INTELLIGENT AND CUSTOMIZABLE CONVERSATIONAL SYSTEMS FOR CLINICAL COMMUNICATION TRAINING

ZHANG XIANG

PhD

The Hong Kong Polytechnic University

2023

The Hong Kong Polytechnic University
Department of Computing

Intelligent and Customizable Conversational Systems for Clinical Communication Training

Zhang Xiang

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
August 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Zhang Xiang

Abstract

The aging of the population and the outbreak of pandemics place the challenges on global healthcare. With advances in artificial intelligence and big data, intelligent healthcare improves the efficiency of the medical system and reduces the workload of the practitioners greatly. In this context, we aim to facilitate autonomous, low-cost, and customizable clinical communication training by developing intelligent techniques.

Effective clinical communication is essential for delivering safe and high-quality patient care, especially under the scenarios that the healthcare system faces high pressure. Training on standardized clinical communication helps to organize the conversation in a structured and focused way that ensures clinical staff get timely and appropriate responses without missing any important information. Traditional classroom teaching on clinical communication requires substantial human and medical resources, and more importantly, lacks enough high-fidelity practice. Therefore, the primary function of this intelligent training system is to simulate the dialogue among the clinicians. We develop a task-oriented multiturn chatbot, which can play various roles to practice conversations with clinical staff. The key research problem addressed here is the detection of sentence-level intents referring to the context of clinical communication standards. Compared to the existing works on intent detection, the sample dialogues for standardized clinical

handover are insufficient. Moreover, these dialogues are inherently sequential and their intents are interrelated. Given this feature, we propose Intent-aware Long-Short Term Memory (IA-LSTM) to incorporate context information into intent detection. In the experiments, IA-LSTM outperforms all baseline methods of intent detection on clinical handovers. Moreover, the proposed intent-aware mechanism can be expanded to other deep learning models, thereby improving their performances.

The second piece of work we have delivered is a timely assessment model, which can automatically evaluate the performance of individual clinical staff in a conversation. The research problem addressed here is the accurate recognition of the conversation content by integrating the information from both the domain knowledge and the learning examples. Based on the biomedical ontology for general purpose, we construct a specific knowledge graph for the clinical communications. A novel method called Knowledge-infused Prompt Tuning is proposed to infuse the external knowledge into prompts. The empirical validation in real world application shows that the proposed method not only achieves superior performance, but also proves more robust with limited data or complex components.

The third function we have developed for intelligent clinical communication training is a friendly platform that enables users to define new training tasks by themselves. The key research problem addressed here is the customizable conversational system with insufficient training data. We propose a novel data augmentation methods for user-defined scenarios, such as the clinical handover under the COVID-19. Based on the pre-trained conversational system with user-defined knowledge, the proposed Data Augmentation with User-Defined Knowledge (UDK-DA) significantly boosts the performance of the clinical training system with only a few samples.

By integrating the aforementioned modules, we develop Heallo, an intelligent, au-

tonomous, and customizable conversational system for clinical communication training. Now Heallo has been incorporated into junior staff training programs at local hospitals and largely benefits the promotion of intelligent healthcare.

Publications

- **Xiang Zhang**, Bruce X.B. Yu, Yan Liu, George Wing-Yiu Ng, Nam-Hung Chai, Eric Hang-Kwong So, Sze-Sze So, and Victor Kai-Lam Cheung. "Heallo: Conversational System for Communication Training in Healthcare Professional Education". In: *2022 10th International Conference on Information and Education Technology (ICIET)*. IEEE. 2022, pp. 32–36.

- **Xiang Zhang**, Yan Liu, Gong Chen and Sheng-hua Zhong. "Lightblue: Nurture Your Personal Chatbot". In: *The 9th International Conference on Artificial Intelligence and Applications (AIAP2022)*, Vol. 12, Feb 2022, pp. 01-21.

- **Xiang Zhang**, Bruce X.B. Yu, Yan Liu, George Wing-Yiu Ng, Nam-Hung Chai, Eric Hang-Kwong So, Sze-Sze So, and Victor Kai-Lam Cheung. "Intent-Aware Long Short-Term Memory for Intelligent Training of Clinical Handover". *2022. In 2022 7th International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 11-16. IEEE, 2022.

- **Xiang Zhang**, Bruce X.B. Yu, Yan Liu, Gong Chen, George Wing-Yiu Ng, Nam-Hung Chia, Eric Hang-Kwong So, Sze-Sze So, Victor Kai-Lam Cheung. "Conversational System for Clinical Communication Training Supporting User-defined

Tasks". *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE) 2022*, pp. 396-403. IEEE, 2022.

- Bruce X.B. Yu, Yan Liu, **Xiang Zhang**, Sheng-hua Zhong, and Keith C.C. Chan. "MMNet: A Model-based Multimodal Network for Human Action Recognition in RGB-D Videos". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 3 (2022): 3522-3538.

- Bruce X.B. Yu, Yan Liu, **Xiang Zhang**, Gong Chen, and Keith C.C. Chan. "EGCN: An Ensemble-based Learning Framework for Exploring Effective Skeleton-based Rehabilitation Exercise Assessment". In: *Proceedings of the 31st International Joint Conference on Artificial Intelligence* (2022): 3681-3687.

- Baixi Xing, **Xiang Zhang**, Kejun Zhang, Xinda Wu, Hui Zhang, Jun Zheng, Lekai Zhang, and Shouqian Sun. "PopMash: an automatic musicalmashup system using computation of musical and lyrical agreement for transitions". In: *Multimedia Tools and Applications* 79.29 (2020), pp. 21841–21871.

- Gong Chen, Yan Liu, Sheng-hua Zhong, and **Xiang Zhang**. "Musicality-Novelty Generative Adversarial Nets for Algorithmic Composition". In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 1607–1615.

- Yang Liu, Yan Liu, **Xiang Zhang**, Gong Chen, and Kejun Zhang. "Learning Music Emotion Primitives via Supervised Dynamic Clustering". In: *Proceedings of the 24th ACM international conference on Multimedia*. 2016, pp. 222–226.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Fiona Liu, for her academic and personal guidance. When I first joined her group, I had no prior research experience. It is her support and encouragement that enables me to develop as a researcher along the way. Thank Fiona for allowing me the most freedom to work on projects I am passionate about, for her meticulous care and unreserved assistance, and for her wisdom in a variety of fields beyond academia. I am always astonished by her big-picture ideas of research and razor-sharp insights; I have frequently realized that a suggestion I initially rejected from Fiona was actually spot-on. She not only approaches research with an open and pure mind, but she also cares for every member like a family.

The majority of my time in this research group led by Fiona is spent as a learner, gaining knowledge from many wonderful group members. I remember the countless discussions, sharing, group meetings, among which I enjoy the summer learning sessions most. Whether or not the topic is related to my own research, I value the moments we discuss, ask questions, and freely exchange ideas. They keep me intellectually curious and free while inspiring me in my field. Specifically, I would like to thank Yang Liu, Sheng-hua Zhong, Song-tao Wu, Gong Chen, Bruce Yu, Yongxu Liu, Zhejing Hu, Zhi Zhang, and Xiao Ma for their selfless contribution of time and valuable inputs. Collabo-

rating with these colleagues has been not only productive but also incredibly enjoyable.

In addition, it has been a privilege to collaborate with the practitioners at Queen Elizabeth Hospital (QEH). It is a fascinating experience to communicate and gain knowledge across various professional fields. And I am filled with admiration and gratitude for their valor and dedication during the difficult period of COVID-19. Without their efforts, this thesis would not exist. I'd like to thank the multi-disciplinary simulation and skills centre team at QEH, especially Dr. George Ng, Dr. Chia, Dr. Eric So, Sze-Sze, and Victor.

I consider myself extremely fortunate to have served as the hall tutor at PolyU's student hostel for four years, where I met a number of exceptional tutors, wardens, and managers, as well as a group of enthusiastic and entertaining hallmates. The retreats, competitions, festivals, and activities at hall have brightened my life and brought me innumerable joy.

My time at PolyU has been fantastic not only for academics, but also for many wonderful friends outside of the PhD program who have kept me happy over the past many years: we enjoy sports, games, brunches, afternoon teas, sunsets, outings and so on. I would like to thank Jasper, Shirmmy, Xian Zhan, Jun, Fengze, Linwan and Eddie in particular. Their companies have rendered the journey unique, memorable, and meaningful.

Lastly, I'd like to express my gratitude for the love and support of my families. Big hugs to my parents, Mr. Heqing Zhang and Mrs. Yunping Li, who are always there for me, care for me, and love me. And thank you to my little sister, Yushu Zhang, for the enjoyable and loving times we shared. This thesis would not have been possible without the support of many wonderful people.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The escalating pressures on the global healthcare system, driven by an aging population, a surge in chronic disease patients, and a rising demand for improved quality of life, have been further exacerbated by the COVID-19 pandemic [4]. The pandemic has precipitated the collapse of numerous medical systems, placing an unprecedented strain on healthcare worldwide.

Modern healthcare is characterized by its collaborative nature, often involving a diverse team of physicians, nurses, and other medical professionals. As such, effective communication within these teams has emerged as a critical prerequisite for successful collaboration [85]. This importance of clear communication becomes even more pronounced in high-stress situations, such as during pandemic response or in the management of complex chronic conditions, where the ability to accurately convey information can directly impact patient outcomes.

In response to these challenges, substantial investments have been made in intelligent

healthcare with the aim of alleviating practitioner workload and enhancing the efficiency of medical care [24]. It is also observed that there was a sudden increased demand of intelligent healthcare with the COVID-19 advent, and the adoption rate of intelligent healthcare is further projected to grow in the post-COVID-19 scenario [42].

With the rapid advances in computational power and the widespread digitization of medical data, Artificial Intelligence (AI) technologies have emerged as a powerful tool with the potential to revolutionize numerous facets of patient care, as well as administrative and operational procedures [27]. And AI has already facilitated transformative advancements in areas such as drug discovery, clinical trials, and personalized medicine [113]. However, despite these significant strides, the potential application of these intelligent techniques to enhance clinical communication remains largely unexplored.

In the subsequent sections of this thesis, we delve deeper into our motivation for this research by examining the current landscape of clinical communication. Based on this analysis, we then formulate specific research problems with the ultimate goal of enhancing the efficiency and effectiveness of clinical communication through the application of AI technologies.

1.2 Scenario Description

1.2.1 Clinical Communication

Clinical communication, a vital component of patient care, encompasses the dynamic exchange of ideas, messages, or knowledge between healthcare providers and patients or among members of a clinical team. This interaction can transpire through various mediums, including oral and written forms, as well as non-verbal cues and signals. In the

context of thesis,our primary focus is on oral communication, a prevalent yet potentially error-prone activity in the realm of patient care.

The effective communication is essential for the provision of high-quality, patient-centered care, especially in a team-based healthcare setting [104]. The dialogue that takes place among healthcare professionals forms a substantial part of the information flow within the healthcare system, impacting every facet of care delivery, from the execution of professional skills to the determination of patient outcomes [38]. Any lapses or errors in communication can have far-reaching consequences, including delayed or incorrect treatment, medication errors, and in the most severe cases, increased patient mortality [70].

In times of crisis, when the healthcare system is stretched to its limits, the role of clinical communication becomes even more pronounced. It becomes the linchpin that determines the efficiency of medical resource utilization. For instance, Hong Kong reported more than 300,000 new confirmed cases within a single week at the peak of the pandemic [25]. In stark contrast, the city's public hospitals could only accommodate about 9,000 beds specifically for COVID-19 patients [53]. This disparity in patient numbers and resources nearly paralyzed the medical system.

In such emergency situations, effective clinical communication becomes the lifeline for optimal resource allocation and utilization. It ensures that despite the strain on resources, patient care is not compromised, and the healthcare system continues to function effectively [37]. Thus, the importance of clinical communication in healthcare cannot be overstated. It is not just a tool for information exchange but a critical factor in patient safety, healthcare efficiency, and overall system resilience.

1.2.2 Standardized Clinical Communication Protocols

To aid clinicians in communicating accurately and succinctly during various tasks, standardized communication protocols have been developed. For example, ISBAR (Identify, Situation, Background, Assessment, Recommendation), a standardized communication framework recommended by the World Health Organization [109], has demonstrated great potential for enhancing the accuracy and transparency of inter-professional and non-face-to-face handover in hospitals [85].

This protocol offers a systematic approach to clinical handover by breaking it down into essential components: ”Identify,” ”Situation,” ”Background,” ”Assessment,” and ”Recommendation”. Each component represents a discrete intent for conveying relevant clinical information, thereby ensuring the integrity of the clinical handover procedure. A detailed explanation of the five elements in the ISBAR Framework is provided in Table 1.1.

Element	**Description**
Identify	Identify yourself, the patient and verify the receiver.
Situation	Clarify the problem or reason for contact.
Background	Briefly summarize patient’s previous history relevant to the current problem.
Assessment	Share the latest clinical assessment, investigation, and your interpretation of the current situation.
Recommendation	Ask for advice or intervention; state your expectation.

Table 1.1: Description of ISBAR standardized communication framework.

However, the teaching of communication protocols remains in the theoretical realm, Theoretical knowledge often fails to translate into practice [15]. These protocols have not been successfully applied in real-world clinical scenarios due to a lack of training: First, conducting exercises in an actual clinical setting is prohibitively expensive, re-

quiring patient consent and professional supervision. Experienced doctors often lack the time to practice communication with junior staff. Second, the practice processes are difficult to monitor and assess. Practices between medical students often lack fidelity, and they rarely receive timely and accurate feedback.

The COVID-19 outbreak has further exacerbated these challenges [6]. The pandemic has led to a surge in the number of junior doctors and medical volunteers requiring training, while existing medical experts are overwhelmed with pandemic response duties. Furthermore, the pandemic has given rise to new and complex clinical scenarios, necessitating the development and implementation of new communication protocols [90].

However, the overburdened medical system, coupled with limited time and resources, often prevents clinicians from collecting sufficient data or carefully designing and implementing comprehensive training courses. Without adequate practice and training, healthcare professionals may struggle to provide effective communication, especially in emergency situations. This underscores the need for innovative solutions to more effective and accessible communication training.

1.2.3 AI-powered Clinical Communication Training

Given the scarcity of medical resources and high demand for training, an AI-powered conversational system emerges as a promising solution. This system can provide autonomous, cost-effective, and adaptive clinical communication training, thereby addressing the existing gaps in traditional training methods. In replace of human practitioners, the conversational system would simulate clinical communication scenarios, evaluate the dialogue quality of trainees, and adapt to new training tasks, offering a

dynamic and interactive learning environment. Figure 1.1 illustrates a diagram of a customizable conversational system for clinical communication training.

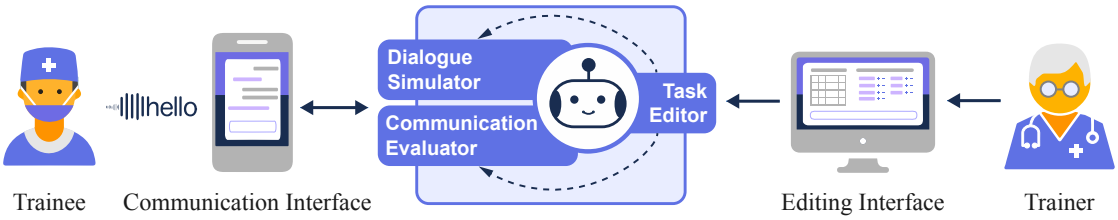

Figure 1.1: Diagram of AI-powered clinical communication training.

The system is composed of three primary modules, each designed to address a specific aspect of clinical communication training.

The first module is a dialogue simulator, essentially a task-oriented, multi-turn chatbot, capable of assuming various roles in practice conversations with clinical staff. This module provides a communication interface through which junior doctors or medical students can engage in simulated scenarios with the AI agent. By doing so, it enables autonomous communication practice, eliminating the need for costly medical resources and allowing for flexible, on-demand training.

The second module, termed the communication evaluator, is designed to automatically evaluate and analyze the performance of trainees. This module provides real-time feedback, thereby maximizing the effectiveness of the training and facilitating continuous improvement. By automating the evaluation process, this module eliminates the need for expert raters, a resource that is often scarce, particularly during public health crises such as the COVID-19 pandemic.

The third module, the task editor, is designed to support user-defined tasks. This feature allows non-IT professionals to easily customize new tasks to suit various communication scenarios. With a user-friendly editing interface, medical professionals can

create new tasks through simple editing, thereby expanding the application of the system to different communication tasks.

1.3 Research Problems

Based on the clinical communication scenario and the proposed diagram for autonomous communication training (Figure 1.1), we further analyze the challenges in each module and elucidate the research problems.

