of patients and doctors separately. Each language model consists of three branches: the context branch, response branch, and knowledge branch. These branches take input from dialogue history, response utterance, and searched knowledge, respectively, which are firstly encoded by using an utterance encoder and context encoder and then a conditional Wasserstein Auto-Encoder to model the prior and posterior distributions based on the three inputs. Two discriminators are used to minimize prior distribution and posterior distributions.

The proposed framework effectively distinguishes between the different roles in patient-doctor conversations. Compared with other sequence-to-sequence models also based on VAE structures, the framework captures the latent information of input utterances more effectively and generates more diverse responses. Crucially, by integrating external medical knowledge into the knowledge branch of each WAE model, the framework grounds generated responses in evidence-based medicine. This reduces hallucinations and ensures alignment with domain-specific standards, a critical requirement for clinical applicability. The experimental analysis confirms that removing this knowledge integration significantly degrades performance across all key metrics, demonstrating its essential role in maintaining both medical accuracy and conversational quality. Evaluation of the model on two real

medical dialogue datasets demonstrates that it outperforms other baseline methods.

The main contributions presented in this chapter can be summarized as follows:

1. A multi-turn dialogue generation model is proposed, consisting of two separate models to represent patient-doctor roles in conversations, connected through a memory mechanism.

2. A Knowledge-based Conditional Wasserstein Auto-Encoder is employed in each model to effectively integrate dialogue history and external medical knowledge. This approach ensures that the framework generates responses that are both contextually accurate and medically relevant, while also enhancing the diversity of generated responses.

3. The proposed framework is evaluated using two real-world medical dialogue datasets, demonstrating superior performance compared to other baseline models.

4.2 Methodology

4.2.1 Problem Definition

Given a multi-turn dialogue c between a patient and a doctor, the dialogue can be represented as follows: $c = \{p_1, d_1, p_2, d_2, ..., p_n, d_n, ..., p_N, d_N\}$ where N denotes the maximum turn number, $p_n = \{w_1^{p_n}, w_2^{p_n}, ..., w_m^{p_n}, ..., w_{|p_n|}^{p_n}\}$ denotes the utterance in n^{th} turn from the patient, $d_n = \{w_1^{d_n}, w_2^{d_n}, ..., w_m^{d_n}, ..., w_{|d_n|}^{d_n}\}$ denotes the utterance in n^{th} turn from the doctor, and w_m represents the

m^{th} word in the utterance. The proposed framework aims to model these two different roles by using two different language models whose architectures are identical. The language model for patients is defined as L_p and the language model for doctors is defined as L_d. These two language models are connected to each other by using a memory mechanism. The probability distribution of generating utterances p_n and d_n from these two language models can be defined as follows:

$$\begin{aligned} &\prod_{n=1}^{N} P_{L_p}\left(p_n \mid p_{<n}, d_{<n}, K, memory\_state\right) \\ =&\prod_{n=1}^{N}\prod_{m=1}^{M} P_{L_p}\left(w_m^{p_n} \mid p_{<n}, d_{<n}, w_{\leq m}^{p_n}, K, memory\_state\right), \end{aligned} \tag{4.1}$$

$$\begin{aligned} &\prod_{n=1}^{N} P_{L_d}\left(d_n \mid d_{<n}, p_{\leq n}, K, memory\_state\right) \\ =&\prod_{n=1}^{N}\prod_{m=1}^{M} P_{L_d}\left(w_m^{d_n} \mid d_{<n}, p_{\leq n}, w_{\leq m}^{d_n}, K, memory\_state\right), \end{aligned} \tag{4.2}$$

where K denotes the knowledge based on the input utterances. The goal is to maximize these two probability distributions.

4.2.2 Overview

Figure 4.1 illustrates the overall structure of the proposed framework. The L_p language model represents the patient's role and the L_d language model represents the doctor's role. The proposed framework combines these two language models in an alternating order to capture and learn the speaking behavior of both patient and doctor through memory recurrence.

Each language model consists of three branches: the context branch, the

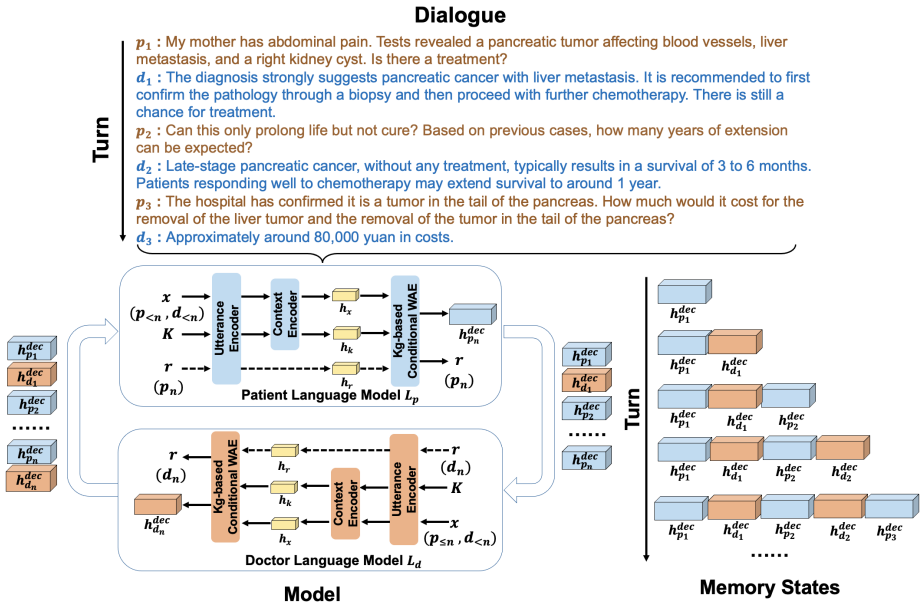


Figure 4.1: Overall architecture of the proposed framework.

response branch, and the knowledge branch. The input of the context branch is the dialogue history denoted as x, which includes multiple utterances exchanged between the patient and the doctor. The response branch takes the response utterance r as its input, representing the ground truth for the current turn. The knowledge branch receives context-dependent documents K, sourced from a structured medical guidance book through a detailed document selection process outlined in Section 4.3.2. Both the context branch and knowledge branch utilize an utterance encoder and a context encoder for input encoding. The utterance encoder processes each utterance, while the context encoder further encodes the cumulative information from all utterances after the initial processing by the utterance encoder. In the response branch, where only one utterance is present, encoding relies solely on the utterance encoder. Following the encoding of all inputs, a knowledge-based conditional Wasserstein Auto-Encoder is employed to model the response distributions, taking into consideration both contextual information and external knowledge. To be more specific, the knowledge represented by the searched documents and dialogue history are

used to separately model the prior distributions of the response. Furthermore, the dialogue history, searched documents, and response are combined to model the posterior distribution of the response. The knowledge-based conditional Wasserstein Auto-Encoder is employed to reconstruct the response utterance from a latent variable sampled from the posterior distribution, aligning the two prior distributions to be closer. This alignment ensures similar prior distributions with the posterior distribution during the inference stage. Further details regarding the utterance encoder, context encoder, and knowledge-based conditional Wasserstein Auto-Encoder are provided in Section 4.2.3 and Section 4.2.4. The detailed structure of each language model is illustrated in Figure 4.2.

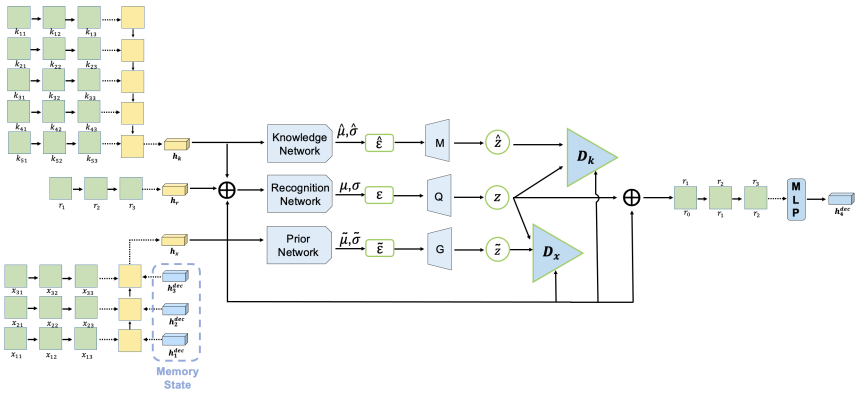


Figure 4.2: Detailed architecture of each language model.

Regarding memory recurrence, during the generation of each response in a complete dialogue, a linear transformation is applied to the final hidden state of the decoder, and the result is stored in a memory list. Subsequently, when generating the next utterance in the dialogue, this memory state serves as the input for the context encoder. This memory state offers an alternative representation of the corresponding response utterance. Consequently, in the generation of the next turn, both the memory state and the current context are utilized as inputs. The detail of this memory mechanism is introduced in Section 4.2.5

4.2.3 Input Representation

Bidirectional Gated Recurrent Unit (Bi-GRU) is selected as the encoder to the input due to its ability to model the sequential dependencies in both forward and backward directions, thus providing a richer contextual representation of each utterance. BiGRU has been widely adopted in dialogue systems, as it can effectively handle complex dependencies across words within an utterance by considering both past and future tokens.

In the proposed framework, the BiGRU-based utterance encoder processes each utterance x_i within the dialogue history, capturing both forward and backward context as shown in Equations 4.3 and 4.4. The input utterance is first tokenized into a sequence of words, with each word w_m being represented as a word embedding $e^{w_m^{x_i}}$. The hidden states from the forward and backward GRUs are concatenated to form h_i^{utt}, the final representation of the i^{th} utterance. This ensures that the model can leverage the information from the entire utterance, enhancing its ability to capture long-term dependencies, which is essential for understanding medical dialogues.

The utterance representations generated by the BiGRU are further processed by a unidirectional GRU, referred to as the context encoder. This context encoder captures the sequential relationships between utterances within the dialogue. The hidden state of the context encoder at each timestep, h_x, represents the cumulative information of the dialogue up to the current turn. Similarly, the same BiGRU is used to encode searched documents into their respective representations h_k, following the same encoding process used for dialogue utterances.

For the response branch, where only a single utterance is present, the context encoder is not necessary. Instead, the concatenation of the forward

and backward hidden states of the BiGRU is used directly as the response representation h_r, as described in Equation 4.5. This approach ensures that both the dialogue history and external knowledge are efficiently represented for subsequent stages in the framework.

$$\overrightarrow{h_{i,m}^{utt}} = \overrightarrow{GRU_{utt}}\left(\overrightarrow{h_{i,m-1}^{utt}}, e^{w_m^{x_i}}\right), \tag{4.3}$$

$$\overleftarrow{h_{i,m}^{utt}} = \overleftarrow{GRU_{utt}}\left(\overleftarrow{h_{i,m-1}^{utt}}, e^{w_m^{x_i}}\right), \tag{4.4}$$

$$h_i^{utt} = [\overrightarrow{h_{i,|x_i|}^{utt}}; \overleftarrow{h_{i,1}^{utt}}], \tag{4.5}$$

$$h_i = GRU_{cxt}\left(h_{i-1}, h_i^{utt}\right). \tag{4.6}$$

The bidirectional nature of the BiGRU allows the model to capture both preceding and succeeding contextual information for each word, which is crucial in medical dialogue systems, where accurate interpretation often depends on both prior and subsequent tokens in an utterance. By processing the input utterances bidirectionally, the BiGRU encodes richer, more informative representations that facilitate improved downstream performance in generating contextually appropriate responses.

4.2.4 Knowledge-based Conditional Wasserstein Auto-Encoder

The Knowledge-based Conditional Wasserstein Auto-Encoder (CWAE) is utilized to generate responses by modeling both the contextual and external knowledge influences. This encoder excels in capturing the variability and complexity of human language, particularly in medical dialogues, where responses are heavily influenced by both the dialogue history and external medical knowledge, such as searched documents or patient-specific infor-

mation. The CWAE consists of several interconnected networks, including the Recognition Network, Prior Network, Knowledge Network, and Discriminators, which is introduced below.

Recognition Network

The Recognition Network is responsible for modeling the posterior distribution $q_\theta(z|x, r, K)$. It receives as input the concatenation of hidden states h_x, h_r, and h_k, which represent the context, the response, and the knowledge, respectively. The network outputs the mean μ and the logarithm of the variance $\log \sigma^2$ of a Gaussian distribution.

The random noise ϵ is sampled from this Gaussian distribution using the re-parameterization trick, allowing for efficient gradient propagation during training. The output of the Recognition Network is then fed into the generator Q to produce the latent variable z. The procedure for obtaining ϵ is defined in Equation 4.7.

$$z = Q_\theta(\epsilon),\ \epsilon \sim \mathcal{N}(\epsilon; \mu, \sigma^2 I),\ \begin{bmatrix} \mu \\ \log \sigma^2 \end{bmatrix} = W g_\theta \left(\begin{bmatrix} h_x \\ h_r \\ h_k. \end{bmatrix} \right) + b. \quad (4.7)$$

Prior Network

The Prior Network models the context-dependent prior distribution $p_\phi(z|c)$. It takes as input the context representation h_c and produces the latent variable $\tilde{z}$. Similar to the Recognition Network, it also outputs the mean and variance parameters from which random noise $\tilde{\epsilon}$ is sampled. The process is

detailed in Equation 4.8.

$$\tilde{z} = G_\phi(\tilde{\epsilon}),\ \tilde{\epsilon} \sim \mathcal{N}\left(\tilde{\epsilon}; \tilde{\mu}, \tilde{\sigma}^2 I\right),\ \begin{bmatrix} \tilde{\mu} \\ \log \tilde{\sigma}^2 \end{bmatrix} = \tilde{W} f_\phi\left(h_c\right) + \tilde{b}. \tag{4.8}$$

Knowledge Network

The Knowledge Network generates the document-dependent prior distribution $p_\omega(z|K)$. It processes the knowledge representation h_k to produce the latent variable $\hat{z}$. Similar to the previous networks, it outputs the mean and variance parameters for sampling the noise $\hat{\epsilon}$, as shown in Equation 4.9.

$$\hat{z} = M_\omega(\hat{\epsilon}),\ \hat{\epsilon} \sim \mathcal{N}\left(\hat{\epsilon}; \hat{\mu}, \hat{\sigma}^2 I\right),\ \begin{bmatrix} \hat{\mu} \\ \log \hat{\sigma}^2 \end{bmatrix} = \hat{W} f_\omega\left(h_k\right) + \hat{b}. \tag{4.9}$$

Discriminators

Two adversarial discriminators, denoted as D_x and D_k, are introduced to align the approximate posterior distribution with the prior distributions from the Prior Network and Knowledge Network. These discriminators are implemented as feed-forward neural networks and are trained to distinguish between posterior samples and prior samples.