1.3.1 Intent detection in Standardised Clinical Communication

The primary function of this intelligent training system is to simulate the dialogue among the clinicians. During communication training, our conversational system typically serves as the information receiver. Therefore, the system's responses are designed to be relatively simple, primarily intended to guide trainees in continuing the conversation. Rather than pursuing realistic and diverse responses, the key objective of the dialogue simulator is to understand the user's intent during communication.

Although intent detection has been a central topic in Natural Language Processing (NLP) for a long time, its application to standardised clinical communication is relatively unexplored, with no publicly available datasets for this specific context. In our work, the intents are defined by different parts of a standardized clinical communication protocol. Unlike existing works on intent detection, sentences in standardised clinical communication are inherently sequential and the intents derived from the standards are interrelated. This means that the intent of each sentence is not only dependent on the preceding dialogue but also shapes the interpretation of the subsequent intents. Given these unique characteristics, we formulate the problem as sentence-level intent detection

referring to the context of clinical communication standards.

1.3.2 Content Recognition with Domain Knowledge

An intelligent training system also contains a real-time assessment module, which automatically evaluate the performance of individual clinical staff in a conversation. Standardized clinical communication can be evaluated from two aspects: whether it strictly follows the communication protocol and whether it omits vital information. In addition to intent detection, we therefore need to identify the information conveyed during the communication. For each task scenario, a document indicating the required information will be provided, with each piece of information represented by sample words. However, these words can only serve as examples of the elements' expressions and cannot fully encompass them. Therefore, the challenge of the communication evaluator is the accurate recognition of the conversation content by integrating the information from both the domain knowledge and the learning examples.

1.3.3 Customizable Conversational System with Insufficient Data

For a new communication task, there may be different communication protocols, required information and evaluation schemes. Designing a new model not only necessitates professional IT personnel and a significant investment of time, but the vast majority of deep learning models also rely on a large quantity of data and careful parameter tuning. Users of a clinical communication training system would likely be medical practitioners who are unable to process large amounts of data and optimize learning models. So we need a pipeline that allows non-IT users to create new models by providing a small number of samples without adjusting the parameters and structure. The key research problem

addressed here is the customizable conversational system with insufficient training data.

1.4 Organization

- Chapter 2 provides a comprehensive review of the related work on conversational systems, the techniques, and the applications in the healthcare domain.
- Chapter 3 details the collection of a clinical handover dataset based on the ISBAR protocol and introduces the IA-LSTM for intent detection in standard clinical communication. This chapter also presents extensive experiments to validate the efficacy and generalizability of the intent-aware design.
- Chapter 4 proposes a method for constructing knowledge graphs using user-defined elements and biomedical ontology. This chapter introduces knowledge-infused prompt tuning for content recognition in clinical communication. Leveraging the created knowledge graph and pretrained language models, knowledge-infused prompts achieve significant improvement in content recognition performance, particularly demonstrating their strengths when faced with limited data.
- Chapter 5 introduces UDK-DA, a method that enables non-IT users to customize the conversational system for new tasks. This chapter also discusses the integration of the intent-aware design into the large pretrained model BioBERT. With UDK-DA and pretrained language models, a conversational system can achieve satisfactory performance on new tasks with a limited number of samples.
- Chapter 6 presents Heallo, a customizable conversational system for autonomous clinical communication training. Heallo, developed as a web application, is integrated into junior staff training programs at the local hospital. In actual use,

Heallo demonstrates performance comparable to that of clinical experts in communication evaluation. User feedback indicates that Heallo is capable of providing effective clinical communication training.

- Chapter 7 provides a summary of the work and contributions of this thesis, and discusses potential directions for future work.

Chapter 2

Related Work

In this chapter, we review the conversational systems, with a particular focus on their application in the healthcare domain. We also explore the underlying technologies that drive the development of these systems, providing a comprehensive overview of the current landscape.

2.1 Conversational Systems

Conversational systems, often referred to as chatbots, dialogue systems, or virtual agents, are computer programs designed to interact with users in natural language [115, 57]. These systems have been a subject of interest and exploration since the early days of Artificial Intelligence (AI) [1]. The journey of conversational systems began with the introduction of the first chatbot, ELIZA, in the 1960s [125]. This pioneering development sparked a wave of research and innovation in the field, leading to the creation of numerous chatbots.

Many of these early developments were driven by the desire to pass the Turing test, a

challenge designed to assess a machine's ability to exhibit intelligent behavior indistinguishable from that of a human [82]. Notably, PARRY, a chatbot that simulated paranoid behavior, managed to deceive its judges, marking a significant milestone in the field [23]. In 1995, the Loebner Competition conducted the first unrestricted Turing test, devoid of any limitations on the subject matter. The advancements in the field led to the creation of more generalized chatbots like CONVERSE [9], A.L.I.C.E [2], and Mitsuku [50].

The advent of messaging platforms has popularized social and small talk chatbots (e.g., A.L.I.C.E, Cleverbot, Simsim, Tay), while also increasing interest in task-oriented chatbots [13]. Task-oriented bots such as personal assistants have seen significant adoption, assisting users with practical daily tasks [22]. Renowned examples include Siri, Alexa, and Google Assistant [3]. This growth has catalyzed the implementation of conversational systems across various industries, including education, social media, finance, catering, and healthcare [26, 19, 121].

Chatbot developments generally fall into two categories: task-oriented and general-purpose chatbots. Task-oriented chatbots are constrained to specific subjects, such as hotel reservations or technical support services. In contrast, general-purpose chatbots aim to pass the Turing test or engage in social chitchats without any specific theme or objective [20]. These bots can also foster engaging user experiences in the open domain for entertainment or emotional connection [130].

However, the advent of ChatGPT [94] has blurred the boundaries between these two categories. Developed by OpenAI, ChatGPT is the cutting-edge application based on the Generative Pre-trained Transformers (GPT) language model [64]. It has garnered widespread recognition for its ability to generate coherent and realistic responses across a broad array of topics [81]. ChatGPT not only possesses a vast knowledge range but also exhibits understanding of numerous tasks, even those it was not specifically designed

for. It demonstrates the capabilities of foundational large language models and presents a possible trajectory towards general artificial intelligence.

Following the release of ChatGPT, a myriad of conversational systems based on large language models has emerged, including notable models like Google's BARD [84], Baidu's Ernie Bot [105], and the open-sourced ChatGLM [135] from Tsinghua University.

The rapid evolution and robust development of conversational systems underscore the immense potential and promising future of this technology. As we continue to harness the power of AI, we can expect these systems to play an increasingly pivotal role in various domains, including healthcare.

2.2 Conversational Systems in Healthcare

Conversational systems have also found numerous applications in the healthcare domain. Benifitting from their capabilities for text-based and voice-enabled interactivity, these systems can effectively cater to diverse demographic groups, making them invaluable tools for health-related interventions [18]. By facilitating instant access to medical assistance via smartphone applications or online services, they provide scalable and economically viable health support solutions [99, 12].

Studies suggest that incorporating conversational systems in healthcare settings can contribute to broader access to medical care, foster enhanced communication between patients and healthcare providers, and serve as a resource to handle the escalating need for health services [35]. This assistance can take several forms, such as facilitating remote testing, monitoring medication adherence, and offering teleconsultations.

Conversational systems have been employed across a wide array of functionalities

in healthcare, including treatment protocols [44], patient monitoring [106], diagnostic aid [41], health education [91], and as support systems for healthcare services [61]. Among these diverse applications, systems conceived for healthcare education or skill training in clinical settings share the greatest resemblance to our task.

For instance, Campillos et al. [16] provide a notable example, having designed a conversational system that impersonates a virtual patient, thereby allowing healthcare professionals to hone their medical history-taking skills. This system leverages a combination of frame- and rule-based methodologies to orchestrate medical dialogues. In a similar vein, Foster et al. [39] utilized virtual patients to impart empathic communication skills to medical students. Li et al. [73] introduced a mobile-based chatbot dedicated to helping nursing students assimilate obstetric vaccination knowledge. This innovation creates an immersive learning environment for students, fostering engagement with vaccination cases and offering a more effective learning experience than traditional lecture-based teaching.

Despite the vast potential of large language models and the existence of various successful conversational system applications in healthcare, there is a conspicuous absence of systems specifically designed for training in standardized clinical communication. Applying these models directly to our intended task—clinical communication training—is challenging. The existing models do not cater explicitly to standardized clinical communication and hence, fail to provide the meticulous understanding of user interactions that a training system necessitates. Such training prioritizes the user's ability to communicate effectively using standardized language and accurate information transmission. Furthermore, while models like ChatGPT excel in diverse tasks, they often generate deceptive illusions by providing plausible but erroneous information, which can potentially hinder the training process. Finally, these models do not readily support

user customization or the integration of user-defined tasks, a feature that is crucial when the system must adapt to new training scenarios.

Therefore, it is evident that further scholarly exploration is needed to develop conversational systems that can effectively facilitate training in standardized clinical communication. The goal is to harness the power of AI to create a system that not only understands and simulates clinical communication but also evaluates and provides feedback on the trainee's performance, thereby enhancing the quality of healthcare delivery.

2.3 Techniques in Conversational System Development

The development of conversational systems encompasses techniques ranging from traditional rule-based methods to the latest data-driven learning models [132].

Rule-based methodologies rely on predefined rules or scripts to guide the system's responses. For instance, ELIZA [125] utilized pattern matching techniques to analyze user utterances, matching them against predetermined keywords, and then generating responses based on corresponding keyword-response rules. Later developments saw the introduction of the Artificial Intelligence Markup Language (AIML) to facilitate the efficient development of rule-based chatbots [2]. While rule-based systems are relatively straightforward to construct and excel in handling structured dialogues, they often struggle with complex and unstructured interactions, limiting their versatility and adaptability.

The advent of neural networks and Deep Learning (DL) technologies has led to a shift towards end-to-end learning methods for conversational system development. Drawing from machine translation, the Sequence-to-Sequence (Seq2Seq) [120] model has emerged as a highly-utilized structure for neural language synthesis. For exam-

ple, [114] developed a neural answering machine employing the Seq2Seq architecture trained with a Sina Weibo dataset. Later, neural attention mechanisms were introduced to enhance Seq2Seq models by correlating prominent elements in the source sequence with the generated item in the target sequence [133, 7, 95]. Zhou et al. [141] integrated a memory mechanism to address the emotional nuances in large-scale conversation generation. Such deep learning methodologies have considerably advanced the design of intelligent chatbots, resulting in more human-like utterances and a more natural conversational style [132].

The recent trend in the field has shifted towards the utilization of large pre-trained language models as the foundational base in conversational systems. This approach develops applications by leveraging inherent capabilities of the large language models. For instance, the globally recognized chatbot, ChatGPT, relies on the GPT-3.5 model. When combined with Reinforcement Learning from Human Feedback (RLHF), ChatGPT is trained to produce responses more aligned with human preferences. Other prominent large models include the Bidirectional Encoder Representations from Transformers (BERT)[29], typically employed for language understanding tasks, and T5 (Text-to-Text Transfer Transformer)[102], which is designed to solve text generation problems. These models have demonstrated impressive performance across various benchmarks and can provide a robust baseline for downstream tasks.

In addition to enhancing the capabilities of the foundational models, researchers are also actively investigating the learning methodologies of these models and the nature of their interactions with users. This includes exploring Learning After Deployment, a technique that allows the model to continue learning and adapting after it has been deployed, and studying User-chatbot Interactions to better understand how users interact with these systems and how these interactions can be improved.

2.3.1 Learning After Deployment

The concept of learning after deployment acknowledges the impracticality of gathering substantial conversational data for new tasks or for interactions between a single user and a personal chatbot. Given the novelty of task scenarios, specific knowledge and interaction logic cannot always be predetermined [131]. Consequently, the most intuitive solution is to design chatbots that evolve post-deployment.

For instance, Evorus, developed by Huang et al. [55], is a crowd-powered chatbot that progresses to automate itself during its operational phase. Despite the innovation, the learning process of Evorus depends on the contributions of paid crowd workers rather than end-users. This approach not only demands substantial human supervision, but it also raises potential privacy issues.

Commonly, learning from users is accomplished by requiring their feedback during engagement [138, 72]. Existing methodologies typically engage compensated annotators to provide scalar rewards or to adhere to specific templates to ensure the input is constructive for the model [107, 136, 76]. For example, some chatbots actively learn during conversation in the question-answering (QA) scenario [127, 71].

There are also studies that allow chatbots to learn directly from natural conversations. For instance, Hancock et al. [48] propose a self-feeding chatbot capable of generating new training instances from its own conversations. Nevertheless, leaning heavily on multiple paid workers or collecting feedback via crowd-sourcing does not necessarily assure the quality of the model. This is especially true for a communication training system, which needs to be intricately designed. And indiscriminate learning can lead to undesirable outcomes, as demonstrated by the notorious incident with Microsoft's Twitter-based chatbot Tay [100]. For our communication training system, users are the

recipients of education rather than a source of learning. The cost might be significantly higher if we were to rely on professional physicians to incrementally train the chatbot through feedback.

2.3.2 Chatbot-User Interactions

The majority of contemporary research on chatbot-user interactions primarily emphasizes on interface design to enhance usability and user experience. For instance, Jain et al. [58] developed a context view to alleviate disparities between the chatbot's understanding and the user's perception of that understanding, offering users a straightforward method to edit context values. Candello et al. [17] explored the influence of different fonts on perceptions of chatbots' humanity.

Simultaneously, there is a growing interest in multi-modal interaction, incorporating visual data for example [56, 86, 28]. Some studies also investigate interactions under unique circumstances, such as Seering et al. [111] examining chatbots that facilitate or participate in multiparty or group interactions, or [5] studying the most effective methods for repairing conversational failures. Luger et al. [80] outline several design challenges arising from the gap between user expectations and experiences, including how a chatbot can reveal its current status or how to design system feedback to effectively convey the system's intent.

In many healthcare problems, emerging scenarios and requirements often result in the inadequate performance of existing models. Developing a new conversational system not only requires substantial resources, but it may also be impractical due to the challenges in obtaining dialogue data within the healthcare domain, especially for newly-emerged cases. As such, an optimal solution is to design a customizable system that

maximizes the utilization of existing resources and minimizes the cost of redevelopment.

Chapter 3

Intent Detection in Standardized Clinical Communication

Intent detection is often regarded as a semantic utterance classification problem and traditional approaches include rule-based templates [32], support vector machine (SVM) [46] and Naive Bayes [87]. With the advent of Deep Learning (DL), neural networks, such as Recurrent Neural Network (RNN) [10], have found widespread use in intent detection, resulting in significant performance gains [77].

However, there are still obstacles to applying existing intent detection algorithms directly to standardized clinical communication. First, these models require enormous volumes of labeled data [11], and extending existing intent detectors to new target domains is a resource-intensive process [103]. There are currently no publicly accessible datasets on standardized clinical handover.

Second, unlike generic intents, intents derived from clinical communication protocols (e.g., ISBAR) are not isolated but are contextually interconnected. This means that the intent of a given sentence is not only determined by its content but also by its position

within the sequence of the conversation. This interconnectedness of intents in standardized clinical communication contrasts with most previous models and underscores the need to consider sequence information when detecting intents.

To address the aforementioned challenges, we first collect a clinic handover dataset, called CLINIC-ISBAR, with the help of clinical experts from the Queen Elizabeth Hospital. CLINIC-ISBAR consists of 100 handover conversations from two real clinical emergency cases (a medical and surgical case), where sentences are labeled as different intents based on the ISBAR framework. It is worth noting that different intents in the ISBAR framework may contain the same information. For example, as shown in Table 1.1, the *interpretation of the current situation* in "Assessment" may be the same thing as the *problem* in "Situation". Therefore, the intent of a sentence can not be determined solely by its content. Along with the content of a sentence, the order in which it is presented is crucial for intent detection. This is because the meaning and intent of a sentence in a conversation can change depending on its position within the dialogue sequence. Early sentences might set the context or identify a problem, while later ones might assess the situation or make recommendations. This dynamic nature of conversation highlights the importance of considering the sequence of the conversation in intent detection. To incorporate the sequential feature in ISBAR standardized communication, we further propose a model called Intent-Aware LSTM (IA-LSTM). The Main contribution of this work is threefold:

- To the best of our knowledge, our work is the first to explore intent detection techniques for facilitating the automation of standardized clinical handover training.
- We collect a clinical handover dataset, CLINIC-ISBAR, of real-world cases with the collaboration of clinical experts. In addition to encouraging the development

of standardized clinical communication training systems, this dataset sheds light on prospective NLP applications.

- Our IA-LSTM greatly improves the performance of intent detection on the CLINIC-ISBAR dataset. The proposed intent-aware mechanism can be expanded to other baseline DL models, which further improves their performances.