The discriminator loss for the context discriminator D_x is defined as follows:

$$\mathcal{L}_{\text{disc\_ctx}} = E_{\epsilon \sim \text{RecNet}(h_x, h_r, h_k)}[D_x(Q(\epsilon), h_x)] - E_{\tilde{\epsilon} \sim \text{PriNet}(h_x)}[D_x(G(\tilde{\epsilon}), h_x)]. \tag{4.10}$$

Similarly, the discriminator loss for the knowledge discriminator D_k is given

by:

$$\mathcal{L}_{\text{disc\_dc}} = E_{\epsilon \sim \text{RecNet}(h_x, h_r, h_k)}[D_k(Q(\epsilon), h_x)] - E_{\hat{\epsilon} \sim \text{KgNet}(h_k)}[D_k(M(\hat{\epsilon}), h_x)]. \tag{4.11}$$

The overall objective function of the CWAE incorporates these discriminators and aims to minimize the divergence between the posterior and prior distributions while maximizing the log-probability of reconstructing the response r from the latent variable z. This objective is illustrated in Equation 4.12.

$$\min_{\theta,\phi,\psi,\omega} -E_{q_\theta(z|x,r,K)} \log p_\psi(r|z,x) + W(q_\theta(z|x,r,K) \| p_\phi(z|x)) + W(q_\theta(z|x,r,K) \| p_\omega(z|K)). \tag{4.12}$$

During training, the GRU decoder reconstructs the response r using the concatenation of z and h_x as the initial hidden state. The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = -E_{z=Q(\epsilon), \epsilon \sim \text{RecNet}(h_x, h_r, h_k)} \log p_\psi(r|z,x). \tag{4.13}$$

At inference time, the model generates latent variables $\tilde{z}$ and $\hat{z}$ based on the context utterances and searched documents. An average sum is then applied to these latent variables, which is combined with h_x to form the initial hidden state of the GRU decoder for response generation.

4.2.5 Memory Recurrence

In medical dialogue, the speaking behaviors between patients and doctors are different. The doctor's utterance often contains abbreviations, profes-

sional medical terms, whereas the patient's utterance often contains ambiguous, informal text, and patient usually plays a questioner role in the medical dialogue. Traditional dialogue generation models normally ignore this or use simple binary role id to model this. The goal of the proposed framework is to design an architecture that effectively simulates the distinct speaking behaviors of patients and doctors. As mentioned in section 4.2.2, two different language models, L_p and L_d, are utilized with identical architectures to separately model the patient and doctor. These two models are connected to each other by a memory mechanism and are trained recurrently when generating the whole sequence of dialogue. Specifically, during the training phase, given a dialogue $c = \{p_1, d_1, p_2, d_2, ..., p_n, d_n, ..., p_N, d_N\}$, the L_p language model firstly predicts the p_1 utterance based on start utterance tokens s_0. When generating the p_1 utterance, the final decoder's hidden state $h_{p_1}^{\tilde{dec}}$ is taken by a linear transformation and the output $h_{p_1}^{dec}$ is stored as a memory state. $h_{p_1}^{dec}$ not only contains the p_1 utterance's information but also contains the searched document information and previous context's information. All this information is incorporated into the representation of the p_1 utterance during the next turn generation.

After predicting p_1 utterance, the language model L_d starts to predict d_1 utterance and the context x now is $\{p_1\}$. When generating the context representation, the average sum is applied to the p_1 utterance's last utterance encoder's hidden state and the memory state $h_{p_1}^{dec}$. Then this representation is fed into the context encoder to obtain h_x for the context branch.

Suppose the objective is to generate the n^{th} turn's doctor utterance; the context at this point includes $\{p_1, d_1, p_2, d_2, ..., d_{n-1}, p_n\}$. Memory states generated from previous turns are stored in a memory list M:

$$M = [h_{p_1}^{dec}, h_{d_1}^{dec}, h_{p_2}^{dec}, h_{d_2}^{dec}, ..., h_{d_{n-1}}^{dec}, h_{p_n}^{dec}]. \tag{4.14}$$

Each context utterance is initially encoded using the utterance encoder. Then, an average sum is performed involving each memory state and the utterance encoder's output. This summation serves as the final input for the context encoder. Consequently, the input for the context encoder mentioned in Section 4.2.3 undergoes a modification, as expressed in Equation 4.16 where j represents the j^{th} utterance in the context.

$$h^{utt} = [h^{utt}_{p_1}, h^{utt}_{d_1}, h^{utt}_{p_2}, h^{utt}_{d_2}, ..., h^{utt}_{d_{n-1}}, h^{utt}_{p_n}], \tag{4.15}$$

$$h_j = GRU_{cxt}\left(h_{j-1}, \frac{(h^{utt}_j + m_j)}{2}\right). \tag{4.16}$$

4.2.6 Training Details

The proposed language models L_p and L_d are trained iteratively until the whole sequence of dialogue is fully generated. Each language model is trained by alternating three phases: training auto-encoder by using reconstruction loss, training generators by gradient ascent on the two discriminator losses, and training discriminators by gradient descent on the two discriminator losses. When training discriminators, the training steps are repeated n_{dis} times. The detailed training procedure of the proposed model is shown in algorithm 1.

4.3 Experiments

4.3.1 Dataset

The proposed model is evaluated on two different medical dialogue datasets. The first one is collected from an online Chinese medical consultation web-

Input: Dialogue Corpus $C = \{c_1, c_2, ..., c_i, ..., c_{|C|}\}$,discriminator training iterations n_{dis}. Initialize L_p language model's parameters:$\{\theta^{L_p}_{uttEnc}, \theta^{L_p}_{cxtEnc}, \theta^{L_p}_{Dec}, \theta^{L_p}_{PriNet}, \theta^{L_p}_{RecNet}, \theta^{L_p}_{KgNet}, \theta^{L_p}_{M}, \theta^{L_p}_{Q}, \theta^{L_p}_{G}, \theta^{L_p}_{D_k}, \theta^{L_p}_{D_x}\}, \theta^{L_p}_{MLP}\}$. Initialize L_d language model's parameters:$\{\theta^{L_d}_{uttEnc}, \theta^{L_d}_{cxtEnc}, \theta^{L_d}_{Dec}, \theta^{L_d}_{PriNet}, \theta^{L_d}_{RecNet}, \theta^{L_d}_{KgNet}, \theta^{L_d}_{M}, \theta^{L_d}_{Q}, \theta^{L_d}_{G}, \theta^{L_d}_{D_k}, \theta^{L_d}_{D_x}, \theta^{L_d}_{MLP}\}$, document corpus B, number of discriminator iterations n_{dis}.

while model not convergence **do**
  Initialize C
  Sort C based on each dialogue's turn number
  while C has unsampled batches **do**
    Sample a batch $\{c_j\}_{j=1}^{N}$ from C with N dialogues
    Initialize context $\{x_j\}_{j=1}^{N}$ with start token s_0
    Initialize memory state m_j
    for each $i \in [1, 2, ..., |c_N|]$ **do**
      search top five documents k_j based on x_j from B
      set response r_j equal to $\{c_j[i]\}$
      if r_j is patient utterance **then**
        $L = L_p$
      else
        $L = L_d$
      end if
      Encode x_j,r_j and K_j. $h_{x_j} = cxtEnc^L(uttEnc^L(x_j), m_j)$, $h_{r_j} = uttEnc^L(r_j)$, $h_{k_j} = cxtEnc^L(uttEnc^L(K_j))$
      Generate z,$\tilde{z}$ and $\hat{z}$ by using equation 5 to 7
      Update $\theta^L_{uttEnc}, \theta^L_{cxtEnc}, \theta^L_{Dec}, \theta^L_{RecNet}, \theta^L_{Q}, \theta^L_{MLP}$ by gradient descent on reconstruction loss (equation 9)
      Update $\theta^L_{PriNet}, \theta^L_{RecNet}, \theta^L_{G}, \theta^L_{Q}$ by gradient ascent on $\mathcal{L}_{\text{disc\_ctx}}$ loss (equation 10)
      Update $\theta^L_{KgNet}, \theta^L_{RecNet}, \theta^L_{M}, \theta^L_{Q}$ by gradient ascent on $\mathcal{L}_{\text{disc\_dc}}$ loss (equation 11)
      Add r_j to context x_j
      Add $h^{dec}_{r_j}$ to memory state m_j
    end for
    for $n \in [1, ..., n_{dis}]$ **do**
      Repeat 5-22 by replacing the 18-20 with Update $\theta^L_{D_k}$ and $\theta^L_{D_c}$ by using gradient descent on $\mathcal{L}_{\text{disc\_ctx}}$ loss and $\mathcal{L}_{\text{disc\_dc}}$ loss.
    end for
  end while
end while

Algorithm 1: Training Procedure of the proposed model

site called Haodaifu. Each dialogue in the dataset contains the questions from the patients and the answers from the online physicians. The other dataset is a public medical dialogue dataset named MedDialog which is built by Zeng et al. (2020). Their original Chinese medical dataset contains millions of dialogues which are also collected from the Chinese medical question-answering forum. For each dataset, $20,000$ dialogues are randomly selected for training and $5,000$ dialogues for testing. Here is a dialogue example of the dataset.

A dialogue example of the proposed dataset.
Patient: Hello, doctor, I have high aminotransferases. can I take anlotinib? **Doctor**: I recommend you to do a liver protection treatment before taking it. **Patient**: Okay, can I do the liver protection treatment in a local hospital for infusion? What kind of medicine should I use. **Doctor**: It depends on the local hospital.

4.3.2 Document Selection

The above datasets only contain the patients' and doctors' utterances without any additional medical information. However, when the doctors reply to the patients' questions, their professional clinical knowledge is used. In other words, clinical knowledge is helpful for generating the doctors' replies. Therefore, a Chinese medical book series named "Medical Guidance Books for Clinical Doctor Qualification Examination" is used as additional knowledge. The book series is the official guidebook for the National Examination for Physicians License in China. The books contain comprehensive medical knowledge and almost cover all aspects of the clinical domain. OCR techniques are first employed to convert the original textbook into electronic text. Then the book is divided into multiple paragraphs; each paragraph is annotated with its chapter name, and section/sub-section names. Table

4.1 illustrates the examples of the extracted paragraphs with their chapter name and section names. A database is built using Apache Lucene (Białecki et al., 2012) to store these paragraphs. During the training of the proposed model, the input context utterances are flattened into a query sequence to retrieve the top five related paragraphs from the textbook. Each of these retrieved paragraphs is considered a document in the proposed framework. These documents are then fed into the searched knowledge branch of the model to enhance response generation with domain-specific clinical context.

Chapter	**Level-1 section**	**Level-2 section**	**Level-3 section**	**Level-4 section**	**Content**
Anatomy	Locomotor system	Osteology	Classification of bones	Short bone	The short bones are cuboidal in shape and are mostly found in clusters in areas that require both stability and flexibility of movement, such as the carpal bones and tarsal bones.
Physiology	Digestion and Absorption	Gastric Digestion	Nature, main components, and functions of gastric juice	Cauda Composition of gastric juice	The main components of gastric juice include water, hydrochloric acid, pepsinogen, mucus, bicarbonate, and intrinsic factor.
Medical Microbiology	Diagnostic methods for viral infections and respiratory viruses	Respiratory viruses	Rubella virus	Principles of prevention and treatment	Vaccination with attenuated rubella vaccine is the primary measure for preventing rubella. Typically, the measles, mumps, and rubella (MMR) trivalent vaccine is used. The target population for vaccination is primarily young women of childbearing age, especially those who have not yet married. Rubella vaccination is contraindicated for pregnant women.
Internal Medicine	Pulmonary arterial hypertension and pulmonary heart disease	chronic car pulmonale	Pathogenesis	Cardiac changes	Due to pulmonary arterial hypertension, right ventricular hypertrophy and right heart failure can occur. In some patients, late stage left ventricular hypertrophy and left heart failure may also develop.

Table 4.1: Example of the structuralized medical guidance book.

4.3.3 Experimental Settings

The Pytorch library (Paszke et al., 2019) is utilized to implement the proposed model. The number of hidden unit for all GRU encoders is set to 300. Whereas the decoders' hidden units are set to 500. The prior, knowledge and recognition networks are two-layer feed-forward neural networks with *tanh* activation function which contains 200 hidden units. All the generators and discriminators are three-layer feed-forward neural networks with ReLU activation function. The number of hidden units in generators is 200 and the number of hidden units in generators is 400. The hidden size of latent variable z, $\hat{z}$ and $\tilde{z}$ is 200. All fully connected layers' parameters are initialized from the uniform distribution $[-0.02, 0.02]$. Gradient penalty is used when training the discriminators and the lambda hyper-parameter is set to 10. The maximum utterance length is set to 40. The Chinese word2vec embedding (Li et al., 2018) is used to initialize the embedding layer and the embedding size is 300. The Chinese vocabulary size is $40,000$. The "Jieba" tokenizer (Sun, 2020) is utilized to tokenize the input Chinese utterances. The batch size is set to 32 and the number of training epoch is set to 150. SGD is used to train the auto-encoder and decoder with 1.0 initial learning rate. For the Haodaifu dataset, the learning rate is decayed by 20% every 10 epochs. For the MedDialog dataset, the learning rate is decayed by 40% every 30 epochs. For the prior network, a GMM mechanism is used, as proposed by (Gu et al., 2019), with the number of prior modes set to 5. The number of discriminator iterations n_{dis} is set to 5. The RMSprop is used in training the generators and discriminators, whose learning rates are set to $5e-5$ and $1e-5$ respectively. During the decoding phase, greedy decoding method is used to generate the response.

During the testing phase, 10 responses are sampled for each context. The

evaluation metrics employed include BLEU score, Bag-of-Words (BOW) embedding, inter-distinct, and intra-distinct scores. The BLEU score (Papineni et al., 2002) calculates the n-gram overlap score between the generated response and the reference utterance. In this chapter, the BLEU-n ($n < 4$) score is calculated with the smoothing technique 7 (Chen and Cherry, 2014). Following the work of Gu et al. (2019), for the sampled 10 responses, the average BLEU-n score is defined as precision, and the maximum BLEU-n score is defined as recall. The harmonic mean of the precision and recall is defined as the F1 score. The BOW embedding (Liu et al., 2016b) measures the cosine similarity between the words in the generated responses and the reference utterance. There are three metrics to calculate the BOW embedding score: average method, which calculates the similarity score between the averaged word embeddings of the two utterances; extrema method, which calculates the similarity score between the largest extrema value among the two utterances' word embeddings; greedy method, which greedily match the words in two utterances based on the cosine similarity scores of their word embeddings and the scores are finally averaged across all words. The distinct-n reflects the degree of diversity of the generated responses. It calculates the ratio of distinct unigrams (distinct-1) and bi-grams (distinct-2) over all unigrams/bi-grams in the generated response utterance. Following Gu et al. (2019), the distinct value for each sampled response is defined as intra-dist and the distinct value for all sampled responses is defined as inter-dist. Note that this chapter is interested in creating a smart medical dialogue framework that can better reply patient's medical related questions. Consequently, the following tables present only the model performance for the doctor's language model.

To evaluate the effectiveness of the language model, performance compar-

isons are made against the following baseline models:

HRED This method (Serban et al., 2016) is a general seq2seq model. It contains a hierarchical RNN encoder to encode the utterance-level information and cotext-level information of the input sentences.