3.1 Related Work

Traditional machine learning algorithms for intent detection include support vector machines, K-nearest neighbours, and decision trees [66]. With the development of deep learning, neural networks began to be utilized extensively for this task [126]. In the pipeline of DL methods, text data is originally represented using word embedding, which turns sparse word representations into dense, low-dimensional vector representations [89]. Word2vec [88], GloVe [98], and FastText [59] are examples of typical word embedding techniques. Following word embedding, various neural networks can be used for intent detection, such as convolutional neural networks [139] and recurrent neural networks (RNN) [10]. Long Short-Term Memory (LSTM) network [52], a well-known variation of RNN, has proven to be highly effective at modeling the temporal relationship of text and identifying long-term relationships. It utilizes memory cells and gates to control the flow of information and addresses the gradients vanishing and exploding issues encountered in standard RNN training [52]. On the basis of the LSTM structure, additional advancements have been made by incorporating bidirectional mechanisms [43], attention mechanisms [8], hierarchical structures [93], and convolutional layers [68].

Despite their superiority in processing sequential data, RNN-based models require

input data to be processed sequentially, hence limiting training speed. Transformer [122] resolved this problem by utilizing just self-attention blocks - all tokens are processed simultaneously and attention weights between them are calculated. In this manner, Transformer supports increased parallelization during training and permits training on larger datasets. It is currently replacing older RNN models and resulting in the creation of huge pre-trained models [128]. One of the most well-known pre-trained models, BERT (Bidirectional Encoder Representations from Transformers) [29], has attained cutting-edge performance in a wide range of NLP tasks. The BERT-base model was pre-trained on a 3.3 billion word corpus and consists of 12 layers of transformer blocks with 100 million in parameters [29]. Then, it can be fine-tuned for the downstream tasks without altering the architecture.

In addition to the advances in network designs, major advancements have been made in intent detection by incorporating additional information, such as learning with external knowledge [123], and taking slot-filling as joint tasks [63]. As described in Section 1.2.2, the sequence in which sentences appear is crucial to understanding the intent of the present sentence in standardized clinical handover. Hence, the sequencial information can be used as additional knowledge for intent detection. The way this sequential information is modeled varies between classification tasks. For example, [129] developed a propagation graph to illustrate the chronology of message distribution for online rumor detection, and [142] utilized headings and sentence positions as the sequential information for distinguishing portions of the medical report.

ISBAR is recommended to organize the contents of a standardised clinical handover conversation. Specifically, ISBAR divides clinical handover to five intents (i.e., Identify, Situation, Background, Assessment, and Recommendation), which are interconnected in a particular order. Thus, the sequence in which these intents appear can repre-

sent the sequential information in a clinical handover. In contrast to earlier work where the sequential information is derived from known factors such as message chronology or article structure [134, 142], in our task, the actual order of intents is unknown. Therefore, order of detected intents can be used to simulate the sequential information that is revealed as intent detection continues.

Based on this observation, we incorporate the sequential information by adding an additional feature that indicates the detected intents. One of the most common methods of representing intent labels is by using one-hot vectors [10]. However, one disadvantage of using one-hot encoding is that it ignores information in the non-dominant dimensions and may magnify errors when the intent prediction results are incorrect. To overcome this defect, we use the probability distribution after the SoftMax layer to represent the intent label in our proposed IA-LSTM model.

3.2 The CLINIC-ISBAR Dataset

This section presents the CLINIC-ISBAR dataset, a clinical handover dataset of real clinical emergency cases. This dataset serves as a foundation for the development and evaluation of computational models aimed at processing and analyzing standardized clinical handover. The processes of data acquisition, annotation, and analysis are discussed in detail in the following subsections.

3.2.1 Data Collection

Our CLINIC-ISBAR dataset was collected in Queen Elizabeth Hospital, Hong Kong, involving two real clinical handover scenarios between junior and senior doctors. The first scenario was a medical (MED) patient with respiratory failure who may require

elective intubation. The second scenario was a surgical (SURG) instance in which a patient with an acute abdominal injury would require emergent surgery.

During the handover procedure, a junior doctor conveyed the case to a senior doctor based on medical records, notes, and testing reports (e.g., hematology reports, chemical pathology reports, CT scans). To prevent any breach of patient confidentiality, all personally identifiable information was replaced with synthetic data. In total, 100 high-quality audio recordings were collected, comprising 48 MED scenarios and 52 SURG scenarios.

3.2.2 Annotation

A rigorous annotation pipeline was developed for the conversion of audio recordings into annotated textual data. The audio recordings were first transcribed manually, with transcriptions subsequently segmented into sentences. A panel of domain experts, with at least three annotators per sentence, labeled each sentence according to the ISBAR framework.

The ISBAR framework comprehensively covers the possible intents in a clinical handover, making it unlikely for irrelevant content to appear in a handover of an emergency case. However, there were instances where a sentence could be categorized under multiple labels or did not clearly fall into any of the five intents. These sentences were deemed controversial and were filtered out, accounting for less than 1% of the total.

The resultant dataset comprises 1895 textual sentences, each annotated and categorized into one of five label classes as per the ISBAR framework. Table 3.1 provides an overview of the distribution of samples across different intent categories. For illustrative purposes, Figure 3.1 showcases snippets of annotated data from the CLINIC-ISBAR

Intent	MED (48 recordings)	SURG (52 recordings)
I (Identify)	112	118
S (Situation)	49	48
B (Background)	251	262
A (Assessment)	423	339
R (Recommendation)	145	148

Table 3.1: Statistics of the CLINIC-ISBAR dataset.

dataset. It should be noted that not all conversations in the dataset strictly adhere to the ISBAR framework in terms of the sequence or inclusion of all five components. This variation is reflective of real-world clinical communication.

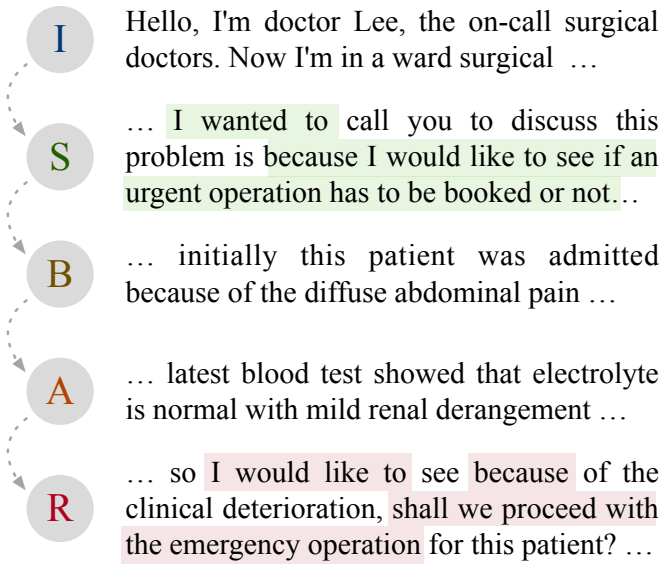

Figure 3.1: Fragments of a conversation in CLINIC-ISBAR with corresponding intents. I: Identity; S: Situation; B: Background; A: Assessment; R: Recommendation.

3.2.3 Analysis

A range of baseline DL models are investigated on the CLINIC-ISBAR dataset (see Section 3.4 for details).Among these, BERT [29] yields the best classification accuracy,

scoring 84%. However, there is still considerable room for improvement. We attribute this to several intrinsic challenges associated with intent detection on the CLINIC-ISBAR dataset:

- Our CLINIC-ISBAR dataset is relatively small and it contains many domain-specific terminologies, making it challenging to scale existing intent detectors from other domains to our target domain.

- The content within one conversation in our dataset is semantically concentrated, meaning that the sentences are all related to the same biomedical case and thus share similar themes and vocabulary. This is in contrast to other intent detection tasks, such as those for a personal assistant, where intents like 'search a restaurant' and 'play music' can be very different and easy to distinguish. The similarity of the content within a conversation and the variety of expressions across different conversations require a more accurate and precise understanding of the sentences.

- Five intents of ISBAR (i.e., identify, situation, background, assessment, and recommendation) have semantically-vague boundaries and are sequentially related to each other.

In addition to the content, we also observe that the sequential position of the sentence in a clinical handover is essential for intent detection. To illustrate how a sentence's position influences sentence understanding, we highlight two sentences in a sample conversation #48 (see Figure 3.1). The highlighted sentences are very similar in terms of vocabularies, patterns and even semantics, but are labelled as different intents. The green sentence at the top of the handover conversation (behind intent I) can be interpreted as the reason for calling (labeled as S). While located at the end of the handover conver-

sation (behind intent A), the red part intends to express the clinician's recommendation (labeled as R).

Considering the sequential structure of the ISBAR framework, we propose to incorporate previous intent information into intent detection and design an intent-aware method based on LSTM structure.

3.3 Intent-Aware LSTM (IA-LSTM) Model

In this section, we formulate the problem of sentence-level intent detection during an ongoing clinical handover and describe in detail our proposed IA-LSTM model.

3.3.1 Problem Formulation

Given a conversation with N sentences from the clinician's side, we denote it as $\mathcal{D} = \{(s^{(n)}, y^{(n)}) \mid n \in \mathbb{Z},\ 1 \leq n \leq N\}$, where $s^{(n)}$ is the n-th sentence and $y^{(n)}$ is the corresponding intent label represented in one-hot encoding. We further denote the n-th sentence as a sequence of word embedding $s^{(n)} = (w_1, \ldots, w_t, \ldots, w_T)$, where T is the number of words in $s^{(n)}$, and $w_t \in \mathbb{R}^D$ is a D-dimensional word embedding of the t-th word. In the ongoing setting, we only have the first n sentences of the conversation when $s^{(n)}$ is given out. Thus, we formulate the problem as an objective of learning a model G for intent detection on a subset of the conversation, which can be written as

$$\hat{\boldsymbol{y}}^{(n)} = G(\boldsymbol{s}^{(n)}, \Theta) \tag{3.1}$$

where Θ is the parameters of the model G, $\hat{\boldsymbol{y}}^{(n)} = (\hat{y}^{(1)}, \ldots, \hat{y}^{(n)})$ is the model predictions given the input sentences $\boldsymbol{s}^{(n)} = (s^{(1)}, \ldots, s^{(n)})$ in the conversation $\mathcal{D}$.

As depicted in Figure 3.2, applying Deep Learning (DL) models to intent detection typically entails three key stages: preprocessing, vectorization, and classification. In the preprocessing stage, raw input sentences undergo transformation through standard techniques such as capitalization normalization, noise removal, and tokenization, converting the text into a series of tokens. These tokens collectively form the vocabulary for the subsequent stages. During the vectorization stage, each token is mapped to a vector through word embeddings, resulting in a sequence of vectors. Using embedded vectors as input, the final step classifies the intent using neural networks like RNN and CNN. In a clinical handover dialogue, sentences are given consecutively in the ongoing manner, corresponding to a chain of intents.

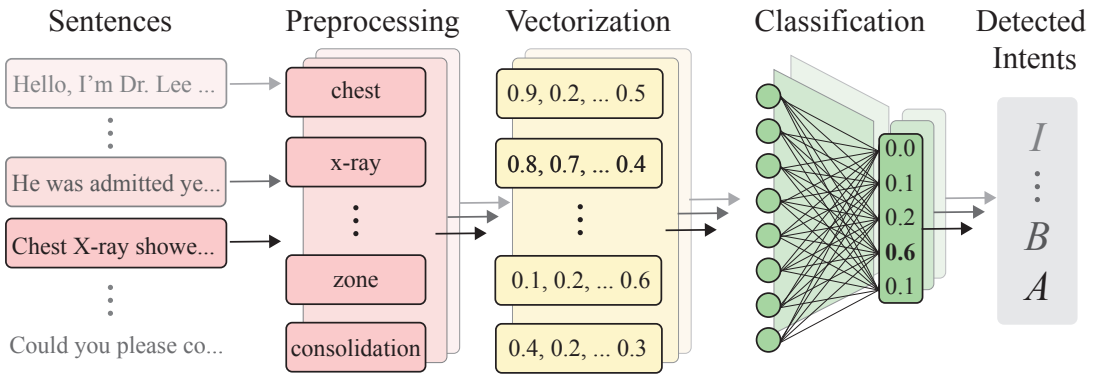

Figure 3.2: Procedures of intent detection. There are three phases involved: preprocessing, vectorization, and classification. Under the ongoing setting, we predict the intent vector of the first n sentences in the clinical handover.

3.3.2 Intent-Aware LSTM

3.3.2.1 Sentence Representation

Given the vectorized word embedding $s^{(n)} = (w_1, ..., w_t, ..., w_T)$, many DL models could be adopted to learn representations of the input sentences. It is still debatable

which DL model performs better, particularly on relatively small datasets [34]. For the ease of understanding, we adopt a basic LSTM [52] as the backbone model. The LSTM unit regulates the flow of information from past stages to the present phase using three gates: an input gate, an output gate, and a forget gate. At each time step $t \in [1, ..., T]$ for its corresponding embedding w_t, LSTM calculates its current hidden state output vector h_t based on a memory cell c_t and an output gate g^o as

$$\begin{aligned} g^o &= \sigma(W^o h_{t-1} + I^o w_t) \\ h_t &= tanh(g^o \odot c_t) \end{aligned} \tag{3.2}$$

where W^o and I^o are weight and projection matrices, respectively. σ represents the logistic sigmoid function, and $\odot$ is the element-wise multiplication. While the memory cell c_t is calculated with three gates that can be defined as

$$\begin{aligned} g^c &= \sigma(W^c h_{t-1} + I^c w_t) \\ g^f &= \sigma(W^f h_{t-1} + I^f w_t) \\ g^u &= \sigma(W^u h_{t-1} + I^u w_t) \\ c_t &= g^f \odot c_{t-1} + g^u \odot g^c \end{aligned} \tag{3.3}$$

where g^c, g^f, and g^u are the activation vectors of the cell state, output, and input gates, respectively; The recurrent weight matrices are denoted by W^c, W^f, and W^u; The projection matrices are represented as I^c, I^f, and I^u. For the input sequence $s^{(n)} = (w_1, ..., w_t, ..., w_T)$, the latest hidden state $h_T$of the LSTM model is used as the sentence representation $\hat{s}^{(n)}$ for intent detection.

3.3.2.2 Intent-Aware Design

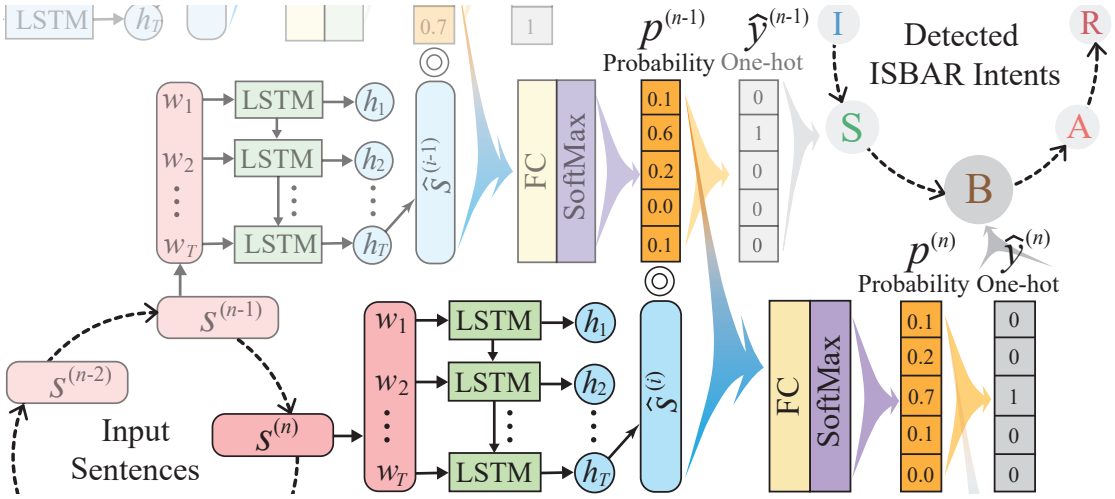

Figure 3.3: Structure of IA-LSTM. The intent-aware mechanism incorporates intent information (i.e., $p^{(n-1)}$, the probability distribution following the SoftMax layer) of the preceding sentence $s^{(n-1)}$ with the current sentence representation $\hat{s}^{(n)}$ learnt from the LSTM backbone model (i.e., the last hidden state h_T).

Figure 3.3 outlines the architecture of the proposed IA-LSTM model, which incorporates an intent-aware mechanism. Unlike standard intent detection, our approach accounts for the contextual relationship between consecutive intents in clinical handover dialogues. By incorporating the probability distribution of the preceding intent into the current intent detection process, IA-LSTM encapsulates this relationship.

Let $p^{(n)} \in \mathbb{R}^C$ denote the probability distribution vector of the intent information for the n-th sentence $s^{(n)} \in \mathbb{R}^{D \times T}$, where C is the number of intent labels. Given the current input sentence $s^{(n)}$, its preceding intents are denoted as $\boldsymbol{p}^{(n-1)} = (p^{(n-k)}, \ldots, p^{(n-1)}), 1 \leq k < n-1$ in a probability distribution format.

Figure 3.3 illustrates the model structure when $k = 1$ (i.e., $\boldsymbol{p}^{(n-1)} = p^{(n-1)}$). To compute $p^{(n)}$, the model first concatenates the most recent intent probability distribution vector $p^{(n-1)}$ with the current sentence representation $\hat{s}^{(n)}$. Then, the IA-LSTM makes

predictions using a fully connected layer and a SoftMax layer based on this concatenated representation. This process can be formulated as

$$p^{(n)} = SoftMax(F([\boldsymbol{p}^{(n-1)}, \hat{s}^{(n)}])) \tag{3.4}$$

where F is the fully-connected layer. For the beginning sentence in conversation $\mathcal{D}$, the preceding intent information is specified as a C-dimensional zero vector.

3.3.2.3 Optimization

Given the proposed IA-LSTM model G as defined in Equation 3.4, we use the cross-entropy loss to optimize it, which can be written as

$$\arg\min_{\Theta} - \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} log(p_c^{(n)}) \tag{3.5}$$

3.4 Experiments

In this section, we evaluate our IA-LSTM on the CLINIC-ISBAR dataset from three aspects:

- Effectiveness on intend detection: We run a number of representative baseline models on the CLINIC-ISBAR dataset and compare their performance with IA-LSTM. This is to verify the effectiveness of using proceeding intents as sequential information.
- Generalization to other models: We investigate the generalization ability of the intent-aware design by expanding it to other DL models.

- Robustness of the probability representation: We conduct experiments to verify whether the probability distribution is a better representation of intent information than one-hot encoding.

3.4.1 Baselines

LSTM [52] is identical to the backbone model utilized in IA-LSTM. We use 1 layer and a hidden size of 16 for LSTM and all of its variants.

Bidirectional LSTM (BiLSTM) [43] employs two LSTMs that receive input in both the forward and reverse directions.

Attention-based LSTM (AttLSTM) [8] learns attention information from the embedding representation to direct the model's attention to particular areas of the sequence for the classification task.

TextCNN [139] is an implementation of CNN for NLP applications in which the word embeddings are fed into three distinct convolutional layers and their output concatenated to a linear layer. In accordance with the configuration in the original work, we employ three kernel sizes $(2, 3, 4)$ and five kernels for each.

Recurrent CNN (RCNN) [68] represents a sentence with a concatenation of GloVe [98] word embedding and the output of BiLSTM .

Transformer [122] is a multi-head self-attention structure that has outperformed RNN/CNN based models on machine translation tasks with faster training speed. We used a 1-layer two-head encoder and averaged the output layer of the encoder before connecting it to the fully connected layer.

BERT [29] is a deep bidirectional Transformer architecture that has been pre-trained using a 3.3 billion word corpus.It has demonstrated cutting-edge performance on nu-

merous NLP tasks. Here we fine-tune on the BERT-base model and connect the output of the first token (the [CLS] token) to a fully connected layer.

Concatenate BiLSTM (CLSTM) [142] uses the nearby sentences processed by BiLSTM to aid classification of the current sentence. In our experiment, k sentences before the current sentence are employed to model interdependencies at the sentence level.

3.4.2 Experimental Settings

Here we introduce the experimental setting for CLINIC-ISBAR in terms of dataset splits, hyper-parameter selection, and evaluation metrics.

Dataset Splits. To ensure the integrity of the clinical handover, we split the dataset by conversation. Specifically, We separated all sentences in a ratio of $6:2:2$ into train, valid, and test sets. Table 3.2 depicts the distribution of sentences for each intent.

Split Item	Total	I	S	B	A	R
#Train	1159	141	61	327	455	175
#Valid	366	40	18	91	151	66
#Test	370	49	18	95	156	52

Table 3.2: Dataset splits of CLINIC-ISBAR.

Hyper-parameters Selection. For embedding initialization, we employ glove.6B.50d [98] (save for BERT), which is trained on Wikipedia 2014 and Gigaword5 with $6B$ tokens and $400K$ vocabularies. For Transformer and BERT, the sentence length is set at 32, whereas other models accept inputs with various sentence lengths. We use a batch size of 16 to train BERT and 1 for other models. And $k = 1$ is set for CLSTM and IA-LSTM. During the training process, Adam optimizer is utilized for all models. We set the learning rate at 0.003 and the dropout rate at 0.02. We set the learning rate to $1e-5$ for parameters in the original BERT model.

To provide a fair comparison, we execute five rounds of trials for all models using five random seeds (1, 12, 123, 1234, and 12345) and record the test accuracy when each model performs best on the validation set within 50 epochs. We report the five-round average test accuracy for all models.

Evaluation Metrics. We report the $Accuracy$ and $Macro\ F1\text{-}Score$ as performance measures. The $Accuracy$ is calculated as

$$Accuracy = \frac{\#\ of\ correct\ predictions}{Total\ \#\ of\ intents} \tag{3.6}$$

For $Macro\ F1\text{-}Score$, it averages $F1\text{-}Score$ of each intent class:

$$Macro\ F1\text{-}Score = \frac{1}{C}\sum_{i=0}^{C} F1\text{-}score_i \tag{3.7}$$

where i is the intent index and C the number of intents. The $F1\text{-}Score$ for the given intent class of i can be calculated based on the harmonic mean of precision and recall

$$F1\text{-}Score_i = 2 * \frac{TP_i}{TP_i + 0.5(FP_i + FN_i)} \tag{3.8}$$

where TP_i, FP_i, and FN_i denote the number of True Positive, False Positive and False Negative cases of intent i, respectively.

3.4.3 Results and Discussion

In this section, we present and discuss the experimental results from the three aspects introduced in the beginning of this section.

Model	Accuracy	F1-Score
LSTM [52]	81.84	77.47
BiLSTM [43]	79.78	74.47
AttLSTM [8]	81.62	78.04
TextCNN [139]	79.90	74.14
RCNN [68]	82.86	78.51
Transformer [122]	78.92	73.39
BERT [29]	84.86	81.09
CLSTM [142]	83.68	84.36
Our IA-LSTM	**88.43**	**85.76**

Table 3.3: Performance of baselines and IA-LSTM (%).

3.4.3.1 Effectiveness on Intent Detection

Table 3.3 shows a comparison of baseline models and the proposed IA-LSTM on the CLINIC-ISBAR dataset. Among all baseline models, BERT achieves the highest accuracy, 84.86%. And CLSTM improves BiLSTM by considering the interdependencies of nearby sentences, reaching an accuracy of 83.68%. Representing the sequential information by intent labels, our IA-LSTM outperforms all baselines with noticeable improvements: our model's performance surpasses the state-of-the-art BERT with an enhanced accuracy of 3.57%. Our IA-LSTM also significantly improves the results of its backbone model LSTM (from 81.84% to 88.43%), demonstrating the effectiveness of our intent-aware design. With an accuracy of 88.43%, our model can feasibly be deployed to a standardized clinical communication training system.

One interesting finding here is the performance of BERT. A large DL model with millions of parameters like BERT tends to overfit on small datasets, resulting in limited advantages [119]. Work of [34] has shown that the simplest LSTM architecture consistently outperforms BERT with small datasets. However, on our CLINIC-ISBAR – a relatively small dataset – BERT yields the best performance among all baselines when

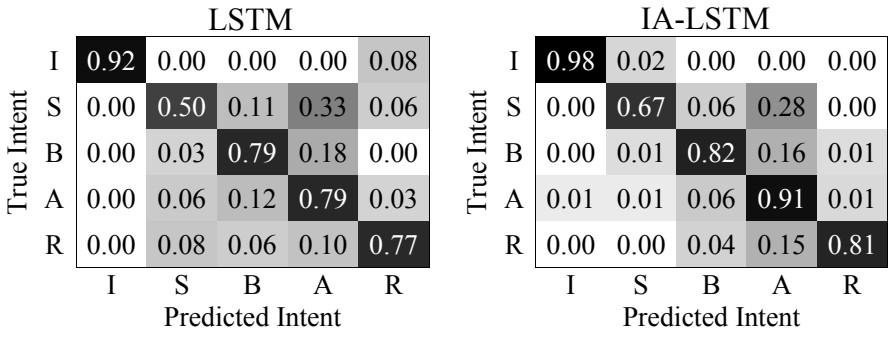

Figure 3.4: Confusion matrices of LSTM and IA-LSTM on CLINIC-ISBAR.

sequential information is not involved. This evidences the complexity of our dataset from another side. More concretely, the sentences in a clinical handover are semantically similar – they all provide information related to a specific clinical case. Thus, detecting the intent behind them requires a precise and comprehensive understanding. A huge and deep structure like BERT can thus detect subtler differences and deeper meaning, leading to better performance on our dataset.

Discussion: how does LSTM benefit from the intent-aware design? We present the confusion matrices for LSTM and IA-LSTM to study further how IA-LSTM enhances intent detection in each class (see Figure 3.4). These matrices demonstrate that all detection accuracies are improved along the diagonal, with the detection accuracies of intents A and S exhibiting the most notable enhancements. This result is consistent with the fact that S and A are closely related and difficult to differentiate. Our intent-aware design enables the model to look at the preceding sequences and infer the difference.

	Sentence	**Intent**
(Previous)	"And I have revealed the CT scan ..."	A
(Target)	"So I think the problem is that the patient suffered from an acute gangrenous appendicitis and probably and very likely perforation."	A

Table 3.4: Two consecutive sentences in a conversation.

Table 3.4 exhibits two consecutive sentences from a conversation that LSTM incorrectly classifies. Without knowing the intent of the previous sentence, LSTM classifies the target sentence as S (Situation). Indeed, this sentence could serve as a summary of the patient's situation or a reason for calling. Nevertheless, it is evident from the preceding statement that the next sentence presents an assessment. IA-LSTM is able to correctly classify the target sentence as A (Assessment) when provided with the intent information.

We have analyzed the performance of IA-LSTM when $k = 1$, which outperforms all baseline models (see Table 3.3). To further validate the effectiveness of our IA-LSTM, we perform ablation on the value of k, which is the number of preceding intents incorporated in the model. Figure 3.5 shows the performance of IA-LSTM when choosing different values of k. With varing values of k, the model maintains a stable performance that is much superior to when no intents are incorporated (Accuracy 81.84%, F1-Score 77.47%). Observations indicate that raising the value of k can further enhance performance, with the best accuracy and F1-Score being being 90.91% and 90.51% when $k = 4$. This again highlights the effectiveness and potential of intent-aware design. From the sharp improvement before $k = 4$ and the slight decline after, we can infer that the most recent preceding intents are more crucial to understanding the current sentence, whereas the farther-off intents only provide a limited amount of information. This conforms to the characteristics of typical human communication.

3.4.3.2 Generalization to other DL Models

Given the efficacy of our intent-aware design, we intend to apply it to more baseline models. Most DL models for text classification contain a fully connected layer as the final layer taht predicts the label. We refer the input of the final layer as a sentence rep-

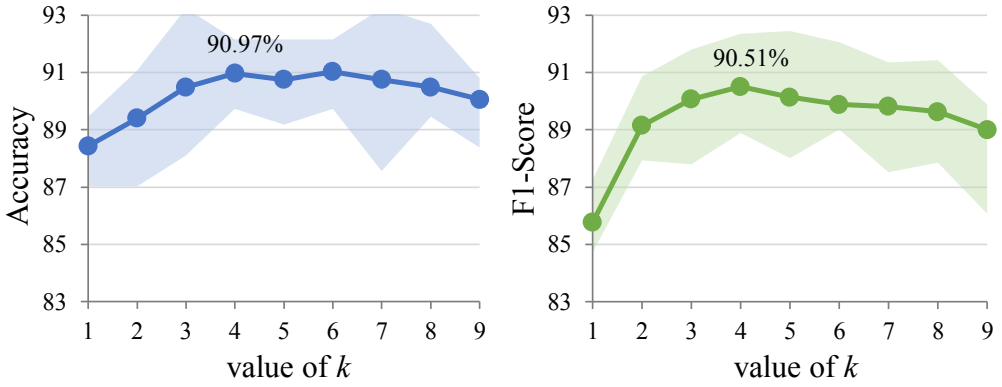

Figure 3.5: The performance of IA-LSTM with varing k values (%). The nodes on the lines indicate the average results of implementations with 5 different seeds. The shaded regions represent the results' upper and lower bands.

resentation $\hat{s}$. Based on the same idea in IA-LSTM, we extend our intent-aware design to general DL models by concatenating the intent vector with $\hat{s}$ and passing it to a fully connected neural network. For the ease of comparison, we set k to 1 for all the expanded implementations.

Model	Without Intent-aware		With Intent-aware ($k = 1$)	
	Accuracy	F1-Score	Accuracy	F1-Score
LSTM	81.81	77.51	88.41 (↑**6.61**)	85.81 (↑**8.31**)
BiLSTM	79.81	74.51	88.41 (↑**8.61**)	86.61 (↑**12.11**)
AttLSTM	81.61	78.01	90.11 (↑**8.41**)	87.91 (↑**9.91**)
TextCNN	79.91	74.11	88.41 (↑**8.51**)	86.71 (↑**12.61**)
RCNN	82.91	78.51	90.31 (↑**7.51**)	88.81 (↑**10.31**)
Transformer	78.91	73.41	88.11(↑**9.21**)	85.81 (↑**12.41**)
BERT	84.91	81.11	88.21 (↑**3.31**)	86.41 (↑**5.31**)

Table 3.5: Performances of baselines with and without the intent-aware design (%).

A comparison of baseline models and their intent-aware versions is shown in Table 3.5. Notably, our method for adding proceeding intent information is adaptable to a variety of model topologies and model sizes. The performance of all models improve

significantly, with RCNN achieving the highest accuracy of 90.31%. Consistent with our observation, this general improvement is attributable to substantial sentence correlations.

Discussion: why is the improvement of BERT not that significant? The BERT-base model used in this paper contains 110M parameters, which is about a thousand times more than the other baselines. In order to extend our intent-aware design to BERT, we concatenate the intent vector p with $\hat{s}$ generated by BERT (the output of the [CLS] token) and pass it to a fully connected layer. It is worth noting that different models may have different dimensions of $\hat{s}$, but the same dimension of p (i.e., 5). For other DL models in the experiment, the dimension of $\hat{s}$ is between 16 to 50: LSTM has $\hat{s}$ of dimension 16, BiLSTM 32, AttLSTM 16, TextCNN 15, RCNN 16, and Transformer 50. However, BERT generates a $\hat{s}$ of dimension 768, which is significantly larger than other baseline models. When confronted with this dominant vector size (768 vs. 5), the intent-aware design still increases BERT's accuracy by 3.3%, confirming the effectiveness and generalizability of our intent-aware design.

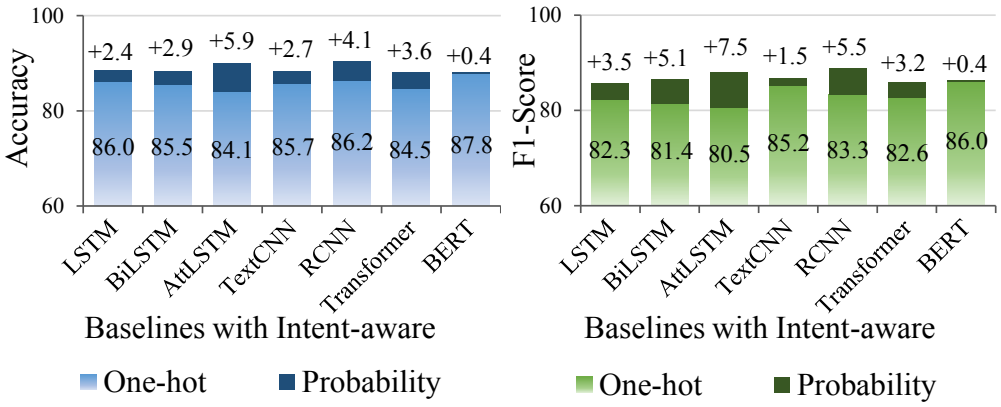

Figure 3.6: Performance of intent-aware models when using one-hot encoding and the probability distribution as the intent vector (%).