VHRED This method (Serban et al., 2017) adds an additional element to the HRED model. It adds a stochastic latent variable upon the context encoder as an additional input of the decoder, which increases the models' variability.

VHCR This method (Park et al., 2018) uses a hierarchical latent variable structure based on the VHRED model.

CVAE Zhao et al. (2017) utilizes a conditional VAE model with KL-annealing and bag-of-word loss.

CVAE_CO A collaborative variational encoder-decoder model with two learning phases proposed by (Shen et al., 2018).

DialogWAE This method is a conditional Wasserstein autoencoder which is proposed by Gu et al. (2019).

Dior-CVAE This method (Yang et al., 2023) is an innovative variational dialog model that incorporates a diffusion model to enrich the prior distribution. It employs the pre-trained language model BART (Lewis et al., 2020) to infer the posterior and likelihood distributions within the CVAE framework.

4.3.4 Ablation Studies

Three variants of the proposed technique are designed to evaluate the effectiveness of the two language models and the knowledge branch. The first model is the proposed technique without memory state. The second model is the proposed technique with one language model. The knowledge branch is added in this variant. The third model uses the two language models without the knowledge branch. The final one is the proposed technique. Table 4.2 shows the experimental results of these models. The proposed technique attains the highest BLEU scores. In addition, the proposed technique attains the highest intra-dist scores, signifying its ability to generate more diverse words in the sampled responses. However, this may reduce the BOW embedding score since the BOW embedding score represents the semantics relationship between the generated response and the reference. Generating more diverse words in the sampled responses may cause a reduction of the BOW embedding score. In terms of the inter-dist score. The model shows suboptimal performance when compared to the model without knowledge. This could be due to the incorporation of a searched document branch into both language models, resulting in forcing the model to generate the responses that are related to the searched documents, which may potentially limit the model's capacity to generate diverse responses.

	BLEU			**BOW Embedding**			**Intra-dist**		**Inter-dist**		
Model	Recall	Precision	F1	Average	Extrema	Greedy	dist-1	dist-2	dist-1	dist-2	avg_len
Ours (w/o memory state)	0.176	0.121	0.144	0.851	0.583	0.790	0.762	0.910	0.407	0.709	**22.651**
Ours (w/o two models)	0.176	0.118	0.141	0.846	**0.589**	0.799	0.730	0.891	0.362	0.678	20.066
Ours (w/o knowledge)	**0.185**	0.106	0.135	0.849	0.584	**0.813**	0.784	0.937	**0.450**	**0.769**	16.524
Ours	0.180	**0.127**	**0.149**	**0.853**	0.586	0.796	**0.807**	**0.960**	0.372	0.640	19.985

Table 4.2: Ablation studies evaluating the effectiveness of searched document branch and two models in the proposed framework on Haodaifu dataset.

	BLEU			BOW Embedding			Intra-dist		Inter-dist		
Model	Recall	Precision	F1	Average	Extrema	Greedy	dist-1	dist-2	dist-1	dist-2	avg_len
Ours (w/o memory state)	0.186	0.128	0.152	0.870	0.597	0.814	0.793	0.957	0.348	0.660	19.200
Ours (w/o two models)	0.183	0.118	0.144	**0.873**	**0.603**	**0.818**	0.699	0.875	0.379	0.712	**23.746**
Ours (w/o knowledge)	**0.194**	0.123	0.151	0.863	0.588	**0.818**	0.723	0.892	0.398	**0.727**	19.988
Ours	0.189	**0.135**	**0.157**	0.866	0.592	0.804	**0.753**	**0.900**	**0.408**	0.703	22.882

Table 4.3: Ablation studies evaluating the effectiveness of searched document branch and two models in the proposed framework on MedDialog.

4.3.5 Comparisons to Baselines

Tables 4.4 and 4.5 show a comparison of the proposed framework with baseline models on two datasets. These results indicate that the proposed model significantly outperforms the baseline models on both datasets concerning the BLEU score, even when the knowledge branch is omitted. A similar trend is observed with the BOW embedding metric across both datasets, reinforcing the model's effectiveness in generating coherent and contextually relevant responses. In terms of the intra-dist scores, the VHRED and Dior-CVAE perform well on the Haodaifu dataset, while VHCR and VHRED perform well on the MedDialog dataset. However, the BLEU score and BOW embedding for these models are much lower than the proposed method. This suggests that they may generate irrelevant and meaningless words that increase their intra-dist scores. In terms of the inter-dist scores, the proposed method also outperforms the baseline models. The model without the knowledge branch achieves the highest score, as explained in Section 4.3.4. The DialogWAE model is the second-best model, achieving high scores in terms of the BOW embedding and inter-dist score. This suggests that using Wasserstein autoencoders can generate more diverse responses that are related to the reference utterances. Notably, despite utilizing a large pre-trained language model as both encoder and decoder, the Dior-CVAE model falls short in BLEU F1 score and BOW embedding compared to both the proposed model and the DialogWAE model. This disparity may stem from the large pre-trained language model generating

irrelevant words unrelated to specific medical domain, thereby diminishing its BLEU score and BOW embedding performance.

	BLEU			BOW Embedding			Intra-dist		Inter-dist		
Model	Recall	Precision	F1	Average	Extrema	Greedy	dist-1	dist-2	dist-1	dist-2	avg_len
HRED	0.120	0.120	0.120	0.788	0.501	0.750	0.819	0.988	0.082	0.099	15.757
VHRED	0.151	0.121	0.134	0.818	0.545	0.785	0.826	**0.991**	0.171	0.259	15.307
VHCR	0.142	0.114	0.127	0.809	0.528	0.767	0.820	0.990	0.160	0.233	15.923
CVAE	0.149	0.075	0.100	0.831	0.530	0.804	0.603	0.752	0.256	0.426	22.664
CVAE_CO	0.145	0.121	0.132	0.808	0.541	0.795	0.671	0.834	0.120	0.186	19.058
DialogWAE	0.178	0.105	0.132	0.847	0.585	0.809	0.804	0.906	0.416	0.736	16.634
Dior-CVAE	0.144	**0.144**	0.144	**0.899**	0.526	0.649	**0.834**	0.967	0.010	0.063	**134.85**
Ours (w/o knowledge)	**0.185**	0.106	0.135	0.849	0.584	**0.813**	0.784	0.937	**0.450**	**0.769**	16.524
Ours	0.180	0.127	**0.149**	0.853	**0.586**	0.796	0.807	0.960	0.372	0.640	19.985

Table 4.4: Evaluation results on Haodaifu dataset.

	BLEU			BOW Embedding			Intra-dist		Inter-dist		
Model	Recall	Precision	F1	Average	Extrema	Greedy	dist-1	dist-2	dist-1	dist-2	avg_len
HRED	0.134	0.134	0.134	0.795	0.497	0.755	0.775	0.942	0.077	0.094	19.213
VHRED	0.166	0.130	0.146	0.848	0.564	0.788	0.837	**0.991**	0.182	0.274	14.565
VHCR	0.152	0.123	0.136	0.833	0.538	0.791	0.832	**0.991**	0.154	0.217	14.490
CVAE	**0.194**	0.087	0.120	0.862	0.581	0.806	0.720	0.908	0.348	0.598	15.215
CVAE_CO	0.163	0.105	0.128	0.826	0.486	0.766	**0.853**	0.970	0.085	0.097	16.978
DialogWAE	0.189	0.125	0.150	0.873	**0.603**	0.816	0.713	0.870	0.352	0.661	21.862
Dior-CVAE	0.140	**0.140**	0.140	**0.876**	0.498	0.639	0.802	0.958	0.011	0.056	**145.819**
Ours (w/o knowledge)	**0.194**	0.123	0.151	0.863	0.588	**0.818**	0.723	0.892	0.398	**0.727**	19.988
Ours	0.189	0.135	**0.157**	0.866	0.592	0.804	0.753	0.901	**0.408**	0.703	22.882

Table 4.5: Evaluation results on MedDialog dataset.

4.3.6 Case Study

Table 4.6 presents examples of responses generated by the DialogWAE model and the proposed model on the Haodaifu dataset. The proposed model demonstrates stronger performance in terms of relevance, diversity, and fluency. In the first example, the patient inquires about the surgical procedure being minimally invasive and its cost. The proposed model directly addresses these questions in its first response. Additionally, the second and third responses provide the patient with a predicted price for the procedure. In contrast, the DialogWAE model simply suggests that the patient undergo examinations, and it sometimes generates responses that are inappropriate for a doctor, such as asking about the cost of the surgery. By utilizing two separate language models to simulate the roles of the patient and doctor, the model can avoid such situations to some extent.

In the second example, the response from the DialogWAE model still revolves around requesting the patient to undergo examinations or offering a brief suggestion to undergo surgery. In contrast, the proposed model incorporates more medical knowledge, such as suggesting the patient undergo an MRI test and providing guidance on routine blood indexes. These examples illustrate the improved ability of the proposed framework to produce contextually relevant, linguistically diverse, and fluent responses.

Context	**DialogWAE**	**Ours**
Patient: Facial nerve spasms. Doctor: Hello, the cause of facial spasms is due to vascular compression of the facial nerve. If you want to cure it, microvascular decompression surgery is required. I suggest you come to our hospital for treatment. Patient: How much does the surgery cost? Is it a minimally invasive procedure?	1. I recommend undergoing a comprehensive examination at a local hospital.	1. The surgery is a minimally invasive procedure. You can undergo interventional treatment at a local hospital.
	2. What examinations did you do after the surgery?	2. Generally, several surgeries are required, and the cost is about 30,000-50,000 yuan.
	3. What is the cost of the surgery?	3. If the surgery is performed in Shanghai, the cost is estimated to be around 4,000 yuan.
Patient: The patient is a 6-year-old female with previously good overall health. She is currently experiencing headache and fever. Doctor: Has the patient visited a hematologist? Patient: The patient underwent two lumbar punctures and was told that there was no issue. However, the laboratory test results show a white blood cell count of 240,000, abnormal C-reactive protein levels, and an erythrocyte sedimentation rate (ESR) of 59. The doctor only mentioned that there is an infection.	1. Recommend surgery.	1. Is it possible to have a cranial magnetic resonance imaging (MRI) done locally? Are blood pressure, blood sugar, and blood lipid levels normal?
	2. Then it is best for you to undergo a comprehensive examination at a local hospital.	2. What medication is the child currently taking? Are the blood pressure, blood sugar, and blood lipid levels normal?
	3. What examinations did you do at the local hospital?	3. I haven't seen the child's imaging results. If it is cerebral white matter dysplasia, surgical treatment can be considered.

Table 4.6: Examples of the generated responses from DialogWAE model and the proposed model on Haodaifu dataset.

4.3.7 Human Evaluation

Human evaluation was conducted by randomly selecting 50 dialogues from the MedDialog dataset. For each dialogue context, 10 responses are generated. These generated responses are then evaluated by three human annotators who were given specific instructions to assess the models' performance in relevance, diversity, and fluency. The annotators assigned a score between 1 and 5 for each criterion, where 5 indicates the best performance and 1 indicates the poorest performance. To maintain impartial judgment, the annotators are intentionally kept unaware of the names of each model. The evaluation results, depicted in Table 4.7, demonstrate that the proposed method surpasses the performance of the baseline models across these three evaluation criteria.

Model	Relevance	Diversity	Fluency
VHRED	2.48	1.93	2.88
VHCR	2.13	1.65	2.95
CVAE	1.93	2.53	2.46
DialogWAE	2.52	3.11	2.53
Dior-CVAE	2.62	1.51	2.49
Ours	**3.15**	**3.45**	**3.03**

Table 4.7: Human evaluation results on 50 samples in MedDialog dataset.

4.4 Conclusion

In this chapter, a novel framework is proposed that leverages two language models, employing the same architecture, to separately represent the roles of doctors and patients. This framework incorporates Wasserstein auto-encoders to model context-dependent priors, knowledge-dependent priors, and posterior distributions, while utilizing a memory state to establish a connection between the two language models. By doing so, the proposed

framework effectively captures the speaking patterns of both roles within patient-doctor conversations. Moreover, it utilizes external knowledge from narratives to generate responses that are both diverse and coherent.

Experimental results obtained on the Haodaifu dataset and the MedDialog dataset demonstrate the superior performance of this framework compared to baseline models across various evaluation metrics. The margin of improvement is substantial, showcasing the capability of the proposed framework. Future research will explore the application of this method on large pre-trained models. Additionally, more complex techniques for incorporating expert knowledge, such as Knowledge Graphs, will be investigated to further enhance the framework's performance.

Chapter 5

Classifying Social Support in Physician Text Using a Rule-Enriched Attention-Based Deep Neural Network

The dialogue generation model introduced in chapter 4 aims to facilitate meaningful interactions between patients and healthcare providers by generating contextually relevant medical dialogues. This capability sets the stage for the subsequent exploration of social support classification in teleconsultation settings. In this chapter, the focus shifts to classifying predefined types of social support offered by physicians during these interactions. Understanding these types of support is crucial for analyzing how various physician communication strategies influence patient engagement and satisfaction. By accurately categorizing social support, deeper insights can be

gained into the dynamics of team-based teleconsultations and their impact on patient outcomes.

5.1 Introduction

Text classification is a pivotal task in the field of Natural Language Processing (NLP), with substantial applications in the healthcare domain. An emerging trend is team-based teleconsultation, an online medical service that enables multiple physicians to communicate with patients through various digital platforms, including text messages, phone calls, and video conferences. Team-based teleconsultation brings substantial advantages to both physicians and patients. By allowing physicians of varying levels of experience to work together, this approach can help distribute the workload more evenly and lessen the pressure on individual doctors (Liu et al., 2020b). This setup enables patients to benefit from the collective expertise of several healthcare professionals. Furthermore, team-based teleconsultation can offer quicker responses than consultations handled by a single physician (Li et al., 2019b), as an individual doctor might struggle to manage the high volume of tasks and frequent interruptions that can delay timely responses to patient inquiries. While team-based teleconsultation has many benefits, it also faces significant challenges. Participation levels among leaders and team members can vary, and these groups are often viewed differently because of their different expertise and reputation. Research has shown that patient satisfaction may decrease when team members primarily handle consultations, threatening the sustainability of this approach. Understanding the impact of feedback from both leaders and team members is crucial for maintaining patient engagement.

This chapter focuses on classifying pre-defined types of social support provided in these teleconsultation interactions to understand and enhance patient engagement. Social support theory (Cobb, 1976) can be used to propose various effects of social support provided by leaders and team members. This support is categorized into three types: direct informational support, indirect informational support, and emotional support. The doctors' responses during the teleconsultation interactions are classified into these three categories. This approach is expected to aid in analyzing and categorizing communication between patients and healthcare providers. Accurate classification of the social support types can reveal information that contributes to a better understanding of how the doctors' responses influence patient satisfaction and engagement in team-based teleconsultations. Therefore, accurately classifying these support types is essential.