ID	Sentence	True	P	O
#859	Mr. Chan is 69-year-old gentleman admitted under medical for shortness of breath and ankle oedema for two days.	B	A	A
#860	Basically he has signs and symptoms of CHF with orthopnea and PND and he was admitted on the 14th which was two days ago.	B	B	A
#861	He has a background history of type 2 diabetes, hypertension, ischemic heart disease with previous PCI done in 2015 and mild renal impairment.	B	B	A
#862	He has previous follow up in the QEH medical.	B	B	A

Table 3.6: Continuous sentences predicted by IA-LSTMs. Three columns on the right indicate true intents, intents predicted by IA-LSTM with the probability distribution (P), and intents predicted by IA-LSTM with one-hot encoding (O).

3.4.3.3 Robustness of the Probability Distribution

In previous work, the intent vector is usually derived from the intent label, using one-hot encoding [10]. Figure 3.6 depicts the performance of all intent-aware models using the intent vector represented by one-hot encoding and the probability distribution. Findings show that the probability representation is consistently better than one-hot representation for all models, potentially due to two reasons. First, the probability distribution reserves more sequential information than the one-hot representation [51] – it contains leaked information of non-dominant intents. Second, the probability distribution is more false-tolerant because it could smooth the prediction error of its former sentence.

Discussion: how the probability distribution functions differently from one-hot encoding? Table 3.6 lists the result of continuous sentences predicted by implementations of IA-LSTM with the probability distribution (P) and one-hot encoding (O). When both models predict the wrong intent for sentence #859, with the probability distribution, this classification error does not influence the subsequent predictions. However, with one-hot encoding, this error is passed on and leads to a series of wrong predictions.

By examining the output after the SoftMax layer of both models (see Figure 3.7), we can see how the probability distribution and one-hot encoding works differently. When using the probability distribution as the intent vector, the classification error was gradually eliminated in the subsequent predictions; however, when the one-hot vector is used, there is little chance to correct the wrong prediction because the error is constantly enlarged every time one-hot encoding is performed.

	Sentence		Value in SoftMax Output					Predict	
	ID	Label	I	S	B	A	R	Result	Correct
Probability	#862	B	0.012	0.053	**0.003**	**0.932**	0.000	A	✗
	#863	B	0.000	0.000	**0.348**	**0.652**	0.000	B	✓
	#864	B	0.000	0.000	**0.599**	**0.401**	0.000	B	✓
	#865	B	0.000	0.000	**0.988**	**0.012**	0.000	B	✓
One-hot	#862	B	0.005	0.010	**0.007**	**0.978**	0.000	A	✗
	#863	B	0.000	0.000	**0.176**	**0.823**	0.000	A	✗
	#864	B	0.000	0.000	**0.016**	**0.984**	0.000	A	✗
	#865	B	0.000	0.000	**0.034**	**0.966**	0.000	A	✗

Figure 3.7: Output after the SoftMax layer when detecting continuous sentences using the probability distribution and one-hot encoding.

3.5 Conclusion

In conclusion, we collect a standardized clinical handover dataset, CLINIC-ISBAR, of real-world cases, and propose a novel intent-aware algorithm IA-LSTM based on the sequential structure of the ISBAR framework. Extensive experiments and comparisons on our dataset have verified the effectiveness, generalization ability and robustness of our intent-aware design. The collected dataset and proposed algorithm lay a foundation for the deployment of clinical communication training systems.

Looking ahead, we see potential for further refining our approach. One possible

direction is to extend our model to handle multi-label intent detection, which would allow a sentence to be associated with multiple intents. Another promising avenue is to adapt our model for outlier detection, enhancing its ability to recognize and handle sentences that do not belong to any of the intent labels, thereby increasing its robustness and adaptability to real-world scenarios.

We hope that this initial attempt of integrating NLP technology into clinical communication training will encourage the creation of intelligent tools for communication training and motivate NLP researchers.

Chapter 4

Content Recognition with Knowledge-Infused Prompt

A standardized clinical communication are usually evaluated from two aspects: whether the conversation follows the steps specified in the communication protocol; and whether all pertinent information is mentioned in the dialogue. We have already identified the sentence-level intents in chapter 3, which correspond to the steps in a communication protocol. And this chapter focuses on detecting the information conveyed in the dialogue.

For each clinical communication scenario, a document is provided by the clinical experts indicating the essential information that need to be conveyed in this case. Table 4.1 illustrates some essential information that should be included in the clinical handover case of a patient with respiratory failure. Essential information is presented as a list of components, and the "Sample Words" column contains example expressions corresponding to each component. Unlike the unified communication protocol, different communication scenarios usually have different required components.

No.	Components	Sample Words
1	Patient's name	Mr. Wong Hong Kin
2	Ward location	Ward E6, Bed 15
3	Patient's age	69 years old
4	Admission time	14th February
5	Respiratory failure	Respiratory failure, Increase breath
6	Ischemic heart disease	Ischemic heart disease, IHD
7	Blood test	ABG, CBC, CRP
8	Diabetes mellitus	Diabetes mellitus
9	Saturation	Saturation, 91%
...	...	...

Table 4.1: Examples of required components and corresponding sample words in the medical scenario.

The examples provided reveal the dissimilar form and granularity of the enumerated components. Certain components closely resemble entities within the general domain, such as names and ages, whereas others align with concepts in the biomedical domain, such as a specific disease or a kind of symptom. Due to the varying granularity of these components, their coverage may differ significantly, and there may even be inclusion relationships between them. Furthermore, certain components refer to descriptive statements, which may not have clear boundaries.

Instead of explicitly marking the precise location of a component within a conversation, the objective of this task is to detect a particular component within a sentence. Intuitively, one might compare the words in the document to those in the conversation (i.e., lexical matching) to determine the presence of corresponding components. However, this approach suffers from two noteworthy limitations. Firstly, lexical matching is extremely sensitive, as even slight changes in morphology or word order may result in failed mappings. Secondly, sample words only serve as examples of the expressions for corresponding components, and thus cannot encompass all possible cases. For ex-

ample, as depicted in Table 4.1, "diabetes mellitus" may also be expressed as "diabetes" or "DM."

Nevertheless, this intuitive approach provides insights to the formulation of the problem into a task of classifying sentence-component pairs. This entails identifying whether a given component is present within a specific sentence. However, the limited amount of training data and the diverse forms of components pose a significant challenge to the model's comprehension ability. To address this challenge, in addition to utilizing sample words from the document, we aim to incorporate external knowledge to assist in the task. Prompts, which serve as cues for language models, can be used to reformat the downstream task and make it more familiar to the language model. Previous research has demonstrated that prompts can enhance the exploitation of knowledge embedded within large-scale pre-trained language models [78, 117]. However, a substantial proportion of the component list comprises clinical terms that may not be adequately addressed by a language model trained on general corpora. In this regard, we propose integrating Knowledge-Infused Prompt with pre-trained language model for content recognition. Initially, we utilize biomedical ontology to augment the sample words and establish a local knowledge graph. Then the proposed infusing method infuses knowledge into prompts by transforming the vertices and edges in the structured graph into textual descriptions.

Our experimental results indicate that using knowledge-infused prompt with large language models can significantly enhance content detection accuracy by leveraging both pretrained language models and knowledge graphs. Particularly in the case of limited training data, knowledge-infused prompt can facilitate faster convergence of the model and achieve superior performance.

4.1 Related work

This section provides an overview of the developments in prompt learning and knowledge graphs, and their relevance to our research.

4.1.1 Prompt Learning

The introduction of large-scale pretrained language models (PLMs) such as BERT [30], GPT [101], and T5 [102], has catalyzed a significant paradigm shift in the domain of natural language processing. The erstwhile *fully supervised learning* approach has been progressively replaced by the *pretrain and fine-tune* paradigm. In recent times, the emergent trend of *prompt learning* has garnered considerable interest within the research community [78].

Prompt learning, by providing task-specific prompts or cues, has demonstrated its efficacy in enhancing the performance of PLMs across a broad spectrum of NLP tasks [79, 47, 14]. As illustrated in Figure 4.1, a typical prompt template consists of the original input, prompt words (which serve as conditions or cues for the language model), and a masking token ⟨X⟩. The prediction is based on the probability that label words are filled in the ⟨X⟩ token. Now that the task follows the same format as the PLM pretrained, prompt-oriented tuning does not necessitate additional neural layers and make the turning process more efficient.

Despite being a nascent field, the construction and utilization of prompts have been subjects of gradual exploration. For instance, Gu et al. [45] proposed pretraining prompts by incorporating prompts into the pretraining stage, thereby achieving a more effective initialization of soft prompts. And Li and Liang [75] proposed prefix-tuning, which learns “virtual tokens” for a prompt.

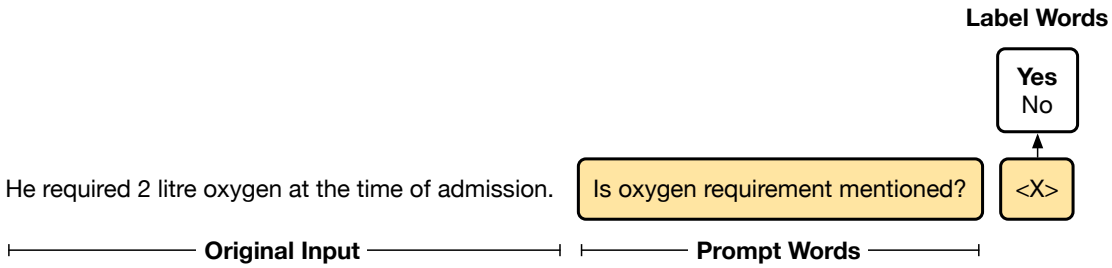

Figure 4.1: An example of a prompt. The yellow rectangles in the figure are prompt tokens, where ⟨X⟩ represents the masking token.

Recent research endeavors have also sought to integrate external knowledge into the design of prompts. For example, Hu et al. [54] incorporated external knowledge into the verbalizer (the mapping from label words to the specific class). And Han et al. [47] designed prompts based on rules extracted from the prior knowledge of a classification task. Chen et al. [21] incorporated knowledge among relation labels into prompt-tuning for relation extraction, utilizing learnable virtual type words and answer words.

These pioneering works underscore the potential of prompts in bolstering the performance of pre-trained language models and offer valuable insights into the design of prompts and the application of knowledge. However, there is a conspicuous absence of studies exploring the integration of knowledge from the biomedical domain for clinical content recognition. The field of prompt tuning is still in its early stages, and there is much room for exploration and improvement.

4.1.2 Knowledge Graph

The term Knowledge Graph (KG) is fist coined by Google when they used semantic knowledge in web search; it has since been adopted by the scientific community [49]. Knowledge graph represents structured information as a semantic graph composed of in-

terconnected nodes and edges [92]. Nodes in a knowledge graph are entities or literals, while edges between nodes represent their semantic relationships. Due to its ability to provide semantically structured information, it has been explored in a variety of applications, including question answering and information retrieval [143]. DBpedia, Google's Knowledge Vault, Wikidata, Microsoft Satori, and Facebook's entity graph are examples of well-known knowledge graphs [97, 33, 92].

These knowledge graphs cannot be directly applied to our problems, because the specific relations between entities needed for content recognition cannot be reflected in those knowledge graphs. Many researchers also investigate mining knowledge from medical documents and constructing knowledge graphs [140, 116, 74]. It typically requires multiple processes, such as named entity recognition, relation extraction, graph embedding, etc [74]. Different from the common process of constructing knowledge graphs from documents, our data sources are relatively structured: one is user-provided documents with defined entities and the other is existing biomedical ontology; and the relations in our knowledge graph are known.

Compared with generic conversation, clinical communication involves complicated medical terminology, posing additional challenges for language modeling [60]. A common approach is to integrate domain knowledge to increase the model's adaptability to a particular field. Ontology, a structured way to represent terminologies with relations and symptoms, can capture biomedical knowledge in a formal and straightforward manner [110]. A prominent ontology in the medical domain is Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [31], which provides structured relationships for more than 300,000 medical concepts. The Human Phenotype Ontology (HPO) [65] contains roughly 12,000 terms describing phenotypes of human genetic diseases. Experimental Factor Ontology (EFO) [83] provides a systematic description of exper-

imental variables, covering aspects of disease, anatomy, cell type, cell lines, chemical compounds and assay information. As a single ontology may contain limited terms and miss valid synonyms, we propose blending concepts from multiple ontologies to create a more comprehensive knowledge graph.

4.2 Dataset Annotation

As previously discussed in section 3.2, we have curated the CLINIC-ISBAR dataset and annotated the sentence-level intent based on the ISBAR communication protocol. For the purpose of this study, we concentrated on the medical cases within the CLINIC-ISBAR dataset and manually annotated the required information based on the components list provided by clinical experts.

Initially, a document analogous to Table 4.1 was provided by the clinical expert. This document contained 22 required components along with their corresponding sample words. Subsequently, each sentence in the clinical handover was meticulously examined to ascertain the presence of a specific component. If a component was identified within a sentence, that particular sentence-component pair was labeled as 'True'; otherwise, it was labeled as 'False'.

Through this rigorous process, we have annotated a total of 21625 sentence-component pairs. The distribution of these components is visually illustrated in Figure 4.2. The figure reveals a considerable variation in the frequency of different components. Furthermore, when juxtaposed with the overall size of the dataset, the frequency of each component appears relatively small. This observation underscores the extreme imbalance within the dataset, characterized by a predominance of 'False' labels and a scarcity of 'True' labels. This imbalance presents a unique challenge that will need to be care-

fully considered in subsequent analyses and modeling stages.

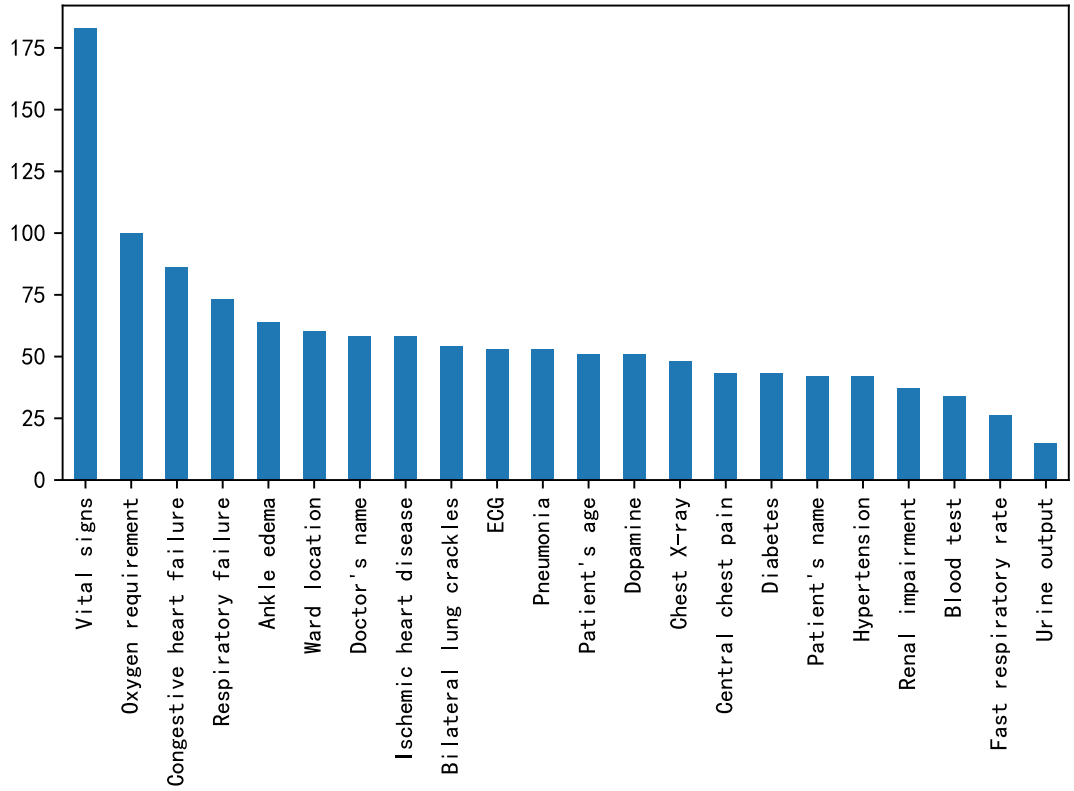

Figure 4.2: Distribution of components.

4.3 Method

In this section, we present how to construct a local knowledge graph based on the given examples and external biomedical ontology. Then we introduce our proposed Knowledge-Infused Prompt Learning (KIPL) which incorporates external biomedical knowledge into a prompt templates.