Traditional machine learning approaches, such as Support Vector Machines (SVM) and Latent Dirichlet Allocation (LDA) (Chen et al., 2020; Tan and Yan, 2020; Zhao et al., 2022), have been widely used for medical text classification. However, these methods rely heavily on manually engineered features, which require extensive and laborious feature engineering to achieve strong results (Minaee et al., 2021). They also struggle with large datasets because these manually crafted features are not well-suited to fully leverage the vast amounts of data (Minaee et al., 2021). While deep learning classifiers typically outperform traditional machine learning models when handling large datasets (Zhou et al., 2023), they often fall short in terms of interpretability and the incorporation of expert knowledge, making it difficult for human experts to fine tune to meet the high precision requirement of medical decision-making.

In this chapter, a novel Rules-enriched Attention-based Deep Neural Network (R-ADNN) approach is proposed for classifying social support types

in teleconsultation text. This innovative method integrates rule-based techniques, contextual information, and deep learning to enhance both the accuracy and interpretability of text classification. Specifically, 37 rules are established to categorize physician responses into 19 predefined domain labels. Subsequently, a deep-learning-based classification framework determines the social support type of each response. There are three branches in the proposed framework, the physician response branch, the contextual information branch, and the label branch. In the physician response branch and the contextual information branch, the text of physician responses and the subsequent neighboring patient responses are first embedded using a BERT model to capture contextual semantics. This is followed by further encoding of the sequential data with a Bi-LSTM model. In the contextual information branch, a word attention mechanism is applied to refine the Bi-LSTM outputs. Meanwhile, in the label branch, the domain labels assigned to the physician responses are encoded as one-hot vectors and passed through an embedding layer. A label attention module then aligns these label embeddings with the encoded vectors of the physician responses. The outputs from both the label attention and word attention modules are combined and fed into a Multi-layer Perceptron (MLP) for the final classification. This integrated approach, combining rule-based techniques with deep learning, enhances the interpretability and accuracy of the model. The proposed R-ADNN framework, evaluated on a real-world dataset, achieves superior performance compared to current state-of-the-art text classification models.

The main contributions presented in this chapter can be summarized as follows:

1. A novel text classification approach specifically designed for under-

standing teleconsultation interactions. This framework combines rule-based techniques with a deep learning approach to effectively classify medical text to social support types, utilizing a hybrid model that integrates domain-specific rules and advanced neural networks.

2. A dedicated lexicon for social support in team-based teleconsultation is developed based on multiple authoritative sources. This lexicon supports the rule-based component and fills a critical gap in resources for this domain.

3. A set of 37 domain-specific rules is created to assign physician responses to 19 predefined domain labels, enhancing the accuracy and interpretability of the initial classification. These rules incorporate expert knowledge and are essential for improving the precision of the subsequent deep learning-based classification.

4. The proposed framework is evaluated using a real-world online medical teleconsultation datasets, demonstrating superior performance compared to other baseline models.

5.2 Methodology

5.2.1 Problem Definition

The problem addressed in this chapter can be formally defined as follows: given the text-based interactions from a teleconsultation record, let R represent the physician's response, and P denote the contextual information derived from the consecutive neighboring patient response. The objective is to develop a framework that, given P and R, classifies the physician's response R into specific social support categories, including direct informa-

tional support, indirect informational support, and emotional support. It is important to note that a single physician response may correspond to multiple labels.

5.2.2 Overview

The overall architecture of the proposed framework is illustrated in Figure 5.1. This framework integrates both rule-based and deep learning methods to classify physician response text. The rule-based method takes the physician response R as input and generates a set of labels, which serve as the input for the deep neural network model. The model consists of three branches: the physician response branch (labeled as ①), the contextual information branch (labeled as ②), and the label branch (labeled as ③). The input to the physician response branch is the physician response text R, while the contextual information branch takes the contextual information P, which comprises the content of the consecutive neighboring patient responses related to the physician response R. The label branch receives the comprehensive set of labels produced by the preceding rule-based method. These labels, closely aligned with social support concepts, leverage domain expertise to prioritize crucial information, addressing the interpretability challenges often associated with deep neural networks (Chau et al., 2020).

For the physician response branch, the input R goes through a BERT-BiLSTM module with label attention where the label (one-hot representation) is the output of Label Branch. For the contextual information branch, the input P goes through another BERT-BiLSTM module with word attention. The outputs of the both branches are concatenated and fed to feed-forward network for the final classification. It is important to note that the framework is designed for binary classification; thus, a separate

identical model is employed for each social support type. Subsequent sections provide a detailed explanation of the BERT-BiLSTM module with attention mechanism or the two branches and the label generation process.

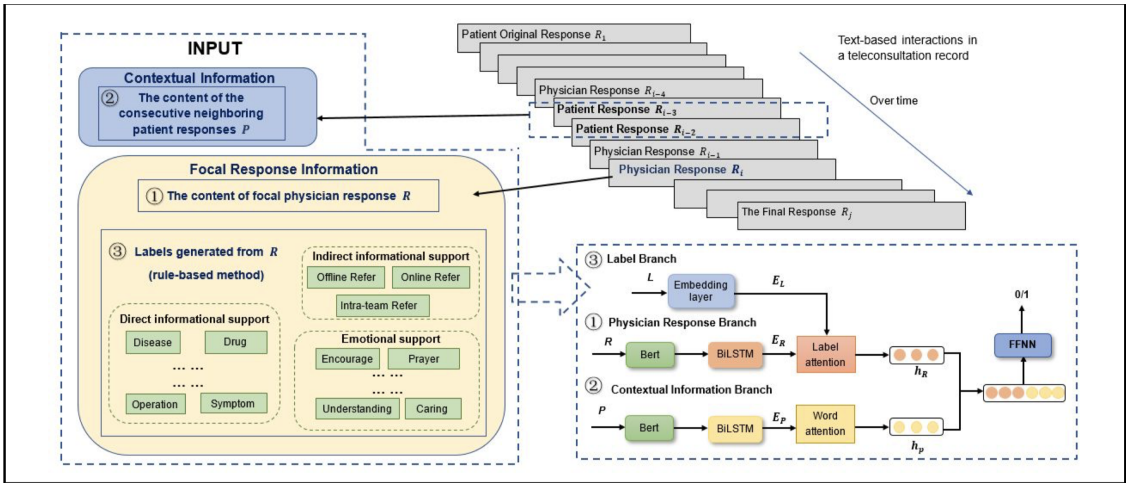


Figure 5.1: The structure of the R-ADNN approach.

5.2.3 BERT-BiLSTM Module with Attention Mechanism

BERT, developed by Devlin et al. (2018), stands as a prominent pre-trained model for natural language processing tasks. BiLSTM builds on the conventional LSTM model by incorporating contextual information from both past and future inputs, demonstrating excellent results in recent text-mining applications (Samtani et al., 2022; Zhou et al., 2016). The structure of the BERT-BiLSTM module is depicted in Figure 5.2. BERT comprises multiple Transformer encoder layers that transform the input text into token, segment, and positional embeddings. According to the method outlined in Devlin et al. (2018), the input sequence is prepared by placing a [CLS] token at the start and a [SEP] token at the end. The text is tokenized using a predefined dictionary, and the tokenized sequence is represented through a learnable embedding matrix. Positional and segment

details are embedded in the same way. These combined embeddings are passed through the self-attention layer, after which the outputs are normalized and sent to the final feed-forward neural network for classification tasks.

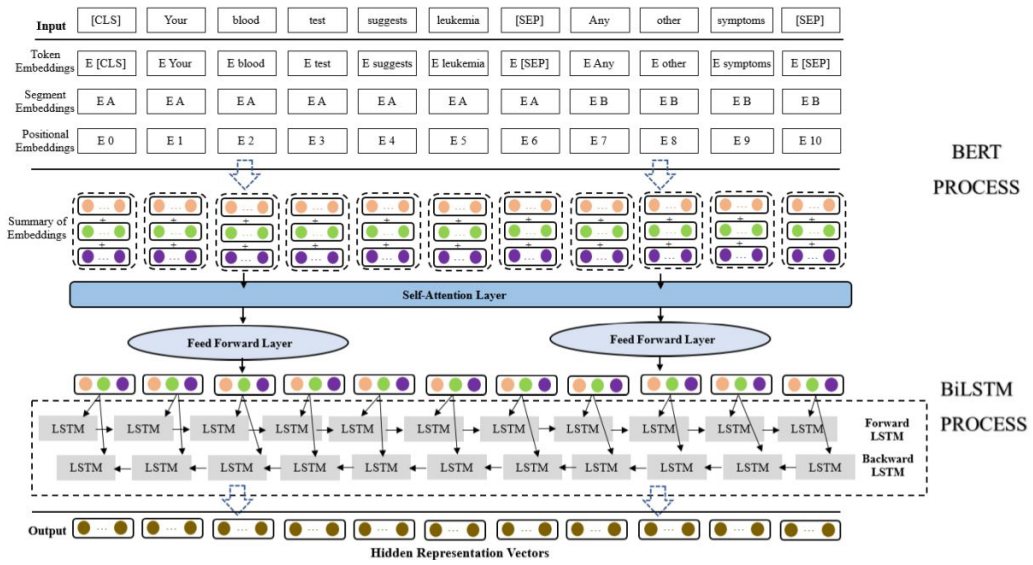


Figure 5.2: Architecture for the BERT-BiLSTM module.

Like many BERT-based methods, the outputs from BERT are passed into a BiLSTM model to capture the hidden representations of the original text. The BiLSTM processes the input bidirectionally, generating two sequences of hidden state vectors for each time step, one for the forward direction and one for the backward direction. These sequences are then combined to form the final hidden representations of the input. Both the Physician Response Branch and the Contextual Information Branch include the BERT-BiLSTM module; however, the BERT model is shared between the two branches, while the BiLSTM models are trained separately for each.

While BERT-BiLSTM treats all words in the input text with equal importance, it is essential to recognize that some words carry more weight depending on the context. To address this, an attention mechanism is introduced to dynamically adjust the weights of the hidden representations

based on their contextual relevance. In the physician response branch, a label-based attention module is implemented, combining the embeddings of domain-specific labels related to social support, denoted as E_L, with the hidden representations E_R. This combination serves as the input to the attention mechanism, as described below.

$$\alpha_i = \text{softmax}\left(\frac{E_{L_i} E_R^\top}{\sqrt{d}}\right) \tag{5.1}$$

$$h_{R_i} = \alpha_i E_R \tag{5.2}$$

$$h_R = \frac{\sum_i h_{R_i}}{|L|}, \tag{5.3}$$

where E_{L_i} denotes the embedding vector for the i^{th} label within the label set L, while E_R represents the hidden states derived from the BiLSTM output of the physician response branch. The attention weights assigned to the i^{th} label are indicated by α_i. The overall hidden representation that integrates all attentions for the physician response R is referred to as h_R, with h_{R_i} signifying the attention vector associated with the i^{th} label related to the physician response R. Once the social support-related label set L for the physician response is established, a one-hot vector is initially employed to encode these labels, which are then embedded using a learnable lookup matrix to calculate the corresponding embeddings E_L.

In the contextual information branch, a word-based attention module is developed to modulate the significance of each word from the patient responses that precede the physician's response in question. The context embeddings, E_P, are first processed through a single-layer fully connected neural network. A learnable vector v_p is subsequently utilized to compute similarity scores for the transformed embeddings, followed by the application of a softmax function to derive the attention weights α_p. The final hidden representation for the contextual information branch, h_p, is ob-

tained by summing the original embeddings E_P, weighted according to their respective attention values.

$$U_P = \tanh\left(W_p E_p^\top + b_p\right) \tag{5.4}$$

$$\alpha_p = \text{softmax}\left(v_p U_P\right) \tag{5.5}$$

$$h_p = \alpha_p E_P. \tag{5.6}$$

As illustrated in Figure 5.1, once the vectors h_R and h_P are obtained, they are concatenated and passed through a two-layer feed-forward neural network (FFNN) for final classification. The FFNN outputs a binary result (0 or 1), indicating whether the physician response includes a particular type of support. The network is trained using cross-entropy loss.

5.2.4 Label Generation

A rule-based approach is employed to create labels related to social support. Each physician response R is assessed against a predefined set of rules to determine its qualification for specific social support labels. The label set includes categories for direct, indirect, and emotional support, in line with established research (Bambina, 2007; Chen et al., 2019; Cutrona and Suhr, 1992). Additionally, several context-specific labels have been introduced to cater to the unique aspects of team-based teleconsultation. For instance, extending Bambina's concept of referral (Bambina, 2007), a label for indirect informational support, termed intra-team referral, has been incorporated to signify when a patient is referred to a specific team member. This label set was validated by six online healthcare experts, resulting in a total of 19 distinct labels. Examples of representative labels for the three

categories of social support are provided in Table 5.1.

Name of label	Relevance of the label to respective type of social support	Source
Direct Informational Support (a total of 10 labels)		
Symptom	A necessary step for a physician to make a diagnosis is to understand and analyze a patient's symptoms	Rakel (2022)
Drug	A typical therapy that a physician uses to prevent and combat a patient's disease is to prescribe some drugs.	Rakel (2021)
Indirect Informational Support (a total of 3 labels)		
Online refer	A physician may refer a patient to some online sources (e.g., a website) to get help.	Bambina (2007)
Intra-team refer	A physician may refer a patient to a designated team member to get help.	Li and Tong (2021)
Emotional Support (a total of 6 labels)		
Encouragement	To reduce a patient's negative emotions, a physician may express his encouragement by inspiring the patient to be positive and stop from indulging in negative emotions.	Cutrona and Suhr (1992)
Prayer	To inspire a patient's hope, a physician may express blessings for a patient's future improvement.	Cutrona and Suhr (1992)

Table 5.1: Representative labels for social support.

Two categories of rules have been developed for label assignment. The first category, based on the methodology proposed by Chau et al. (Chau et al., 2020), involves verifying whether the input R contains specific terms from the social support lexicon to assign labels. The second category encompasses more complex rules that analyze various word patterns using regular expressions. For instance, in the context of indirect informational support, a physician might suggest that the patient "visit xxx hospital for a more thorough examination." Identifying such cases necessitates recognizing a combination of a verb (e.g., "visit") and a noun (such as "hospital"), rather than relying solely on lexicon-based word mapping. In collaboration with two domain experts, a total of 37 rules were formulated. Representative rules are displayed in Table 5.2.

To the best of current knowledge, no existing lexicon has been specifically designed for team-based teleconsultation. As a result, a social support

Label	Corresponding identification rules
Direct Informational Support (a total of 17 rules)	
Symptom	whether the input *R* contains symptom-related words from our social support lexicon
Drug	whether the input *R* contains drug-related words from our social support lexicon
Indirect Informational Support (a total of 6 rules)	
Online refer	whether the input *R* contains a link; whether the input *R* contains keywords such as "website" or "app"
Intra-team refer	whether the input *R* contains "@"; whether the input *R* contains the combination pattern of "family name + designation +instruction" (e.g., "**Doctor Wang**, please **reply** to the patient's question")
Emotional Support (a total of 13 rules)	
Encouragement	whether the input *R* contains positive emotion-related words from our social support lexicon without negation-related words from lexicon; whether the input *R* contains both negative emotion-related words and negation-related words from our social support lexicon
Prayer	whether the input *R* contains prayer-related words from our social support lexicon

Table 5.2: Representative rules for social support.

lexicon was developed for this chapter, drawing from several authoritative sources. These include the Chinese version of the International Classification of Diseases, 11th revision (ICD-11), the Chinese Classification and Codes of Operations and Procedures (ICD-9-CM3), the Chinese LIWC sentiment lexicon, and a disease-centered entity database created by the Chinese Academy of Sciences[1]. Additionally, 14 medical research assistants were recruited to extract keywords from disease descriptions available on the target teleconsultation platform. This platform provides patients with comprehensive, accessible explanations of common diseases to promote patient education. Each disease entry contains informal language, symptoms, relevant medical tests, treatments, and non-medical advice, such as dietary or exercise recommendations. By manually extracting keywords, the lexicon was expanded to include informal terms frequently used by patients, which often differ from academic or medical terminology. For instance, patients might use the colloquial term "Alcoholic Nose (酒糟鼻)" instead of the formal term "Rosacea (玫瑰痤疮)." Finally, with guidance from

[1]https://github.com/liuhuanyong/QASystemOnMedicalKG

domain experts, the social support lexicon was organized into 18 distinct categories. Table 5.3 provides examples of these representative groups.

Group	Representative words	Sources
Direct Informational Support (a total of 7 groups)		
Symptom	bleeding(出血), pain(疼痛)	ICD-11 (Chapter 21); Classification and Codes of Disease (Chapter 18); A disease-centered entity database developed by Chinese Academy of Sciences[1]
Drug	amoxicillin(阿莫西林), ibuprofen (布洛芬)	ICD-11 (subset of Chapter X); National Essential Medicines Directory; A disease-centered entity database developed by Chinese Academy of Sciences
Indirect Informational Support (a total of 3 groups)		
Family Name	Li (李), Wang (王)	A widely used Chinese family name database[2]
Designation	doctor(医生),professor(教授)	HIT (Harbin Institute of Technology) synonyms lexicon
Emotional Support (a total of 8 groups)		
Positive Emotion	brave (勇敢),optimistic (乐观)	The Chinese version of LIWC lexicon; NTUSD Chinese sentiment lexicon (Ku and Chen 2007); Hownet Chinese sentiment lexicon (http://www.keenage.com/)
Prayer	pray(祈祷), bless(祝福)	Chinese emotional lexicon ontology (Xu et al. 2008)

Table 5.3: Representative groups and words of social support lexicon.

5.2.5 Binary Task Specialization

In the proposed approach, BERT is fine-tuned throughout the training process. Each support type is addressed by training a distinct model to predict whether the physician's response corresponds to that specific type. This approach transforms the multi-label classification problem into multiple binary classification problems. The binary Cross-Entropy loss function is applied in this context.

5.3 Dataset Collection

This chapter draws on data from a prominent teleconsultation platform in China, which connects over 887,000 physicians from approximately 9,900

hospitals across the country. In June 2017, the platform launched a team-based teleconsultation service. As depicted in Figure 5.3, each medical team has a dedicated page that showcases the team's name, leader, service volume, pricing, and profiles of the team leader and members. These profiles include names, clinical titles, and their affiliated hospitals or departments. The page also provides a record of previous teleconsultations. Figure 5.4 offers examples of patient teleconsultation records with a specific team. If a patient consults the same team multiple times, the platform arranges the records chronologically, based on the start of each consultation. Consultations are labeled either as "online check-in after offline visits" if the patient had previously seen a doctor in person, or simply as "text-based teleconsultation" for regular online interactions. By selecting the "detaila" button, users can access a full record containing the patient's gender, age, primary complaint, and timestamped text exchanges between the patient and physicians. To ensure privacy, personal identifiers and sensitive information such as names or medical test results are omitted from the records.

The raw dataset consists of 112,111 team-based teleconsultations involving 2,397 teams and 103,867 patients, collected between June 2017 and October 2020. To compile this dataset, links to all teams listed on the platform were first gathered, followed by the extraction of detailed information about each team and their previous consultations with patients. For each patient's interaction with a team, all available data, including physician-patient text exchanges, were retrieved as shown in Figure 5.4.

Consultations where the team did not respond, accounting for about 3% of the dataset, were removed. Additionally, consultations from the last three months of the data collection period, representing around 9% of the total, were excluded due to possible incomplete engagement data. In instances where multiple consultations were mistakenly merged into one record, cor-

Figure 5.3: An example of an online medical team.

rections were made. After these adjustments, the final dataset contains 115,845 team-based teleconsultations involving 2,307 teams and 93,629 patients.

To assess the effectiveness of the R-ADNN approach, an evaluation was carried out using a random sample of 1,000 team-based teleconsultation records, which included 5,847 physician responses. Three medical students annotated these responses. Since a single response may provide multiple types of social support (Chen et al., 2019), the annotators were instructed to evaluate all the three types of support rather than focusing on just one. In total, 2,734 responses were labeled as providing direct informational support, 1,038 as offering indirect informational support, and 310 as delivering

Figure 5.4: Screenshots of a patient's teleconsultation record with a team.

emotional support.

The annotated dataset was then divided into training and test sets, using a 70:30 split. All models were trained on the same training set and evaluated on the corresponding test set. Due to the relatively small number of responses identified as containing emotional support, which could potentially affect model performance, the training set was augmented with an additional 315 responses coded for emotional support.[2].

[2]An additional 1,000 teleconsultations were randomly selected to identify emotional support responses. Initially, a rule-based classification flagged these responses, which were then verified for accuracy by three coders.

5.4 Experimental Results

5.4.1 Experimental Settings

The framework was implemented using the PyTorch library, leveraging BERT-Base-Chinese as the pre-trained model, which consists of 12 transformer encoders, each with 12 multi-head attention heads. The transformers have a hidden dimension of 768, and the maximum input sequence length is capped at 512. For the Bi-LSTM, a single layer was employed with a hidden size of 256, while the MLP's hidden size was set to 2048. Training utilized the Adam optimizer with automatic mixed precision enabled to accelerate the process. A batch size of 10 was used, with a learning rate warm-up proportion of 0.1 and a weight decay coefficient of 0.01. The dropout rate was maintained at 0.1, and the model was trained for 10 epochs for efficiency.

The framework was compared against other models using the F1 score as the primary metric, supplemented by accuracy, precision, and recall for additional insights. The effectiveness of the proposed framework was assessed by evaluating its performance against various baseline models.

Rule Based In this baseline model, classification of the physician's response is done solely based on pre-defined rules, with no use of machine learning techniques.

Random Forest In this baseline model, the physician's input text is broken down into individual words and represented using the bag-of-words approach. A random forest algorithm is then applied for classification.

SVM In this baseline model, the physician's response text is represented using the same bag-of-words approach as previously described. A support

vector machine is then employed for classification, utilizing both linear and radial kernels.

BERT This baseline model consists of a basic vanilla BERT model followed by a single-layer feed-forward neural network as the classifier. The model only uses the physician's response text as input, and the hyperparameters used are identical to ours.

BERT+LSTM This baseline model combines BERT with Bi-LSTM, forming a simple yet commonly used architecture in many NLP tasks. The hidden states from different time steps in the Bi-LSTM are concatenated and then passed to a single-layer feed-forward network for final classification. The hyperparameters used in this model are identical to ours.

5.4.2 Comparisons to Baselines

Table 5.4 presents the performance evaluation results using metrics such as accuracy, precision, recall, and F-measure. From the table, it is evident that the R-ADNN approach outperforms the other baseline models. Specifically, R-ADNN achieves the highest accuracy rates: 94.3% for direct informational support identification, 98.1% for indirect informational support identification, and 96.4% for emotional support identification. It also records the highest F-measure scores: 93.9% for direct informational support, 94.2% for indirect informational support, and 82.1% for emotional support. Additionally, R-ADNN attains the highest precision and recall for both direct and indirect informational support. While R-ADNN does not achieve the top precision and recall for emotional support, its F-measure still surpasses that of the other models, indicating superior overall performance when both precision and recall are considered (Ebrahimi et al.,

2020). In summary, the R-ADNN approach demonstrates better performance compared to the other baseline models.

Social Support Classification	Model	Evaluation Metrics (%)			
		Accuracy	Precision	Recall	F1
Direct Informational Support	Rule-based	0.817	0.805	0.820	0.812
	Random Forest	0.718	0.854	0.502	0.633
	SVM (linear)	0.726	0.845	0.532	0.653
	SVM (radial)	0.736	0.824	0.578	0.679
	BERT	0.904	0.920	0.879	0.899
	BERT-BiLSTM	0.925	0.931	0.913	0.922
	R-ADNN	**0.943**	**0.956**	**0.923**	**0.939**
Indirect Informational Support	Rule-based	0.975	0.929	0.929	0.929
	Random Forest	0.900	0.814	0.555	0.660
	SVM (linear kernel)	0.899	0.882	0.487	0.628
	SVM (radial kernel)	0.903	0.817	0.578	0.677
	BERT	0.904	0.860	0.539	0.663
	BERT-BiLSTM	0.933	0.886	0.708	0.787
	R-ADNN	**0.981**	**0.954**	**0.935**	**0.944**
Emotional Support	Rule-based	0.852	0.394	0.872	0.543
	Random Forest	0.942	0.850	0.511	0.638
	SVM (linear)	0.939	0.814	0.511	0.627
	SVM (radial)	0.943	0.831	0.548	0.660
	BERT	0.912	0.962	0.133	0.234
	BERT-BiLSTM	0.944	0.849	0.537	0.658
	R-ADNN	**0.964**	0.823	0.814	**0.818**

Table 5.4: Social support classification performance.

5.5 Conclusions

In this chapter, the Rule-Enriched Attention-Based Deep Neural Network (R-ADNN) is introduced to classify social support types in team-based teleconsultation text. The proposed approach combines a rule-based module with advanced deep learning techniques, specifically leveraging BERT and BiLSTM models, along with attention mechanisms. The rule-based component contributes to the models's interpretability, as each prediction is transparently linked to domain-specific rules grounded in expert knowledge. While BERT and BiLSTM are inherently less interpretable, the attention mechanism offers partial insight into model decisions by high-

lighting important tokens of the input text. Together, this hybrid architecture enhances both the classification accuracy and the explainability of the model. By incorporating contextual patient responses and domain-specific knowledge, the R-ADNN framework is capable of detecting nuanced social support types often overlooked by traditional methods. Evaluation results show that the model consistently outperforms baseline methods, demonstrating its effectiveness in capturing the complexities of physician-patient communication in teleconsultation scenarios.

For future work, expanding the social support lexicon and refining the rule-based component could further improve the model's performance. Additionally, applying the R-ADNN framework to other domains or different languages could test its generalizability and robustness. Exploring alternative deep learning architectures might also enhance the model's ability to understand and classify social support in more varied and complex scenarios.

Chapter 6

Enhancing Medical Named Entity Recognition Through Prompt Learning and Relational Networks

Following the exploration of physician response classification in chapter 5, which identified different forms of social support in teleconsultations, the focus shifts in chapter 6 to a more granular task—Medical NER. While classifying physician interactions enhances the understanding of communication in telemedicine, accurately identifying medical entities in clinical text is equally critical for processing patient records and supporting decision-making. The proposed NER model, built on prompt learning and a relational network, excels at identifying and classifying complex medical terms by capturing the relationships between words and predefined prompts. This advancement further contributes to the automation of healthcare text processing, ensuring precise extraction of essential medical information for

better clinical outcomes.

6.1 Introduction

The rapid growth of digital health records and the proliferation of unstructured clinical data have led to an increasing demand for efficient techniques to extract valuable information from these vast repositories (Savova et al., 2010). NER is a crucial NLP technique that has shown great potential in identifying and classifying entities such as diseases, medications, procedures, and patient demographics from textual data (Kundeti et al., 2016; Bose et al., 2021). In healthcare, NER can significantly enhance the ability to mine actionable data, streamline clinical workflows, and support decision-making processes (Janowski, 2023; Borchert et al., 2022).

The healthcare domain presents unique challenges for NER techniques due to the complexity and variability of medical terminology, abbreviations, and the need for high accuracy for obvious reasons–medical errors can have critical consequences (Kundeti et al., 2016). Recent development in NER, such as deep learning-based models, has shown promising performance in addressing these challenges (Wu et al., 2017; Zhu et al., 2018; Liu et al., 2021a). These models are usually trained on extensive corpora of medical text and are capable of recognizing intricate patterns and relationships within clinical narratives and facilitating the extraction and categorization of medical entities (de Lima Santos et al., 2021; Chawla et al., 2021; Polignano et al., 2021).

Prompt learning in NLP leverages pre-trained deep learning models by crafting specific prompts to guide these general models to perform specialized tasks such as NER (Liu et al., 2023a). While prompt learning has

become popular in areas like text generation and question-answering (Hou et al., 2024; Cui and Li, 2024), adapting to NER, particularly in healthcare, remains relatively rare. The strength of prompt learning lies in its ability to adapt pre-trained models to new tasks with minimal task-specific data. This is extremely valuable in healthcare where annotated data is often very limited and expensive to obtain. Integrating prompt learning into deep learning models to solve the NER task can significantly enhance the performance of recognizing medical entities. This is done by providing structured input that aligns with the models' existing knowledge(He et al., 2023; Zhu et al., 2022).

This chapter introduces a novel framework that combines prompt learning with relational network technique to enhance medical NER. The relational network in this scenario is crucial in capturing and utilizing the relationships between prompts and medical terms. This is vital to disambiguate medical entities and their context. The approach enables the model to effectively handle synonyms, abbreviations, and context-dependent meanings, leading to more accurate and reliable entity recognition in medical text. The main contributions are as follows:

1. A medical NER framework is introduced that integrates prompt learning into pre-trained deep learning models, aiming to address the complexities of medical entities and the challenges imposed by limited annotated data.

2. A prompt position predictor and prompt type predictor with relational network is proposed to more effectively predict the start and end indices and type for recognized entities by capturing relationships between prompts and medical text.

3. The proposed framework is evaluated using a real-world medical dataset,

showing significant performance improvement compared to other baseline models.

6.2 Methodology

6.2.1 Problem Definition

The primary objective of the proposed technique is to address the challenge of NER from medical text. Formally, let Z represent a given input sentence, which comprises of several entities. Each entity is characterized by its position within the text, defined by its start and end indices, as well as by its type that categorizes the entity according to a predefined set of medical entity categories (e.g., diseases, symptoms, medications, etc.). The task is to develop a model capable of automatically identifying and extracting all the medical entities present in the sentence Z, including determining both their exact positions and corresponding types.