4.3.1 Knowledge Graph Construction

Our methodology employs an automatically constructed knowledge graph, formulated in the conventional triple format, denoted as $KG =< S, P, O >$. Here, S and O denote entities, and P represents relations between entities. These variables—S, P, and O—are aligned with the conventional notion of a knowledge graph, serving as subject, predicate, and object, respectively. Hence, the knowledge graph is defined as $G = (s, p, o) \mid s \in S, p \in P, o \in O$.

Figure 4.3 depicts the procedure of constructing the knowledge graph. Given a document containing required information, we consider its components and sample words as entities. The relation between these entities is encapsulated by the term "*includes*." Thus, during the initialization of the knowledge graph, the triple "*component, includes, sample words*" is inserted into graph G. As indicated in Figure 4.3, we categorize entities in sample words into two types: the green nodes represent objects or concepts, while the blue nodes represent literals with numerical values. We associate these literals with the relevant component by introducing a new triple, "*component, has_value, literal*," to the graph.

Following the initialization phase, the knowledge graph is expanded using three biomedical ontologies – HPO [65], SNOMED-CT [31], and EFO [83]. These ontologies offer a structured representation of biomedical terminologies, enabling us to locate symptoms, hypernyms, and hyponyms associated with a given clinical term. Because hypernyms or hyponyms may bring about changes in the scope of components, only synonym expansion is considered here. As illustrated in Figure 4.3, we search for the synonyms of the green nodes (sample words) and insert the associated triple (*"sample words, has_synonym, synonym*") into graph G. To ensure the reliability of the syn-

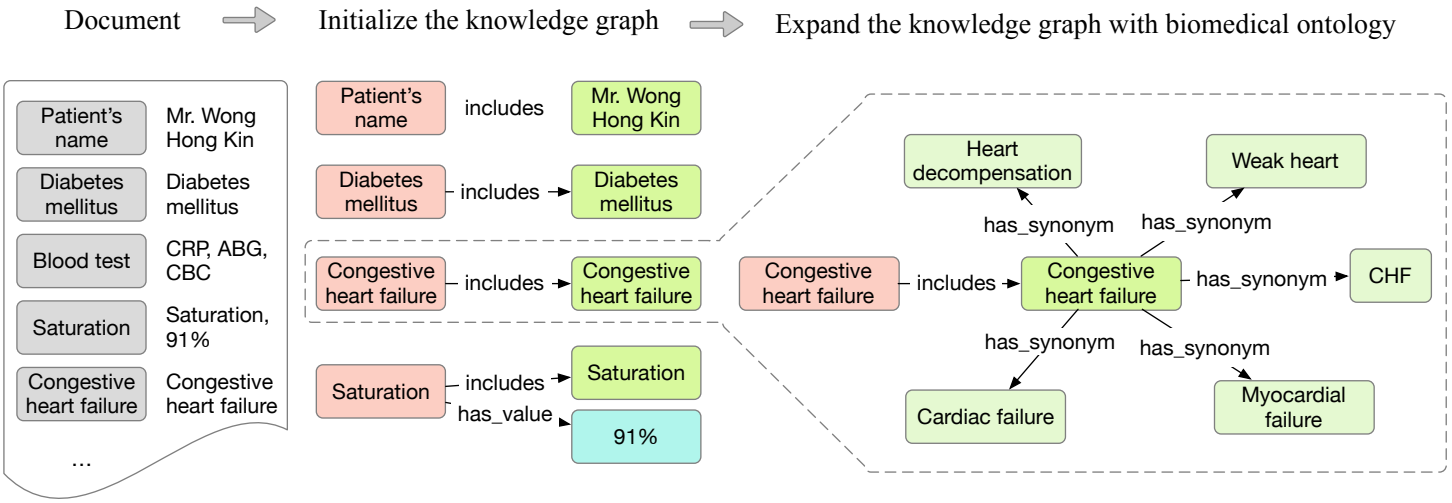

Figure 4.3: Procedure of constructing the knowledge graph. First, a knowledge graph is initialized based on the provided documents. It is then expanded using biomedical ontology.

onyms extracted from these external resources, we only include those that appear in two or more ontologies. During this expansion, if a synonym is found that is already in the graph, the graph is updated such that this sample word inherits all the outgoing relationships from that synonym. Subsequently, the synonym and its associated edges are removed from the graph. During the graph refinement process, we identify and remove any self-referencing edges. Upon the completion of the knowledge graph construction, the graph features a maximum of three types of edges: "*includes*", "*has_value*", and "*has_synonym*".

4.3.2 Knowledge-infused Prompt Learning

In our sentence-component pair classification problem, an input contains a sentence sequence $\mathbf{x} = (x_0, x_1, \ldots, x_n)$ and a component sequence $\mathbf{z} = (z_0, z_1, \ldots, z_m)$. The task

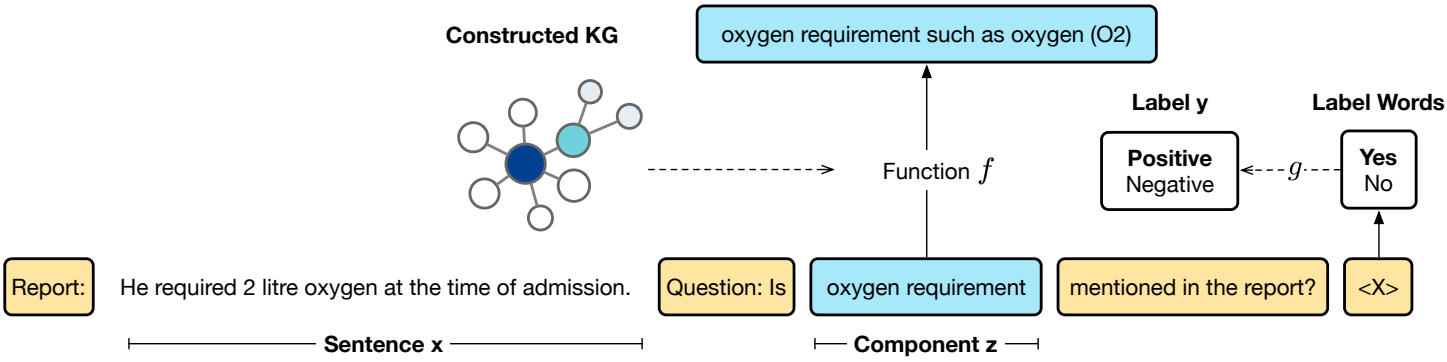

Figure 4.4: Knowledge-infused prompt learning for sentence-component pair classification. The yellow rectangles in the figure denote prompt tokens, where $\langle \mathrm{X} \rangle$ signifies the masking token. The yellow rectangles are components which can be expanded using the knowledge graph.

is to assign this pair to a specific class label $y \in \mathcal{Y}$. In prompt learning, we need to formulate the classification problem as a language modeling problem. As shown in figure 4.4, a prompt template is used to encapsulate the input pair into a coherent piece of natural language text that retains the semantic essence of the original input pair while making it interpretable for the language model.

In this example, we need to ascertain whether the component **z** "oxygen requirement" is referred to in the sentence **x** ("He required 2 litre oxygen at the time of admission"). And the prompt is initially formulated as:

$$\mathbf{p} = \text{Report: } \mathbf{x} \text{ Question: Is } \mathbf{z} \text{ mentioned in the report?} \langle \mathrm{X} \rangle \tag{4.1}$$

This formulation allows the language model to understand the task in natural language terms and provides a focal point ("oxygen requirement") for it to center its attention on. This mechanism can accommodate a variety of sentence-component pairs while providing the context of classification task.

Following this, we aim to infuse biomedical knowledge into the prompt. This is where we introduce function f which is tasked with the expansion of the component **z** by exploring entities and relationships in the knowledge graph G. Function f operates by employing a breadth-first algorithm to search the graph G for each component, transforming the obtained triples into natural language descriptions. Specifically, "*component, has_value, literal*" would be transformed to "equals to literal", "*component, includes, sample words*" becomes "such as sample words", and *"sample words, has_synonym, synonym*" would be expressed as "(synonym)". In situations where an entity possesses multiple outward relations of the same type, we simply concatenate the words it points to with commas. To illustrate, consider the following examples, represented in Table 4.2:

Component	**Initial Prompt**	**Knowledge-infused Prompt**
Hypertension	… Is hypertension mentioned in the report? ⟨X⟩	… Is hypertension (htn, high blood pressure, increased blood pressure, hyperpiesia) mentioned in the report? ⟨X⟩
Saturation	… Is saturation mentioned in the report? ⟨X⟩	… Is saturation equals to 91% mentioned in the report? ⟨X⟩
Blood test	… Is blood test mentioned in the report? ⟨X⟩	… Is blood test such as ABG (blood gas analysis, BGA, arterial blood gas), CBC (blood cell count, full blood count), CRP (C-Reactive Protein) mentioned in the report? ⟨X⟩

Table 4.2: Examples of knowledge-infused prompts.

As shown, the function f expands the initial prompts with useful biomedical information drawn from the knowledge graph, thereby producing the knowledge-infused prompts.Then the pre-trained language model, denoted as $\mathcal{M}$, processes the input $f(\mathbf{p}, G)$ and outputs the probability of each word v in the ⟨X⟩ token:

$$P_{\mathcal{M}}\left(\langle \mathrm{X} \rangle = v \mid f(\mathbf{p}, G)\right) \tag{4.2}$$

The transformation of these probabilities into labels requires a mapping function, g, which connects the label word set $\mathcal{V}$ to the label space $\mathcal{Y}$. In our scenario, we work with two label sets: $\mathcal{V}_1 = \{$ "yes" $\}$ and $\mathcal{V}_2 = \{$ "no" $\}$. The function g maps $\mathcal{V}_1$ to the label "positive" and $\mathcal{V}_2$ to the label "negative". Thus, the probability of label y is calculated using the equation:

$$P\left(y \mid \mathbf{x}, \mathbf{z}\right) = g\left(P_{\mathcal{M}}\left(\langle \mathrm{X} \rangle = v \mid f(\mathbf{p}, G)\right) \mid v \in \mathcal{V}_y\right) \tag{4.3}$$

This mapping function g plays an essential role in bridging the model's output space, expressed in words ("yes" or "no"), with the label space of the classification problem ("positive" or "negative"). This decoupling between the model's output and the label space allows for flexibility in dealing with more complex problems. For instance, it accommodates scenarios where a single label in the classification problem might be associated with multiple words or phrases in the model's output.

Furthermore, it's worth emphasizing that by excluding $\langle \mathrm{X} \rangle$ from the prompt, the problem can still be transformed into a conventional sentence classification task. The methodology for infusing knowledge remains unchanged. The difference lies in the calculation of probabilities. Rather than determining the word distribution probability at the $\langle \mathrm{X} \rangle$ position, we would classify based on the encoding of the whole sentence.

4.4 Experiments

In this section, we detail the experimental setup, compare the performance of various methods for content recognition under different constraints, and present case studies for specific components.

4.4.1 Experiment Settings

The experiments were conducted on the annotated data from the medical case subset of the CLINIC-ISBAR dataset, as described in Section 4.2. The data was randomly partitioned into training, testing, and validation sets in a ratio of 3:1:1. For the pre-trained language models, we employed BERT-base for vanilla fine-tuning and T5-base for prompt-oriented fine-tuning. Knowledge-infused prompts were implemented in both scenarios. The selection of most hyperparameters was guided by previous works. Both BERT-base and T5-base were optimized using the Adam optimizer. The learning rate was set to $1e-5$ for BERT and $1e-4$ for T5. Models were trained for 5 epochs with a batch size of 4, and the best checkpoint was selected based on validation performance.

To ascertain the necessity of large pretrained language models for this task, we also compared the performance with deep learning models and traditional machine learning models. The deep learning models used in these experiments were configured similarly to the baselines described in Section 3.4.1. A wide range of machine learning techniques, namely Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), is chosen to provide a diverse array of algorithmic principles for comparison in the context of our text classification task. An integral preprocessing step involved the transformation of text data into a numeric format via the Term Frequency-Inverse Document Frequency (TF-IDF) approach, with a cap placed

on the maximum feature count at 1200, ensuring a manageable and computationally efficient feature space.

Each machine learning method was selected and configured with due consideration to the specific requirements and characteristics of our task. Logistic Regression, a simple and efficient method for binary classification, was applied with an L2 regularization penalty and a regularization strength parameter of 1.0, offering a balance between complexity and generalization. We utilized the Naive Bayes technique, acknowledging its effectiveness in high-dimensional problems, and operated under the Gaussian likelihood assumption, in line with the continuous nature of TF-IDF-transformed data. The KNN model was configured to use the 9 closest neighbors in the training data for prediction. For SVM, recognized for its flexibility and proficiency with high-dimensional data, we chose the Radial Basis Function (RBF) kernel to capture potential nonlinear relationships in the data. A regularization parameter of 1.0 and a gamma parameter set to 'scale' were adopted.

Since the proportion of positive and negative labels is extremely unbalanced, *Precision* and *Recall* and *F1_Score* are reported as performance measures, which are calculated as follows:

$$\begin{aligned} Precision &= \frac{\text{\# of correctly detected components}}{\text{\# of detected components}}, \\ Recall &= \frac{\text{\# of correctly detected components}}{\text{\# of existing components}}, \\ F1\_Score &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \end{aligned} \tag{4.4}$$

4.4.2 Results and Discussion

The results of content recognition for a variety of algorithms are summarized in Table 4.3.

Method	Original Prompt			Knowledge-Infused Prompt		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
BERT	0.805	0.963	0.877	0.823	0.983	0.896
T5	0.918	0.984	0.950	0.960	0.983	0.971
LSTM	0.284	0.685	0.398	0.272	0.643	0.382
BiLSTM	0.276	0.728	0.400	0.304	0.692	0.421
AttLSTM	0.275	0.681	0.388	0.299	0.696	0.413
RCNN	0.402	0.740	0.520	0.420	0.715	0.527
TextCNN	0.170	0.473	0.250	0.185	0.478	0.266
Transformer	0.151	0.426	0.221	0.151	0.461	0.226
Logistic Regression	0.136	0.362	0.198	0.145	0.388	0.211
Naive Bayes	0.065	0.833	0.121	0.065	0.829	0.120
KNN	0.131	0.742	0.222	0.136	0.865	0.235
SVM	0.153	0.433	0.226	0.261	0.554	0.355

Table 4.3: Performance of different methods on content recognition.

The results demonstrate a substantial disparity in performance between traditional machine learning (ML)/deep learning (DL) methodologies and expansive pre-trained models, thereby underscoring the necessity of employing pre-trained language models for our task of content recognition. In most models, the use of knowledge-infused prompts leads to improved performance. This improvement is particularly pronounced when integrated with pre-trained language models, despite their already superior performance with the original prompt. This demonstrates the profound capacity of knowledge-infused prompts to effectively employ pre-trained models and assimilate external knowledge, consequently achieving superior F1-scores.

In contrast, for DL and ML models, the enhancements associated with the knowledge-

Training Data	Original Prompt			Knowledge-Infused Prompt		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1/2	0.757	0.978	0.854	0.790	0.992	0.880
1/4	0.760	0.963	0.849	0.792	0.971	0.872
1/8	0.678	0.956	0.793	0.689	0.960	0.802
1/16	0.626	0.952	0.755	0.679	0.944	0.790

Table 4.4: F1-score of T5 when using different size of the training data.

infused prompts were less significant and in certain cases, even resulted in a performance decline. This phenomenon can be attributed to the fact that these models, being trained from scratch, do not possess the requisite prior knowledge to effectively exploit the benefits of the infused knowledge. Furthermore, the knowledge-infused prompts, which integrate more semantically-rich data, may inadvertently induce additional ambiguity into these models. Thus, the knowledge-infused prompts are ostensibly more advantageous for models with greater learning capabilities.

The subsequent analysis focused on assessing the performance of the models upon reducing the size of the training data. Table 4.4 delineates the F1-score of the T5 model when employing different fractions of the training data.

From the table, it is evident that the model, when employing knowledge-infused prompts, consistently outperforms the original prompt, irrespective of the size of the training data, thus manifesting the efficacy of knowledge-infused prompts. Incorporating external knowledge imbues the model with robustness, allowing it to sustain relatively higher precision, recall, and F1-scores across varying sizes of training data. This is particularly apparent when only a minuscule portion of the training data is utilized (1/16), where the application of knowledge-infused prompt tuning sustains a relatively high score.

These results, therefore, underscore the significance of integrating knowledge-infused

prompts with pre-trained models in our content-recognition task, especially in scenarios where the availability of training data is limited. However, these empirical findings may not entirely represent the efficacy of knowledge-infused prompts, considering that not all components pertain to the clinical domain, and some clinical concepts might comprise limited extended words. To elucidate the potential benefits of knowledge-infused prompts, an in-depth analysis of the individual component recognition performance was conducted.