6.2.2 Overview

Figure 6.1 illustrates the proposed framework, which consists of several key modules: BERT, Bi-LSTM module, Prompt-Text Fusion module, position predictor, and type predictor. The input sequence S of the framework consists of k prompts (position token and type token) followed by the input medical text Z. Firstly, the BERT model processes the input sequence to produce hidden representations. The hidden representation of medical text is fed to the Bi-LSTM module to further capture the intricate sequential dependencies within the medical text. The output of the Bi-LSTM module and the prompt representation previously generated by BERT are then fed

into the prompt-text fusion module to generate the final representation of prompts. Then the position predictor takes the prompt position token representations and the medical text representations encoded by the Bi-LSTM module to generate the start and end indices of the recognized entities. The type predictor takes the type token representations and the medical text representations to classify each prompt into specific entity type.

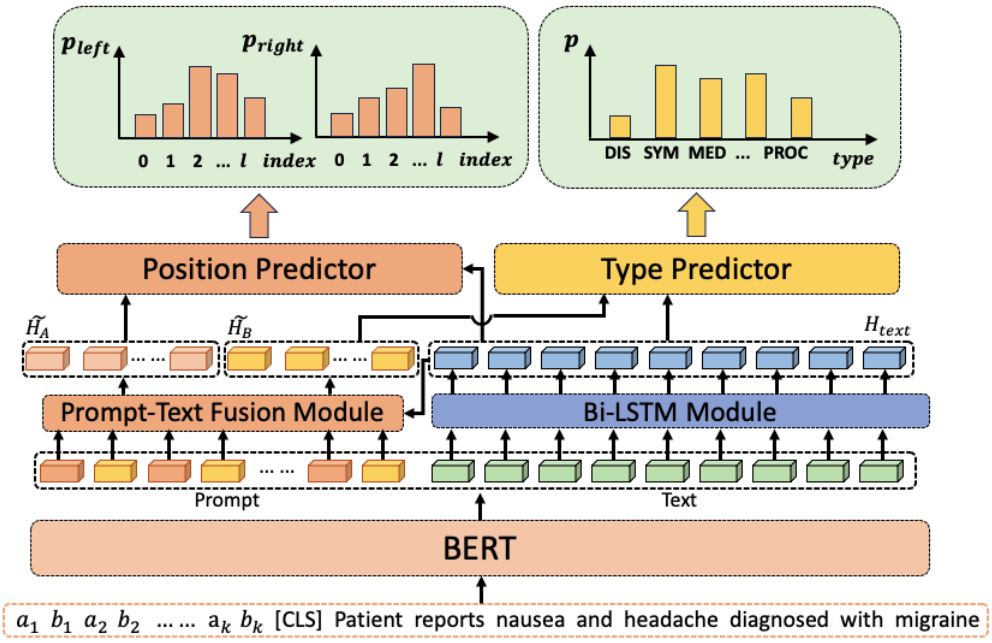


Figure 6.1: Overall architecture of the proposed framework.

6.2.3 Prompt Construction

Following the approach proposed by Shen et al. (2023), given an input medical text Z of length l, the input sequence $\mathbf{S}$ for the framework is constructed by concatenating k prompts with the input medical text Z. The input medical text Z is tokenized into a sequence of text tokens $\mathbf{z} = [z_1, z_2, \ldots, z_l]$. There are k prompts $\mathbf{P} = [p_1, p_2, \ldots, p_k]$ with each prompt p_i consists of a pair of tokens: a position token a_i and a type token b_i, where i ranges from 1 to k. The position tokens a_i are used to represent the positions of the potential entities in the text, while the type tokens b_i are

used to represent the entity types. In the implementation, the position and type tokens are represented by sparse BERT tokens, specifically the unused tokens from the BERT vocabulary, denoted as $[unused1]$ to $[unused100]$, as described in Shen et al. (2023). In addition, a special $[CLS]$ token is used to separate the prompt tokens from the input text. The final input sequence fed into the model is structured as follows:

$$\mathbf{S} = \{a_1, b_1, a_2, b_2, \ldots, a_k, b_k\}\ [CLS]\ z_1, z_2, \ldots, z_l. \tag{6.1}$$

6.2.4 Sequence Encoding

The sequence encoding process plays a fundamental role in the architecture by transforming the input into a feature vector that encapsulates its semantics. BERT Devlin et al. (2018), a widely adopted pre-trained language model, is employed as the initial embedding mechanism due to its strong ability to capture contextual information. BERT uses transformer encoders comprising self-attention layers, normalization layers, and feed-forward neural networks, and has demonstrated exceptional performance across various NLP tasks (Radford et al., 2018).

In the proposed framework, the input sequence $\mathbf{S}$ is initially encoded using BERT. This process is represented as:

$$\mathbf{H} = \text{BERT}(\mathbf{S}) = [h_1, h_2, \ldots, h_{k+l+1}], \tag{6.2}$$

where $\mathbf{S}$ consists of k prompt tokens and l input text tokens, with an additional $[CLS]$ token at the start. Each token's hidden state h_i has a dimension of d. To prevent the prompts from influencing the encoding of the medical text, an attention mask is applied during the encoding process.

Although BERT is highly effective at capturing token-level contextual relationships, transformer-based models often struggle with capturing distant token dependencies (Wang et al., 2020). To address this limitation, a Bi-LSTM module is incorporated to enhance the model's ability to capture sequential dependencies. While conventional RNNs are prone to vanishing gradient issues, particularly with long sequences, LSTMs (Hochreiter and Schmidhuber, 1997) mitigate this problem by employing memory cells and gates that selectively retain historical information.

Given the sequential nature of the medical text, the proposed framework utilize a Bi-LSTM module to further encode the hidden representations of the text portion of $\mathbf{S}$, leaving the prompts unaffected. The hidden representation encoded by BERT for the input medical text tokens $\{z_1, z_2, \ldots, z_l\}$ are defined as:

$$\mathbf{H}_{\text{input}} = [h_{k+1}, h_{k+2}, \ldots, h_{k+l}], \tag{6.3}$$

where $\mathbf{H}_{\text{input}}$ has a dimension of $d \times l$. The Bi-LSTM module, consisting of three layers of Bi-LSTM, processes these hidden representations to capture both forward and backward sequential dependencies. For simplicity, the following shows the process for a single Bi-LSTM layer:

$$\overrightarrow{\mathbf{H}} = \text{LSTM}_{\text{forward}}([h_{k+1}, \ldots, h_{k+l}]), \quad \overleftarrow{\mathbf{H}} = \text{LSTM}_{\text{backward}}([h_{k+1}, \ldots, h_{k+l}]). \tag{6.4}$$

The final representation of the input text is obtained by concatenating the forward and backward hidden states:

$$\mathbf{H}_{\text{text}} = [\overrightarrow{h_{k+1}} \oplus \overleftarrow{h_{k+1}}, \overrightarrow{h_{k+2}} \oplus \overleftarrow{h_{k+2}}, \ldots, \overrightarrow{h_{k+l}} \oplus \overleftarrow{h_{k+l}}], \tag{6.5}$$

where $\mathbf{H}_{\text{text}} \in \mathbb{R}^{l \times 2d'}$ represents the final encoded sequence for the input text, d' is the hidden size of the Bi-LSTM layers, and $\oplus$ denotes concate-

nation.

The BERT hidden representations of the position tokens $\{a_1, a_2, \ldots, a_k\}$ are defined as:

$$\mathbf{H}_A = [h_1, h_3, h_5, \ldots, h_{2k-1}], \tag{6.6}$$

where $\mathbf{H}_A$ has a dimension of $d \times k$ representing the k position tokens. Likewise, the hidden representations of the type tokens $\{b_1, b_2, \ldots, b_k\}$ are defined as:

$$\mathbf{H}_B = [h_2, h_4, h_6, \ldots, h_{2k}], \tag{6.7}$$

where $\mathbf{H}_B$ has a dimension of $d \times k$. The sequence encoding process ensures that both the medical text and prompt tokens are effectively encoded, preparing them for further processing.

6.2.5 Prompt-Text Fusion Module

The Prompt-Text Fusion Module is designed to enhance the interaction between prompt tokens and medical text by employing multi-head attention mechanisms. While BERT generates contextualized embeddings for both prompts and text, this module captures the deeper relationships necessary to accurate entity recognition.

The Prompt-text fusion module consists of three layers, each containing two multi-head attention. The first multi-head attention focuses on the relationships within the prompt tokens themselves, while the second focuses on the prompt tokens with the medical text. Multi-head attention enables the model to capture a wide range of dependencies across different parts of the sequence, making it highly effective for handling the complexity of entity relations in medical text.

The first multi-head attention of each layer applied to the position tokens is defined as follow:

$$\mathbf{H}_{A,\text{self-attn}} = \text{MultiHeadAttention}(\mathbf{H}_A + \mathbf{E}_{bind}, \mathbf{H}_A + \mathbf{E}_{bind}, \mathbf{H}_A), \quad (6.8)$$

where $\mathbf{E}_{bind}$ is a learnable embedding matrix to bind the position and type tokens for each prompt. The second multi-head attention of each layer applied to the position tokens is defined as follow:

$$\mathbf{H}_{A,\text{cross-attn}} = \text{MultiHeadAttention}(\mathbf{H}_{A,\text{self-attn}} + \mathbf{E}_{bind}, \mathbf{H}_{\text{text}}, \mathbf{H}_{\text{text}}). \quad (6.9)$$

Then the output of the second multi-head attention is fed to a feed-forward network to produce the final representation for the position tokens as follow:

$$\tilde{\mathbf{H}_A} = \text{FeedForward}(\mathbf{H}_{A,\text{cross-attn}}). \quad (6.10)$$

Similarly, the process is the same for the type tokens $\mathbf{H}_B$, resulting in the final type token representations $\tilde{\mathbf{H}_B}$. Importantly, the weights of these networks are not shared for the position and type tokens.

6.2.6 Position Predictor

The position predictor is responsible for predicting the exact location of the recognized entities. Specifically for each prompt the position predictor outputs the start and end position in the text. The position predictor consists of two linear transformation layers, which transform the position token representations $\tilde{\mathbf{H}}_A^{(i)}$ and the text representations $\mathbf{H}_{\text{text}}$. There is another linear layer with a Sigmoid activation function is used to generate

the probability distribution for the position as follows:

$$\mathbf{H}_{\text{Fusion}}^{(i)} = \text{Linear}(\tilde{\mathbf{H}}_{A}^{(i)}) + \text{Linear}(\mathbf{H}_{\text{text}}) \tag{6.11}$$

$$P_{\text{left}}^{(i,j)} = \sigma\left(\text{Linear}(\mathbf{H}_{\text{Fusion}}^{(i,j)})\right), \tag{6.12}$$

where $P_{\text{left}}^{(i,j)}$ represents the probability that the j-th word is the start of the entity predicted by the i-th prompt. The probability distribution for the end is calculated similarly.

6.2.7 Type Predictor

Similar to the position predictor, the type predictor is responsible for predicting the category of the recognized entities. A relational network as described in 3.2.4 is adopted to capture the interactions between the type tokens and the medical text, to improve the framework's ability to classify entity types for each prompt.

In the context of this chapter, strong semantic relationships exist between the prompts and the medical text. The relational network can effectively model these relationships, which is critical for accurate entity classification. Specifically, the relationships are captured as follows:

$$\mathbf{R}_i = \frac{1}{l}\sum_{j=1}^{l} g_\phi(\tilde{\mathbf{H}}_{B_i}, \mathbf{H}_{\text{text}_j}). \tag{6.13}$$

There is another linear layer with a Sigmoid activation function is used to generate the probability distribution for each type as follows:

$$\mathbf{P}_i = \sigma f_\theta(\mathbf{R}_i), \tag{6.14}$$

where $\mathbf{P}_i$ represents the type probability distribution of the i-th prompt.

During training, dynamic template filling as outlined in Shen et al. (2023) is used to optimize the model. In the inference phase, the start and end indices, and the type of the entity corresponding to the i-th prompt are determined by using argmax to select the highest probabilities. In cases where multiple prompts identify overlapping entities, only one is retained. If entities share the same start and end index but have conflicting types, the entity with the highest probability is selected.

6.3 Experiments

6.3.1 Dataset

HealthNER, a dataset Lee and Lu (2021) designed for NER in healthcare, is used for experiments. This dataset is assembled by collecting data from multiple sources, including healthcare information sites, online health news, and medical Q&A forums. There are ten distinct entity types relevant to healthcare, including body parts, symptoms, diseases, and medications. For instance, in the sentence "I do not know why every year there is a sudden and severe pain in my heart. Is this a problem with my body?", the annotated named entities are: "heart" (BODY), representing a body part; "pain" (SYMP), representing a symptom occurs twice; and "body" (BODY), another body part. The dataset also provides the start and end indices for these entities, along with their types. In total, the dataset consists of $28,161$ medical text samples for training and $2,531$ samples for testing.

6.3.2 Experimental Settings

The implementation used for experiments uses the PyTorch library (Paszke et al., 2019) and adopts the BERT-Base-Chinese as the pre-trained model. This model consists of 12 transformer encoders, each with 12 heads for multi-head attention. It has a hidden size of 768, with a maximum input sequence length of 512 tokens. Spare tokens are used to as the prompts, with the number of prompt set to 50. The Adam optimizer (Kingma and Ba, 2014) is used with automatic mixed precision to enhance the training speed. The learning rate warm-up proportion is set to 0.1, with a weight decay coefficient of 0.01 and a dropout rate of 0.5. The model is trained over 200 epochs.

The F1 score, precision, and recall (three standard metrics in NLP tasks) are used to compare the framework against other models. Precision measures the proportion of correctly identified entities out of all the entities predicted by the framework. Recall measures the proportion of correctly identified entities out of all the actual entities presented in the data. The F1 score is the harmonic mean of precision and recall, providing a balanced metric that accounts for both false positives and false negatives. A higher F1 score indicates better overall performance, balancing the trade-off between precision and recall.

6.3.3 Comparision to Baselines

The effectiveness of the framework is compared with several baseline models as follows:

BiLSTM-CRF This method is widely-used for sequence labeling, which combines BiLSTM networks with Conditional Random Fields (CRF) to

capture dependencies between labels in a sequence.

BERT BERT is a pre-trained transformer model that has achieved state-of-the-art performance in various NLP tasks. In the comparison, all the hyperparameters used in BERT are the same as the proposed approach.

ME-CNER This baseline, proposed by Xu et al. (2019a), is designed for Chinese NER tasks. The method creates character-level embeddings by integrating information at radical, character, and word levels, aiming to improve the performance of Chinese NER tasks.

Gazetteers This baseline, proposed by Ding et al. (2019), utilizes GNN with a multidigraph structure to integrate information from multiple gazetteers to for NER tasks.

Lattice The Lattice model, proposed by Zhang and Yang (2018), utilizes a lattice-structured LSTM model for Chinese NER, integrating both sequences of input characters and potential words from a given lexicon.

ME-MGNN This baseline, proposed by Lee and Lu (2021), is also specifically designed for Chinese NER in healthcare. The model incorporates embeddings at various granularities, including radical, character, and word levels, and utilizes multiple GNN to improve the recognition performance.

The performance comparison among the proposed method and several baseline models is summarized in Table 6.1. The proposed method achieves an F1 score of 76.37%, outperforming the best-performing baseline, ME-MGNN, by 0.9%. This result demonstrating the effectiveness of the proposed method in improving medical entity recognition, with the help of relational network. Even without the relational network, the proposed method still performs strongly, achieving an F1 score of 76.11%, higher than all the other baselines. Compared to traditional models like BiLSTM-

CRF and the basic BERT model, the proposed method shows a significant improvement in F1 score, by 6.7% and 3.4%, respectively. The incremental gains over more advanced baselines such as Gazetteers and Lattice further show improvement of the proposed approach, as a result of the use of the prompt-text fusion and relational network in capturing complex dependencies within the data.

Model	F1	Precision	Recall
BiLSTM-CRF	71.56	70.38	72.77
BERT	73.82	71.45	76.36
ME-CNER	74.15	73.68	74.62
Gazetteers	74.26	73.00	75.56
Lattice	75.22	74.69	75.76
ME-MGNN	75.69	75.46	75.76
Proposed method w/o relational network	76.11	**76.73**	75.50
Proposed method	**76.37**	76.24	**76.51**

Table 6.1: Evaluation results on HealthNER dataset.

6.4 Conclusion

This chapter presents a novel approach that combines prompt learning with relational networks to tackle the challenges of NER in healthcare. The approach leverages the strengths of prompt learning, which offers flexible and context-sensitive input transformations, alongside relational network that effectively captures intricate dependencies among entities and input text. The proposed approach significantly enhances the performance of the NER model, as demonstrated by experimental results. The method shows substantial improvement in accuracy and robustness across diverse and complex medical text, surpassing traditional NER techniques. Future work can focus on delving deeper into the integration of domain-specific knowledge, particularly through the incorporation of more comprehensive medical ontologies and structured data sources. Additionally, exploring

the scalability of this approach by applying it to larger and more diverse medical datasets will refine the model's ability to generalize across various subdomains within healthcare.

Chapter 7

Conclusions

In conclusion, this thesis harnesses the power of advanced deep learning techniques to address pressing challenges in hearthcare. This thesis effectively demonstrates the transformative potential of deep learning-based NLP techniques in automating critical processes for smart healthcare. By leveraging cutting-edge deep learning techniques, I have contributed to the development of multiple techniques to streamline healthcare operations. This thesis introduces a novel framework to classify patient's chief complaints, a multi-turn dialogue generation framework for simulating realistic patient-doctor interactions, a rule-enhanced deep learning model to categorize physician communication styles, and a prompt-learning approach for NER to extract critical medical information. These developments underscore a commitment to practical, scalable solutions that can significantly improve healthcare efficiency and patient outcomes. This thesis offers a comprehensive exploration of the motivations, challenges, and research contributions while providing a roadmap for future research that aims to further push the boundaries of smart healthcare.

7.1 Thesis Summary

The aim of the research presented in this thesis is to leverage advanced deep learning-based NLP techniques to address critical challenges towards smart healthcare. As outlined in Chapter 1, the research focuses on four key tasks: chief complaint text classification, medical dialogue generation, physician text classification, and medical named entity recognition. This chapter also provides an overview of the challenges faced by the current healthcare landscape and summarizes the key contributions of the thesis.

Chapter 2 offers a comprehensive literature review, examining the current state of NLP techiniques. It contextualizes the advancements in the field that inform this research, including an analysis of existing methods in general text classification, chief complaint classification, genearl dialogue generation, medical dialogue generation, and general named entity recognition, and medical named entity recognition. The chapter highlights the significance of understanding patient language and clinical terminology, emphasizing the need for effective models that can interpret the nuances of medical dialogue. This foundational knowledge sets the stage for the contributions presented in subsequent chapters.

Chapter 3 details the development of a novel framework to classify chief complaint text from patients. By leveraging a hierarchical relational network, the framework effectively captures the complexity of medical terminology, facilitating improved patient triage processes. The evaluation results indicate that this approach effectively enhances the accuracy of chief complaint classification.

Chapter 4 advances the field of medical dialogue generation by introducing a framework designed to simulate patient-doctor conversations. This chap-

ter elaborates on the architecture of the proposed framework which utilizes context-aware techniques to produce relevant and coherent responses. The framework demonstrates its ability to understand the nuances of patient language, addressing challenges such as ambiguous inputs and varying dialects. The effectiveness of the framework is validated through user studies, showing that it significantly improves the quality of patient interactions in telehealth environments, ultimately enhancing the overall patient experience.

In Chapter 5, the focus shifts to the classification of physician text during teleconsultations into the types of social support provided. This chapter categorizes physician responses into direct informational support, indirect informational support, and emotional support, employing a Rule-enriched Attention-based Deep Neural Network for classification. The findings reveal insights into communication strategies that can enhance patient engagement and satisfaction.

Finally, Chapter 6 explores the essential task of medical named entity recognition. This chapter presents a novel approach that combines prompt learning with pre-trained deep learning models to effectively position and categorize medical entities from clinical narratives. The research emphasizes the importance of comprehensive medical ontologies in enhancing contextual understanding and accuracy. The results demonstrate significant performance gains, contributing to improved data management and informed decision-making processes in clinical settings.

Overall, this thesis demonstrates the significant potential of deep learning-based NLP techniques to transform healthcare delivery by automating and refining various processes associated with patient care. Each chapter contributes to the overarching goal of enhancing patient-provider interactions

and optimizing clinical data management, thereby addressing critical challenges towards smart healthcare.

7.2 Limitations

This thesis explores advanced NLP techniques in the medical field, introducing novel frameworks and methods for chief complaint classification, medical dialogue generation, physician text classification, and NER. Despite the promising results, several limitations are evident across these studies. The chief complaint classification framework is constrained by the quality and diversity of the hierarchical label data, relying on well-defined medical categories that may not capture the nuances of clinical language or emerging medical terminology. Additionally, its computational complexity may challenge scalability and efficiency in larger datasets or real-time applications. The medical dialogue generation framework, while showing substantial improvement, lacks multilingual capabilities, which may hinder performance across diverse languages and dialects, and it can struggle with highly diverse or ambiguous patient input. Similarly, the Rule-enriched Attention-based Deep Neural Network for physician text classification demonstrates enhanced accuracy but is limited by the narrow scope of its social support lexicon and the specificity of domain rules, raising concerns about its generalizability across different healthcare settings. In the realm of NER, while combining prompt learning with pre-trained model has yielded significant performance gains, it presents challenges such as the labor-intensive process of crafting effective prompts and the substantial computational resources required by the relational network used in its preditors. Furthermore, the model's reliance on comprehensive medical ontologies and structured data poses limitations in scenarios where such

resources are unavailable or insufficiently detailed. Future research should focus on enhancing these methodologies to improve their applicability and efficiency within diverse healthcare context.

7.3 Future Work

The future direction of this thesis focuses on addressing the identified limitations and advancing the field of medical NLP. For chief complaint classification, efforts will be directed towards enhancing the framework's scalability and efficiency by optimizing the hierarchical relational network and expanding the label dataset to encompass a broader range of medical conditions and emerging terminology. In medical dialogue generation, a key area of interest will be developing multilingual models and incorporating diverse linguistic datasets to ensure effective performance across various language contexts, along with exploring techniques to handle highly diverse or ambiguous patient inputs. For physician text classification, future work will include expanding the social support lexicon, refining rule-based components, and testing the R-ADNN framework across different healthcare settings and languages to assess its generalizability and robustness. For medical NER, research will delve into integrating comprehensive medical ontologies and structured data sources to enhance contextual understanding and accuracy, while also investigating the scalability of the prompt learning and relational network approach with larger and more diverse medical datasets. Efforts will focus on simplifying prompt creation and improving computational efficiency to balance model complexity with practical usability in real-world healthcare applications.

Bibliography

Abdallah, S., Shaalan, K., and Shoaib, M. (2012). Integrating rule-based system with classification for arabic named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 311–322. Springer.

AlMahmoud, R. H. and Hammo, B. H. (2024). SEWAR: A corpus-based n-gram approach for extracting semantically-related words from arabic medical corpus. *Expert Systems with Applications*, 238:121767.

An, Y., Xia, X., Chen, X., Wu, F.-X., and Wang, J. (2022). Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. *Artificial Intelligence in Medicine*, 127:102282.

Avci, E. and Turkoglu, I. (2009). An intelligent diagnosis system based on principle component analysis and anfis for the heart valve diseases. *Expert Systems with Applications*, 36(2):2873–2878.

Bambina, A. (2007). *Online social support: the interplay of social networks and computer-mediated communication*. Cambria press.

Banerjee, S., Akkaya, C., Perez-Sorrosal, F., and Tsioutsiouliklis, K. (2019). Hierarchical transfer learning for multi-label text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300.

Białecki, A., Muir, R., Ingersoll, G., and Imagination, L. (2012). Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*, page 17.

Blanco, A., Casillas, A., Pérez, A., and Diaz de Ilarraza, A. (2019). Multilabel clinical document classification: Impact of label-density. *Expert Systems with Applications*, 138:112835.

Borchert, F., Lohr, C., Modersohn, L., Witt, J., Langer, T., Follmann, M., Gietzelt, M., Arnrich, B., Hahn, U., and Schapranow, M.-P. (2022). Ggponc 2.0-the german clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline ner taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660.

Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., and Ghosh, P. (2021). A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319.

Brown, P., Halász, S., Goodall, C., Cochrane, D. G., Milano, P., and Allegra, J. R. (2010). The ngram chief complaint classifier: A novel method of automatically creating chief complaint classifiers based on international classification of diseases groupings. *Journal of Biomedical Informatics*, 43(2):268–272.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., and Tortora, G. (2024). Can chatgpt provide intelligent diagnoses? a com-

parative study between predictive models and chatgpt to define a new medical diagnostic bot. *Expert Systems with Applications*, 235:121186.

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., and Androutsopoulos, I. (2019). Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322.

Chang, D., Hong, W. S., and Taylor, R. A. (2020). Generating contextual embeddings for emergency department chief complaints. *JAMIA open*, 3(2):160–166.

Chapman, W. W., Christensen, L. M., Wagner, M. M., Haug, P. J., Ivanov, O., Dowling, J. N., and Olszewski, R. T. (2005a). Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial Intelligence in Medicine*, 33(1):31–40.

Chapman, W. W., Dowling, J. N., and Wagner, M. M. (2005b). Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. *Annals of Emergency Medicine*, 46(5):445–455.

Chapman, W. W., Dowling, J. N., and Wagner, M. M. (2005c). Generating a reliable reference standard set for syndromic case classification. *Journal of the American Medical Informatics Association*, 12(6):618–629.

Chau, M., Li, T. M., Wong, P. W., Xu, J. J., Yip, P. S., and Chen, H. (2020). Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS quarterly*, 44(2).

Chawla, A., Mulay, N., Bishnoi, V., and Dhama, G. (2021). Improving the performance of transformer context encoders for ner. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

Chen, L., Baird, A., and Straub, D. (2019). Fostering participant health knowledge and attitudes: an econometric study of a chronic disease-focused online health community. *Journal of Management Information Systems*, 36(1):194–229.

Chen, S., Guo, X., Wu, T., and Ju, X. (2020). Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. *Information Processing & Management*, 57(5):102253.

Chen, W., Gong, Y., Wang, S., Yao, B., Qi, W., Wei, Z., Hu, X., Zhou, B., Mao, Y., Chen, W., Cheng, B., and Duan, N. (2022). DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, Dublin, Ireland. Association for Computational Linguistics.

Chen, Z., Yang, R., Zhao, Z., Cai, D., and He, X. (2018). Dialogue act recognition via crf-attentive structured network. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Choi, B.-J., Park, J.-H., and Lee, S. (2019). Adaptive convolution for text classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2475–2485.

Clifford, C. T., Pour, T. R., Freeman, R., Reich, D. L., Glicksberg, B. S., Levin, M. A., and Klang, E. (2021). Association between covid-19 diagnosis and presenting chief complaint from new york city triage data. *The American Journal of Emergency Medicine*, 46:520–524.

Cobb, S. (1976). Social support as a moderator of life stress. *Psychosomatic medicine*, 38(5):300–314.

Cui, C. and Li, Z. (2024). Prompt-enhanced generation for multimodal open question answering. *Electronics*, 13(8):1434.

Cui, M., Bai, R., Lu, Z., Li, X., Aickelin, U., and Ge, P. (2019). Regular expression based medical text classification using constructive heuristic approach. *IEEE Access*, 7:147892–147904.

Cutrona, C. E. and Suhr, J. A. (1992). Controllability of stressful events and satisfaction with spouse support behaviors. *Communication research*, 19(2):154–174.

de Lima Santos, D. B., de Carvalho Dutra, F. G., Parreiras, F. S., and Brandão, W. C. (2021). Assessing the effectiveness of multilingual transformer-based text embeddings for named entity recognition in portuguese. In *ICEIS (1)*, pages 473–483.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., and Si, L. (2019). A neural multi-digraph model for chinese ner with gazetteers. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1462–1467.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343.

Ebrahimi, M., Nunamaker Jr, J. F., and Chen, H. (2020). Semi-supervised cyber threat identification in dark net markets: A transductive and deep learning approach. *Journal of Management Information Systems*, 37(3):694–722.

Eftimov, T., Koroušić Seljak, B., and Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., and Stamatopoulos, P. (2000). Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 75–78.

Friedman, C. and Johnson, S. B. (2006). Natural language and text processing in biomedicine. In *Biomedical Informatics*, pages 312–343.

Fu, G. and Luke, K.-K. (2005). Chinese named entity recognition using lexicalized hmms. *SIGKDD Explor. Newsl.*, 7(1):19–25.

Garla, V., Taylor, C., and Brandt, C. (2013). Semi-supervised clinical text classification with laplacian svms: an application to cancer case management. *Journal of Biomedical Informatics*, 46(5):869–875.

Graves, A., Jaitly, N., and Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640.

Gu, X., Cho, K., Ha, J.-W., and Kim, S. (2019). DialogWAE: Multimodal response generation with conditional wasserstein auto-encoder. In *International Conference on Learning Representations*.

Gu, X., Yoo, K. M., and Ha, J.-W. (2021). Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12911–12919.

Gupta, D., Suman, S., and Ekbal, A. (2021). Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications*, 164:113993.

He, K., Mao, R., Huang, Y., Gong, T., Li, C., and Cambria, E. (2023). Template-free prompting for few-shot named entity recognition via semantic-enhanced contrastive learning. *IEEE transactions on neural networks and learning systems*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hou, Z., Bi, S., Qi, G., Zheng, Y., Ren, Z., and Li, Y. (2024). Syntax-guided question generation using prompt learning. *Neural Computing and Applications*, 36(12):6271–6282.

Hsu, J.-H., Weng, T.-C., Wu, C.-H., and Ho, T.-S. (2020). Natural language processing methods for detection of influenza-like illness from chief complaints. In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1626–1630.