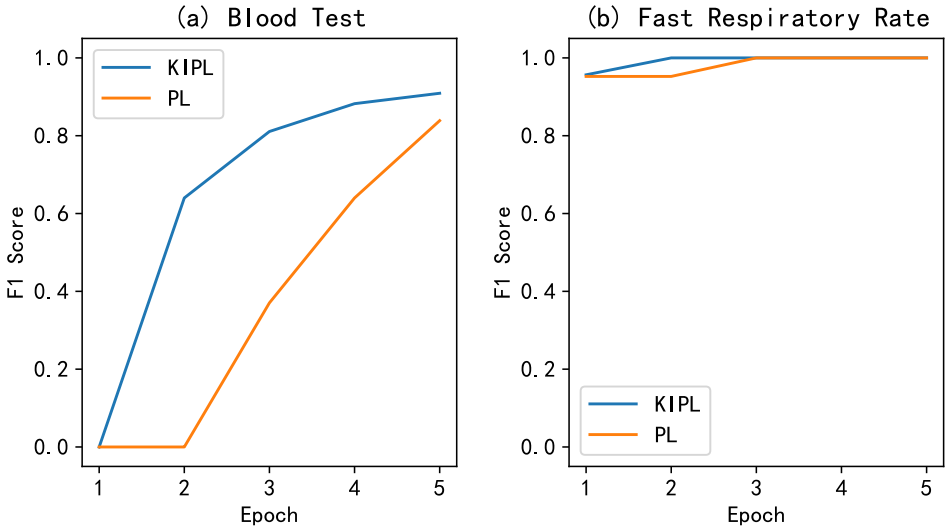

Figure 4.5: Learning curves of Prompt Learning (PL) and Knowledge-Infused Prompt Learning (KIPL) in individual components.

The T5 model was fine-tuned for individual component recognition, and Figure 4.5 illustrates the learning curves of Prompt Learning (PL) and Knowledge-infused Prompt Learning (KIPL) in the components ”Blood Test” and ”Fast Respiratory Rate”. The concept of ”Fast Respiratory Rate” is unambiguous and comprises limited expression variations. Consequently, both the PL and KIPL models converge rapidly and attain high f1 scores. In contrast, the concept of ”Blood Test” is more convoluted, encapsulating

both nebulous terms such as blood culture and specific tests like ABG, thereby imposing greater challenges on the model. The figures suggest that the integration of external knowledge in KIPL fosters quicker pattern identification.

Figure 4.6 illustrates the learning curves for individual components when a smaller subset of the training data is used. It is clear from the figure that in the case of the "Blood Test" component, KIPL consistently outperforms PL, and its learning curve demonstrates superior stability. On the other hand, for the "Fast Respiratory Rate" component, where the performance discrepancy between the two models was negligible in the full data scenario, a pronounced divergence emerges when the training data set is reduced. With the incorporation of external knowledge, KIPL not only exhibits a faster learning rate but also achieves a higher f1-score.

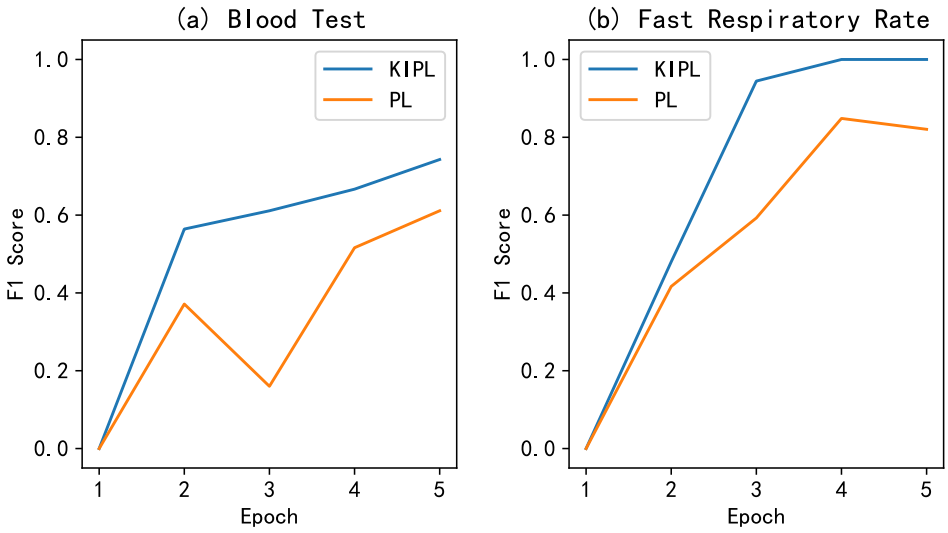

Figure 4.6: Learning curves of Prompt Learning (PL) and Knowledge-Infused Prompt Learning (KIPL) in individual components. 40% of the dataset is used for training.

These examples underline the distinctive advantage of knowledge-infused prompt in scenarios involving complex clinical terms and limited data availability. These find-

ings align with our expectations, as external knowledge is inherently valuable in such contexts.

4.5 Conclusion

In this chapter, we introduced the knowledge-infused prompt as a novel approach to content recognition in the domain of clinical communication. We initiated the process by recasting the content recognition task as a language modeling problem. Utilizing the biomedical ontology and the given document, we proceeded to construct a knowledge graph. Subsequently, the biomedical knowledge was integrated into prompts through a careful exploration of the entities and relationships within the knowledge graph.

The experimental results drawn from clinical handover data analysis provide compelling evidence of the efficacy of our approach. Particularly when it comes to deciphering complex clinical terms and working with limited training data, integrating knowledge-infused prompts with pre-trained language models significantly enhances the performance of content detection. These promising results lay the groundwork for future research to further refine and optimize the use of knowledge-infused prompts across various contexts.

Chapter 5

Customizable Conversational System with Insufficient Data

Creating a healthcare domain-specific conversational training system is a significant undertaking, typically requiring collaboration among domain experts, extensive data annotation, and a prolonged development cycle [96]. As with many AI applications, such systems are often confined to specific task scenarios, and struggle to adapt quickly to new clinical situations and training requirements. For instance, the global outbreak of COVID-19 is affecting healthcare systems around the world: Not only are we confronted with complex and changing clinical communication environments, but also with the emergence of new communication standards. Our medical staff and volunteers are in desperate need of training at the moment. However, this is also the time when manpower and time are most scarce. In these pressing times, it's impractical to engage in lengthy system development involving multiple stakeholders and extensive data collection. Notably, the most current and relevant information often resides with frontline medical professionals. In response to these challenges, we propose a customizable method that

empowers non-expert users to quickly develop AI applications for new tasks.

Several hurdles impede non-expert users from creating new AI models for fresh training tasks in clinical communication. Firstly, they often lack the capability to design and debug models, which restricts the type of models available for the task. Secondly, non-experts may find it challenging to amass, label, and process vast data quantities. This is significant because many contemporary AI algorithms are data-driven; limited data can result in model overfitting and decreased generalization performance. Finally, specific tasks may require specific knowledge that, while readily expressible by humans, cannot be directly instilled into AI models.

Regarding the intent detection task of the conversational system for clinical communication training, it employs similar paradigms and data processing procedures. The advent of large-scale pre-trained models allows serving different downstream tasks by fine-tuning parameters without altering the structure [29]. However, for data-driven methods, the risk of introducing significant bias increases in scenarios with insufficient data. Therefore, we shift the focus of the problem to the construction of an appropriate dataset, from which the model can efficiently learn more effective features.

With this aim, we propose Data Augmentation with User-Defined Knowledge (UDK-DA), a technique that enables non-information technology (IT) professionals to design AI tasks using limited samples and user-defined knowledge. In our intent detection task, infuses user-defined lexical and contextual knowledge into training samples via data augmentation, thereby increasing the robustness and generalizability of machine learning models. Experiments show that UDK-DA can significantly boost performance on intent detection and content recognition, demonstrating the possibility of designing NLP tasks by non-professionals.

5.1 Related work

A typical intent detection pipeline consists of text pre-processing, tokenization, vectorization, and classification. While pre-processing and tokenization can be implemented with a set of rules, vectorization and classification are based on the statistics of the current dataset (e.g. TF-IDF) or learned from an external corpus (pretrained models). Taking advantage of the parallel training ability of Transformers, large pretrained models are emerging. A prime example is Bidirectional Encoder Representations from Transformers (BERT) [29], which has become a standard building block for training NLP models in many tasks. BERT is pretrained on a large corpus and can be used for a variety of NLP tasks by fine-tuning on a given task without modifying the network structure. Using the same idea, pretrained language models in biomedical domain have been developed, such as BioBERT [69]. BioBERT is a BERT-based model pretrained on large-scale biomedical corpora and has been found to outperform earlier models on an array of biomedical text mining tasks [69].

Inadequate or even unavailable training data from emerging classes present a major obstacle to text classification tasks [36]. In low data settings, it is often necessary to increase the size of training data to reduce overfitting and improve the robustness of machine learning models [67]. Data augmentation (DA) is a frequently used technique for increasing the size of training data without explicitly collecting new data [118]. It also provides a simple way to inject prior knowledge into a deep learning system and to improve models' generalizability.

The current DA techniques in NLP can be divided into three categories: rule-based, example interpolation-based and model-based methods. One of the most popular rule-based methods is Easy Data Augmentation (EDA) proposed by Wei and Zou [124]. It

contains a set of token-level random perturbation operations (including random insertion, deletion, and swap) and can improve performance on many text classification tasks. Another class of DA technique interpolates the input examples and labels of two or more real examples, which is also sometimes referred to as mixed sample data augmentation. pioneered by MIXUP [137], The model-based techniques use seq2seq and language models for DA. For instance, the popular backtranslation method [112] entails translating the source language into other languages and then back into the source language to generate a diverse expression. Some researches also investigate domain knowledge to assist DA in a professional field. Kang et al. [62] proposed UMLS-EDA for biomedical named entity recognition, which extends the EDA method by Unified Medical Language System (UMLS) [62] knowledge. The authors incorporated the UMLS knowledge by identifying UMLS concepts and replacing them with synonyms.

However, the majority of DA methods in NLP work with a reasonable amount of data, ranging from a few dozens to tens of thousands of samples per class. In our case, the model may need to work with an extremely limited number of user-provided examples, possibly as few as single-digit samples per category. Additionally, the external knowledge employed in DA methods is either broadly targeted (e.g., synonymies in WordNet) or utilized implicitly (e.g., the translation model in backtranslation [112]), where users cannot modify or monitor the knowledge used.

5.2 Methods

5.2.1 IA-BioBERT

Building upon the successful application of the intent-aware mechanism for intent detection in standardized clinical communication as detailed in Chapter 3, we extend this approach to BioBERT, a pretrained language model tailored to the biomedical domain. This extended model, termed IA-BioBERT, leverages domain knowledge and contextual information, thereby enhancing the model's robustness and performance in the biomedical context.

As discusses in Section 3.4.3.2, applying the intent-aware mechanism to large pretrained models only yielded limited performance improvements. This is primarily due to the substantial dimensional discrepancy between the sentence vector and the intent vector. To mitigate this issue, we compress the sentence representation in IA-BioBERT using a fully connected network. The structure of the IA-BioBERT is illustrated in Figure 5.1.

The i^{th} input sentence is initially tokenized into a series of tokens utilizing the BioBERT Tokenizer. We then append the [CLS] token at the beginning of the token sequence and pad the sequence to achieve a specified length. The tokens are subsequently converted into numeric representations based on BioBERT's predefined vocabulary. The BioBERT model then processes the padded sequence, resulting in a vector representation for each token.

To condense this information, we pass the vector representation of the [CLS] token through a fully connected layer (FC1), yielding a dense sentence representation $\boldsymbol{s}i$. This representation is concatenated with the previous sentence's prediction result $\boldsymbol{p}(i-1)$ and then passed through a second fully connected layer (FC2) and a SoftMax layer.

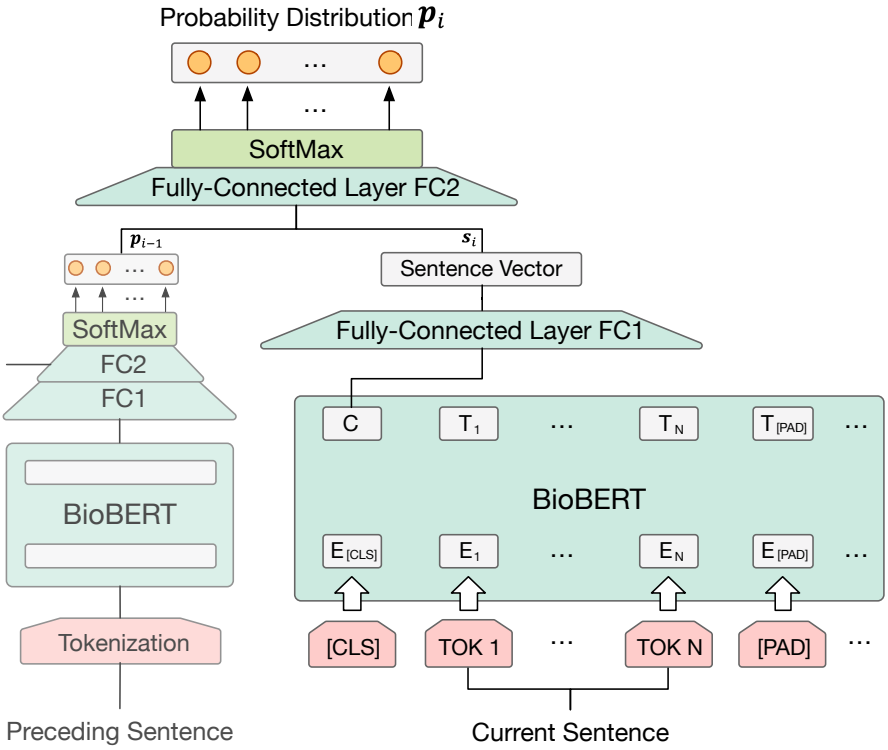

Figure 5.1: Architecture of IA-BioBERT model for intent detection.

This produces the prediction result $\boldsymbol{p}_i$, which is described by:

$$\boldsymbol{p}_i = \text{SoftMax}\left(\left[\boldsymbol{p}_{i-1}\boldsymbol{s}_i\right] W^T + \boldsymbol{b}\right), \tag{5.1}$$

where vector W and $\boldsymbol{b}$ are the parameters of the fully-connected layer FC2 and $[\cdot\cdot]$ is the action of concatenation. The predicted intent is then obtained by selecting the maximum value in the prediction result vector: $y_i = \text{argmax}\left(\boldsymbol{p}_i\right)$.

By employing this strategy, IA-BioBERT encapsulates the strengths of the intent-aware mechanism and the biomedical specificity of BioBERT.

5.2.2 Data Augmentation with User-Defined Knowledge (UDK-DA)

DA techniques are commonly employed to increase the quantity and variety of samples, thereby enhancing the generalizability of learning models. Although common strategies such as Easy Data Augmentation (EDA)[124] and backtranslation[112] have been successful in NLP tasks, they may not fully capture the characteristics of clinical communication data. Thus, we introduce User-Defined Knowledge Data Augmentation (UDK-DA), a DA technique specifically tailored to the standardized clinical communication data. UKD-DA augments the conversational data using user-defined knowledge from two aspects: term replacement improves data at the lexical level, and context expansion generates more diverse combinations at the sentence level.

- **Term replacement** is similar to synonym replacement, but rather than substituting random words, it replaces specific terms with their alternative forms or synonyms defined by users. To streamline the process and reduce manual work, a clinical named entity recognition (NER) model is initially employed to extract clinical entities. In our implementation, the Spark NLP for Healthcare's NER model is utilized. Following extraction, biomedical ontologies, including HPO [65], SNOMED-CT [31], and EFO [83], are used to generate a synonym set for each term. Users are then given these synonym sets as reference and define their own alternative sets. During the augmentation phase, terms in these refined lists are replaced with user-defined alternatives, yielding more diverse linguistic constructions.

- **Context expansion** targets augmentation at the sentence level. Owing to the interdependence of categories within a communication protocol, the sentence's context can substantially impact its classification. To that end, we employ user-defined

rules to facilitate context expansion. Given the sequential nature of conversations, we designate the preceding sentence as the context for the current one. Table 5.1 demonstrates the variability of sentence intent depending on its context; the dash ("-") indicates that the category cannot be determined and that we need to consider specific content.

No Context	Context I	Context S	Context B	Context A	Context R
I	I	I	I	I	I
S	S	S	-	-	-
B	B	B	B	-	-
A	S	S	-	A	-
R	-	-	R	R	R

Table 5.1: Contextual relationships of five intents in the ISBAR protocol.