Huang, L., Ma, D., Li, S., Zhang, X., and Wang, H. (2019). Text level graph neural network for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3444–3450.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hughes, M., Li, I., Kotoulas, S., and Suzumura, T. (2017). Medical text classification using convolutional neural networks. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, pages 246–250.

Janowski, A. (2023). Natural language processing techniques for clinical text analysis in healthcare. *Journal of Advanced Analytics in Healthcare Management*, 7(1):51–76.

Jernite, Y., Halpern, Y., Horng, S., and Sontag, D. (2013). Predicting chief complaints at triage time in the emergency department. In *NIPS*

Workshop on Machine Learning for Clinical Data Analysis and Healthcare.

Jin, Z., Zhang, Y., Kuang, H., Yao, L., Zhang, W., and Pan, Y. (2019). Named entity recognition in traditional chinese medicine clinical cases combining bilstm-crf with knowledge graph. In *Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part I 12*, pages 537–548. Springer.

Kang, N., Singh, B., Afzal, Z., van Mulligen, E. M., and Kors, J. A. (2013). Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5):876–881.

Kim, J., Amplayo, R. K., Lee, K., Sung, S., Seo, M., and Hwang, S.-w. (2019). Categorical metadata representation for customized text classification. *Transactions of the Association for Computational Linguistics*, 7:201–215.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kundeti, S. R., Vijayananda, J., Mujjiga, S., and Kalyan, M. (2016). Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945. IEEE.

Lafferty, J., McCallum, A., Pereira, F., et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Lauriola, I., Lavelli, A., and Aiolli, F. (2022). An introduction to deep

learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470:443–456.

Lee, L.-H. and Lu, Y. (2021). Multiple embeddings enhanced multi-graph neural networks for chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Lee, S. H., Levin, D., Finley, P. D., and Heilig, C. M. (2019). Chief complaint classification with recurrent neural networks. *Journal of Biomedical Informatics*, 93:103158.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Li, B., Chen, E., Liu, H., Weng, Y., Sun, B., Li, S., Bai, Y., and Hu, M. (2021a). More but correct: Generating diversified and entity-revised medical response. *arXiv preprint arXiv:2108.01266*.

Li, D., Ren, Z., Ren, P., Chen, Z., Fan, M., Ma, J., and de Rijke, M. (2021b). Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554.

Li, G., Song, H., Liang, H.-N., Qu, Y., Liu, L., and Bai, X. (2019a). Medical diagnosis by complaints of patients and machine learning. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pages 1–5.

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., Ji, D., and Li, F. (2022a). Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Li, J., Wu, H., Deng, Z., Lu, N., Evans, R., and Xia, C. (2019b). How professional capital and team heterogeneity affect the demands of online team-based medical service. *BMC Medical Informatics and Decision Making*, 19:1–15.

Li, M., Lin, X., Chen, X., Chang, J., Zhang, Q., Wang, F., Wang, T., Liu, Z., Chu, W., Zhao, D., and Yan, R. (2022b). Keywords and instances: A hierarchical contrastive learning framework unifying hybrid granularities for text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4432–4441.

Li, R., Lin, C., Collinson, M., Li, X., and Chen, G. (2019c). A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., and Du, X. (2018). Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143.

Li, X., Cui, M., Li, J., Bai, R., Lu, Z., and Aickelin, U. (2021c). A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing*, 443:345–355.

Li, X., Zhang, H., and Zhou, X.-H. (2020). Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of Biomedical Informatics*, 107:103422.

Liu, B., Tur, G., Hakkani-Tur, D., Shah, P., and Heck, L. (2017). End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. *arXiv preprint arXiv:1711.10712*.

Liu, C., Sun, H., Du, N., Tan, S., Fei, H., Fan, W., Yang, T., Wu, H., Li, Y., and Zhang, C. (2016a). Augmented lstm framework to construct medical self-diagnosis android. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 251–260.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016b). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Liu, J., Bai, R., Lu, Z., Ge, P., Aickelin, U., and Liu, D. (2020a). Data-driven regular expressions evolution for medical text classification using genetic programming. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8.

Liu, J., Gao, L., Guo, S., Ding, R., Huang, X., Ye, L., Meng, Q., Nazari, A., and Thiruvady, D. (2021a). A hybrid deep-learning approach for complex biochemical named entity recognition. *Knowledge-Based Systems*, 221:106958.

Liu, J., Zhang, X., Kong, J., and Wu, L. (2020b). The impact of teammates' online reputations on physicians' online appointment numbers: A social interdependency perspective. In *Healthcare*, volume 8, page 509. MDPI.

Liu, M., Bao, X., Liu, J., Zhao, P., and Shen, Y. (2021b). Generating emotional response by conditional variational auto-encoder in open-domain dialogue system. *Neurocomputing*, 460:106–116.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023a). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, W., Tang, J., Cheng, Y., Li, W., Zheng, Y., and Liang, X. (2022). Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 447–459. Springer.

Liu, X., Yu, H.-F., Dhillon, I., and Hsieh, C.-J. (2020c). Learning to encode position for transformer with continuous dynamical model. In *International Conference on Machine Learning*, pages 6327–6335.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA. Association for Computational Linguistics.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2023b). Gpt understands, too. *AI Open*.

Lu, H.-M., Zeng, D., Trujillo, L., Komatsu, K., and Chen, H. (2008). Ontology-enhanced automatic chief complaint classification for syndromic surveillance. *Journal of Biomedical Informatics*, 41(2):340–356.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., and Wang, J. (2018). An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Luo, Y. (2017). Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, 72:85–95.

Ma, Y., Liu, X., Zhao, L., Liang, Y., Zhang, P., and Jin, B. (2022). Hybrid embedding-based text representation for hierarchical multi-label text classification. *Expert Systems with Applications*, 187:115905.

Mikosz, C. A., Black, S., Gibbs, G., Cardenas, I., and Silva, J. (2004). Comparison of two major emergency department-based free-text chief-complaint coding systems. *Morbidity and Mortality Weekly Report*, pages 101–105.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Miyazaki, T., Makino, K., Takei, Y., Okamoto, H., and Goto, J. (2019). Label embedding using hierarchical structure of labels for twitter classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 6318–6323.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1101–1111.

Nagarhalli, T. P., Vaze, V., and Rana, N. (2021). Impact of machine learning in natural language processing: A review. In *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*, pages 1529–1534. IEEE.

Ohashi, S., Takayama, J., Kajiwara, T., Chu, C., and Arase, Y. (2020). Text classification with negative supervision. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 351–357.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pappas, N. and Henderson, J. (2019). Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

Park, Y., Cho, J., and Kim, G. (2018). A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An

imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., and Spyropoulos, C. D. (2001). Using machine learning to maintain rule-based named-entity recognition and classification systems. In *proceedings of the 39th annual meeting of the association for computational linguistics*, pages 426–433.

Polignano, M., de Gemmis, M., Semeraro, G., et al. (2021). Comparing transformer-based ner approaches for analysing textual medical diagnoses. In *CLEF (Working Notes)*, pages 818–833.

Pomares-Quimbaya, A., Gonzalez, R. A., Velandia, O. M. M., Peña, A. A. G., Rodríguez, J. C. D., Múnera, A. S., and Labbé, C. (2018). Concept attribute labeling and context-aware named entity recognition in electronic health records. *International Journal of Reliable and Quality E-Healthcare (IJRQEH)*, 7(1):1–15.

Qiu, Y., Li, H., Li, S., Jiang, Y., Hu, R., and Yang, L. (2018). Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 209–221. Springer.

Quimbaya, A. P., Múnera, A. S., Rivera, R. A. G., Rodríguez, J. C. D., Velandia, O. M. M., Peña, A. A. G., and Labbé, C. (2016). Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raposo, D., Santoro, A., Barrett, D., Pascanu, R., Lillicrap, T., and Battaglia, P. (2017). Discovering objects and their relations from entangled scene representations. *arXiv preprint arXiv:1702.05068*.

Roy, K., Debdas, S., Kundu, S., Chouhan, S., Mohanty, S., and Biswas, B. (2021). Application of natural language processing in healthcare. *Computational Intelligence and Healthcare Informatics*, pages 393–407.

Saibene, A., Assale, M., and Giltri, M. (2021). Expert systems: definitions, advantages and issues in medical field applications. *Expert Systems with Applications*, 177:114900.

Samtani, S., Chai, Y., and Chen, H. (2022). Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model1. *MIS quarterly*, 46(2).

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Sato, S., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. (2017). Modeling situations in neural chat bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). 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.

Schäfer, A., Blach, N., Rausch, O., Warm, M., and Krüger, N. (2020). Towards automated anamnesis summarization: Bert-based models for symptom extraction. *arXiv preprint arXiv:2011.01696*.

Schweter, S. and Baiter, J. (2019). Towards robust named entity recognition for historic german. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T., and Alinejad-Rokny, H. (2021). A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627.

Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Shen, X., Su, H., Niu, S., and Demberg, V. (2018). Improving variational

encoder-decoders in dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Shen, Y., Tan, Z., Wu, S., Zhang, W., Zhang, R., Xi, Y., Lu, W., and Zhuang, Y. (2023). PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Shen, Y., Wang, X., Tan, Z., Xu, G., Xie, P., Huang, F., Lu, W., and Zhuang, Y. (2022). Parallel instance query network for named entity recognition. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics.

Shimura, K., Li, J., and Fukumoto, F. (2018). HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 811–816.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Sulieman, L., Gilmore, D., French, C., Cronin, R. M., Jackson, G. P., Russell, M., and Fabbri, D. (2017). Classifying patient portal messages using convolutional neural networks. *Journal of Biomedical Informatics*, 74:59–70.

Sun, J. (2020). jieba chinese text segmentation. https://github.com/fxsjy/jieba.

Tan, H. and Yan, M. (2020). Physician-user interaction and users' perceived service quality: evidence from chinese mobile healthcare consultation. *Information Technology & People*, 33(5):1403–1426.

Tang, K.-F., Kao, H.-C., Chou, C.-N., and Chang, E. Y. (2016). Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS workshop on deep reinforcement learning*.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Travers, D. A. and Haas, S. W. (2004). Evaluation of emergency medical text processor, a system for cleaning chief complaint text data. *Academic Emergency Medicine*, 11(11):1170–1176.

Valmianski, I., Goodwin, C., Finn, I. M., Khan, N., and Zisook, D. S. (2019). Evaluating robustness of language models for chief complaint extraction from patient-generated text. *arXiv preprint arXiv:1911.06915*.

Varshney, D., Zafar, A., Behera, N. K., and Ekbal, A. (2023). Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports*, 13(1):3310.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., et al. (2018). Us-

ing clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.

Wang, B., Shang, L., Lioma, C., Jiang, X., Yang, H., Liu, Q., and Simonsen, J. G. (2020). On position embeddings in bert. In *International Conference on Learning Representations*.

Wang, Q., Zhou, Y., Ruan, T., Gao, D., Xia, Y., and He, P. (2019). Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of biomedical informatics*, 92:103133.

Wang, X., Chused, A., Elhadad, N., Friedman, C., and Markatou, M. (2008). Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 783. American Medical Informatics Association.

Wang, Y., Shindo, H., Matsumoto, Y., and Watanabe, T. (2021). Nested named entity recognition via explicitly excluding the influence of the best path. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3547–3557, Online. Association for Computational Linguistics.

Wang, Z., Wang, P., Huang, L., Sun, X., and Wang, H. (2022). Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119. Association for Computational Linguistics.

Wei, Z., Liu, Q., Peng, B., Tou, H., Chen, T., Huang, X., Wong, K.-f., and Dai, X. (2018). Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Wu, Y., Jiang, M., Xu, J., Zhi, D., and Xu, H. (2017). Clinical named entity recognition using deep learning models. In *AMIA annual symposium proceedings*, volume 2017, page 1812. American Medical Informatics Association.

Xia, Y., Zhou, J., Shi, Z., Lu, C., and Huang, H. (2020). Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1062–1069.

Xu, C., Wang, F., Han, J., and Li, C. (2019a). Exploiting multiple embeddings for chinese named entity recognition. page 2269–2272, New York, NY, USA. Association for Computing Machinery.

Xu, L., Zhou, Q., Gong, K., Liang, X., Tang, J., and Lin, L. (2019b). End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Y. Mahajan, P. and Rana, D. P. (2023). Text mining approach for the prediction of disease status from discharge summaries using ccbe and neroa-cnn. *Expert Systems with Applications*, 227:120310.

Yan, G., Pei, J., Ren, P., Ren, Z., Xin, X., Liang, H., de Rijke, M., and Chen, Z. (2022). Remedi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SI-*

GIR Conference on Research and Development in Information Retrieval, pages 3013–3024.

Yan, H., Sun, Y., Li, X., and Qiu, X. (2023). An embarrassingly easy but strong baseline for nested named entity recognition. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1442–1452, Toronto, Canada. Association for Computational Linguistics.

Yang, T., Tran, T. T., and Gurevych, I. (2023). Dior-CVAE: Pre-trained language models and diffusion priors for variational dialog generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yang, Z., Chen, H., Zhang, J., Ma, J., and Chang, Y. (2020). Attention-based multi-level feature fusion for named entity recognition. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3594–3600. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., et al. (2020). Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Zhang, H., Lan, Y., Pang, L., Guo, J., and Cheng, X. (2019). ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Zhang, S., Zhang, X., Wang, H., Cheng, J., Li, P., and Ding, Z. (2017). Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767.

Zhang, X., Xu, J., Soh, C., and Chen, L. (2022). LA-HCN: Label-based attention for hierarchical multi-label text classification neural network. *Expert Systems with Applications*, 187:115922.

Zhang, Y., Fang, Q., Qian, S., and Xu, C. (2020a). Knowledge-aware attentive wasserstein adversarial dialogue response generation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–20.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020b). DIALOGPT : Large-scale generative pre-training for conversational response generation. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhang, Y. and Yang, J. (2018). Chinese NER using lattice LSTM. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.

Zhao, T., Zhao, R., and Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Zhao, W., Liu, Q. B., Guo, X., Wu, T., and Kumar, S. (2022). Quid pro quo in online medical consultation? investigating the effects of small monetary gifts from patients. *Production and Operations Management*, 31(4):1698–1718.

Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480.

Zhou, J., Zhang, Q., Zhou, S., Li, X., and Zhang, X. M. (2023). Unintended emotional effects of online health communities: A text mining-supported empirical study. *MIS Quarterly*.

Zhou, M., Li, Z., Tan, B., Zeng, G., Yang, W., He, X., Ju, Z., Chakravorty, S., Chen, S., Yang, X., et al. (2021). On the generation of medical dialogs for covid-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for

relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Zhu, Q., Li, X., Conesa, A., and Pereira, C. (2018). Gram-cnn: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, 34(9):1547–1554.

Zhu, T., Qin, Y., Chen, Q., Hu, B., and Xiang, Y. (2022). Enhancing entity representations with prompt learning for biomedical entity linking. In *IJCAI*, pages 4036–4042.

Zhu, Z., Li, J., Zhao, Q., and Akhtar, F. (2023). A dictionary-guided attention network for biomedical named entity recognition in chinese electronic medical records. *Expert Systems with Applications*, 231:120709.