The process of context expansion capitalizes on the contextual relationship table, augmenting sentences by adjusting their context and modifying their category according to the corresponding cell in the table if the value is not a dash ("-"). For representing the context, we select a random sentence from the conversation samples in the same category. Algorithm 1 explicates the detailed mechanics of the context expansion process.

5.2.3 Semantic Matching

In Chapter 4, we presented knowledge-infused prompt tuning with pre-trained language models for content recognition. This approach performs admirably with limited training data; however, it still necessitates human annotation. Under circumstances where labeled data is not readily available, the challenge of detecting components can be reframed as a task of identifying corresponding sample words. The rationale behind this is

Algorithm 1: Context expansion

Input: conversation dataset D, contextual relationship table T, label list L
Output: augmented dataset AD

```
AD ← D;
for (sentence, label, context, context label) in D do
    for l ∈ L do
        if context label ≠ l then
            new label ← T(label, l);
            if new label ∈ L then
                new context ← a random sentence with label l in D ;
                new context label ← l;
                add (sentence, new label, new context, new context label) to
                  AD;
            end
        end
    end
end
```

that the mention of a sample word implies the mention of the corresponding component.

Traditional information retrieval systems advocate the use of exact lexical matching, which decomposes the source text into n-gram words and subsequently assesses their presence [108]. However, this approach may be overly rigid due to its lack of higher semantic matching and sensitivity to morphological changes. Consequently, it might overlook mentioned components due to an incomplete word list. To overcome these limitations, we propose an N-gram based Semantic Matching approach, which leverages the pre-trained word embeddings to calculate the semantic similarity between phrases [40].

Let's denote a sample word k as a sequence of word embeddings $k = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_t, \ldots, \boldsymbol{\nu}_T)$, where T represents the word count in k and $\boldsymbol{\nu}_t \in \mathbb{R}^M$ stands for the M-dimensional word embedding of the $t-$th word. The semantic embedding for the sample word k can be calculated by averaging all the word embeddings:

$$\boldsymbol{e}_k = \frac{1}{T}\sum_{t=1}^{N}\boldsymbol{\nu}_t \tag{5.2}$$

For a given input sentence, corresponding 1-gram, 2-gram, 3-gram, and 4-gram sequences are derived. Consider the sentence "*He requires 10 ml dopamine.*"; its 4-gram sequences would be ("*He requires 10 ml*", "*requires 10 ml dopamine*", "*10 ml dopamine.*"). A set W of n-gram words is formed by collating all words across these sequences. The semantic embedding $\boldsymbol{e}_w$ of the n-gram word w can be acquired by applying the word embedding strategy and Equation 5.2. The cosine similarity of k and w's semantic embeddings represents the similarity between k and w:

$$s(k,w) = \cos\left(\boldsymbol{e}_k, \boldsymbol{e}_w\right) = \frac{\boldsymbol{e}_k \boldsymbol{e}_w}{\|\boldsymbol{e}_k\| \, \|\boldsymbol{e}_w\|} \tag{5.3}$$

There might be instances when semantic similarity cannot be computed, such as when a word's semantic embedding is absent. Under these circumstances, lexical matching is employed as a substitute; if the words are identical, the score is set to 1, else it defaults to 0.

The sample word exhibiting the highest similarity to the n-gram word w can be selected by comparing their respective similarity scores. If the similarity surpasses a predefined threshold, w is classified as belonging to the component represented by the sample word. Upon processing all words within the n-gram set, the semantic matching approach yields a list of detected components in the input sentence. The specifics of this process are delineated in Algorithm 2.

Algorithm 2: Semantic matching

Input: component list C, input sentence S, similarity threshold ϑ
Output: detected component list DC

```
for n ← 1 to 4 do
    for n-gram word in S do
        max similarity ← 0;
        for component ∈ C do
            for sample word ∈ component do
                similarity ← s(sample word, n-gram word); // Equation 5.3
                if similarity > max similarity then
                    max similarity ← similarity;
                    detected component ← component;
                end
            end
        end
        if max similarity > ϑ then
            if detected component ∉ DC then
                add detected component to DC;
            end
        end
    end
end
```

5.3 Experiments

The efficacy of our proposed methodology is evaluated on standard clinical handovers taken from the CLINIC-ISBAR dataset 3.2. The dataset comprises 48 clinical handover samples totaling 980 sentences, sourced from the medical case. Figure 5.2 shows the sentence distribution and class distribution for the handover samples. It is noteworthy that these conversations vary significantly in length and the distribution of sentence categories is unbalanced, posing difficulties for subsequent intent detection tasks.

Our experiments are designed to emulate a scenario of limited sample availability for intent detection tasks. We utilize a single conversation sample from the clinical

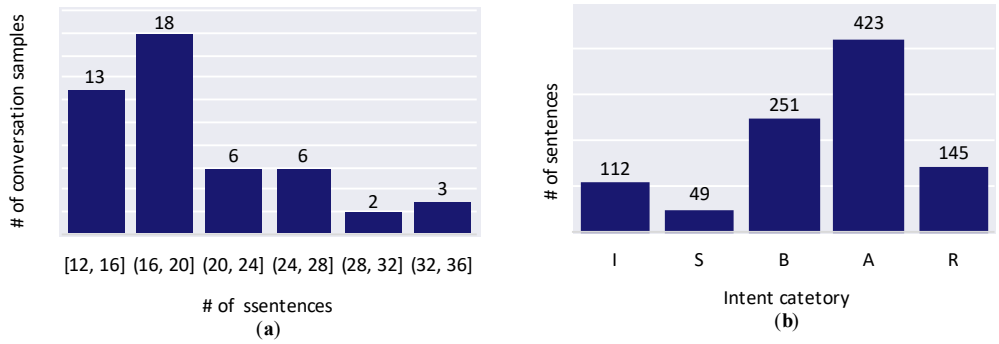

Figure 5.2: Sentence distribution in clinical handover. (a) Distribution of the number of sentences contained in conversation samples. (b) Sentence distribution by category.

expert as training data, while the remaining samples serve as performance evaluators. For the content recognition task, we employ the same document outlined in Section 4.2, which encompasses the required components. Figure 4.2 presents the distribution of components in the clinical handover.

Apart from executing experiments on the collected clinical handover, we also illustrate the establishment of a new model, specifically focusing on a new case of COVID-19.

5.3.1 Data augmentation

After acquiring data, the system augments conversation samples using EDA, backtranslation, term replacement and context expansion. We adopt the original EDA implementation[1], where the ratio of augmented sentences to original sentences is set to 9:1. Thus, the data volume amplifies tenfold post-EDA. Each operation, encompassing synonym replacement, random insertion, random swap, and random deletion, impacts 10% of the sentence words.

[1] https://github.com/jasonwei20/eda_nlp

For backtranslation, we leverage the English-Chinese and Chinese-English models of Google Cloud's Translation API[2], with each sentence undergoing a single round of backtranslation.

Regarding term replacement,each term is substituted with equivalent or synonymous terms, each substitution generating a distinct sentence. And context expansion of each sentence is executed as per Algorithm 1.

The augmented conversation samples are employed for fine-tuning the IA-BioBERT model for intent detection. The model is trained using the Adam optimizer with a maximum sequence length of 32 and a batch size of 16. We set the learning rate at 1e – 5 for original BioBERT model parameters and 1e-3 for fully-connected layer parameters. Owing to data scarcity, the dataset is not divided into training, validation, and test subsets; instead, all augmented data is used for training. The stopping criteria is defined as reaching either 50 training epochs or 100% training set accuracy.

In the context of content recognition tasks, we employ a different strategy. Instead of substituting terms within the conversation, we leverage term replacement techniques on the document comprising the components and their associated sample words (refer to Table 4.1). This methodology serves to expand our repository of sample words. We also assume a scenario wherein users are capable of identifying and marking all the components embedded within each sentence of the provided conversation. Consequently, we can construct a concise sentence-component pair dataset, akin to the one outlined in Section 4.2. Subsequent to the application of term replacement, this dataset is expanded. We then evaluate prompt tuning with the T5 model on this augmented dataset. The setting of the hyperparameters remains consistent with those mentioned in Section 4.4.1.

[2]https://cloud.google.com/translate

5.3.2 Intent detection models

In our study, we benchmark our proposed IA-BioBERT against various common neural networks utilized in NLP classification tasks, as outlined in section 3.4. Furthermore, we evaluate IA-BioBERT in relation to two of its variant models, aiming to substantiate the effect of the design of the model components:

- **BioBERT** [69] is the backbone model IA-BioBERT. It has been trained on biomedical corpora using the pre-trained BERT model. In this paper, all BioBERT models utilize BioBERT-Base 1.1[3].
- **IA-BERT** shares IA-BioBERT's structure, but replaces the pre-trained BioBERT-Base v1.1 with the BERT-base model.
- **IA-BioBERT without FC1** is another variant of IA-BioBERT; it circumvents FC1 by directly combining the representation of the [CLS] token, outputted by BioBERT, with the probability distribution of the preceding sentence's intent. This concatenated vector is then fed into the final fully-connected layer, FC2.

We adhere to the experimental settings outlined in section 3.4.2 during implementation and maintain the same hyperparameters.

5.3.3 Results on CLINIC-ISBAR

- **Intent Detection**

  In scenarios where training data sets are unbalanced, the tendency of a neural model to favor categories with abundant data samples may negatively impact the

[3] https://github.com/dmis-lab/biobert

performance of categories with fewer samples. To neutralize this bias, we employ the *Macro F1-score* as our performance metric.

Table 5.2 summarizes the Macro F1-score results for intent detection. As context expansion has no effects on models that do not take context as input, we limit its evaluation to context-based models. The four DA methods used in the task editor module enhanced the performance of all models. When context is not considered, BioBERT outperforms other models with augmented data due to its biomedical corpus pre-training. In addition, our proposed IA-BioBERT — which amalgamates the strengths of the medical corpus and contextual relationships — achieves leading performance under all DA methods. IA-BioBERT's performance benefits further from context expansion, yielding the highest overall Macro F1-score (0.791).

Model	**DA Methods**				
	w/o DA	**EDA**	**EDA+BT**	**EDA+BT+TR**	**EDA+BT+TR+CE**
LSTM [52]	0.475	0.527	0.553	0.566	-
BiLSTM [43]	0.423	0.484	0.505	0.520	-
AttLSTM [8]	**0.551**	0.569	0.611	0.602	-
RCNN [68]	0.512	0.566	0.584	0.604	-
TextCNN [139]	0.477	0.581	0.591	0.594	-
Transformer [122]	0.433	0.389	0.471	0.598	-
BERT [29]	0.368	0.600	0.612	0.670	-
BioBERT [69]	0.449	0.645	**0.693**	**0.701**	-
IA-BERT	0.470	0.615	0.652	0.682	0.769
IA-BioBERT w/o FC1	0.447	0.637	0.678	0.691	0.750
IA-BioBERT	0.473	**0.671**	0.682	0.690	**0.791**

Table 5.2: F1-scores of all intent detection models using different DA methods. BT: backtranslation; TR: term replacement; CE: context expansion.

We notice that context-based models might overly rely on context. This reliance becomes problematic especially when training samples are scant and lack diverse

context relations, resulting in overfitting and suboptimal test set performance. However, the augmentation of diverse context cases through context expansion significantly improves the performance of all context-based models.

- **Content Recognition**

  In the content recognition experiment, we set an empirical value of 0.9 as the threshold for semantic matching, and evaluated the performance using precision, recall, and F1 score as metrics. The results from various content recognition methods are displayed in Table 5.3.

Model	w/o DA			w/ DA		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Lexical Matching	0.991	0.621	0.763	0.970	0.690	0.806
Semantic Matching	0.970	0.673	0.795	0.857	0.786	0.815
T5 (Basic Prompt)	0.934	0.842	0.886	0.946	0.870	0.907
T5 (Knowledge-infused Prompt)	0.918	0.881	0.899	0.961	0.865	0.910

Table 5.3: Performance on content recognition.

The table reflects that the performance of both lexical and semantic matching outstrips that of DL and ML models (Tabel 5.3) even that they're trained on the full training dataset. This underscores the vital role of sample words in content recognition. Despite high precision, the traditional lexical matching method, reliant on manually provided words, suffers from low recall. Two critical limitations are inherent to this method. First, lexical matching is sensitive to slight modifications in morphology and order, resulting in potential matching failures. Second, the sample words utilized merely serve as component expression examples and fail to cover all possible scenarios. The incorporation of semantic word embedding ensures robustness in semantic matching when matching components, resulting in an enhanced F1 score, albeit with a minor compro-

mise in precision. By managing the threshold of semantic vector similarity, semantic matching can adjust the matching ”looseness” of terms. Furthermore, our observations affirm that the performance of both methods can be bolstered by augmenting sample words through term replacement.

In an alternate setting, we presume that users can pinpoint the component in each provided conversation sample. Accordingly, we can fabricate a sentence-component pair dataset analogous to the procedure described in section 4.2, using the provided clinical communication sample. Although a solitary dialog sample is available for training, the pre-trained T5 model, when tuned using designed prompts, exhibits high performance, which is further amplified with a knowledge-infused prompt. The table also demonstrates that the optimal F1-score is procured by integrating a knowledge-infused prompt with data augmentation (term replacement, in this instance). This reinforces the notion that these two methods to incorporate knowledge are not exclusive but rather work together.

5.3.4 Demonstration of a COVID-19 Case

In addition to the collected dataset, we present a case study demonstrating the methodology employed to construct the model using a real-world example provided by the clinician. Table 5.4 demonstrates a clinical handover instance involving a COVID-19 patient potentially requiring intensive care unit (ICU) management due to respiratory failure. This case exhibits a dialogue between the resident doctor and the ICU doctor, annotated with the ISBAR framework to label the resident doctor’s inputs.

In our training framework, the role of the dialogue simulator mirrors the role of the ICU doctor, primarily as the recipient of the message. Therefore, our focus is mainly on

Speaker	Content	Intent
Resident doctor	This is Dr. Jeffery Lee, resident of AED. Are you ICU Dr. Ng?	I
ICU doctor	*Yes, I am.*	
Resident doctor	We have a patient in the resuscitation room right now, Ms. Yuen Mei Ho, a 65-year-old housewife.	I
Resident doctor	I am calling to ask if you can provide ICU care for Ms. Yuen.	S
Resident doctor	She has respiratory failure secondary to COVID infection and is now confused.	S
ICU doctor	*Would you please tell me more?*	
Resident doctor	She came to the AED at 9:30 this morning because she has fever and she finds trouble with her breathing.	B
Resident doctor	She deteriorated rapidly in the AED because of the acute respiratory failure secondary to COVID-19, and we found her saturation was about 85% in room air.	B
	...	
Resident doctor	We brought her to our resuscitation room at 1 pm as her saturation had further dropped to 85%. Now she is confused, E3M6V4.	A
Resident doctor	Temperature 38.6. SpO2 89% on 15L non-rebreathing mask, BP normal but tachycardia with heart rate about 140 beat per minute.	A
Resident doctor	the pH of her arterial blood gas is 7.20, pCO2 is 5.0, Po2 8.0, Creatinine is 230 umol/l, K is 4.5 mmol/l, WBC 12.1.	A
Resident doctor	Our impression is she has severe respiratory failure secondary to COVID infection and we will intubate the patient.	R
Resident doctor	May I ask if you would like to take this patient to ICU for further care?	R
ICU doctor	*Certainly. We will immediately get the bed ready and arrange transfer.*	

Table 5.4: A clinical handover case of the COVID-19 patient.

the communication style and the content of the dialogue conveyed by the resident doctor. The resident doctor's statements and their associated labels are initially augmented in accordance with the procedure outlined in Section 5.3.1. Subsequently, the IA-BioBERT model is trained on these augmented dialogues.

To augment the data, clinicians are required to provide an equivalent list and a context relationship table. Table 5.5 depicts an example of an equivalent list, where each row contains a set of synonyms or equivalents, separated by semicolons. Given that the COVID-19 case follows the ISBAR protocol as well, we utilize the same context relationship as shown in Table 5.1.

Ms. Yuen; Yuen Mei Ho
Resuscitation room; AED
ICU; Intensive care unit
Diabetes mellitus; DM; Diabetes
COVID-19, COVID infection
ADL; ADLs
DTS; Deep throat saliva
...

Table 5.5: Equivalent list of the COVID-19 case.

For content recognition, a list of essential components that should be addressed during the conversation is provided (Refer to Table 5.6). Each component contains one or more samppe words, akin to Table 4.1. These sample words are then augmented by term replacement.

After obtaining the intent detection model and sample words from the provided document, we demonstrate a use case of intent detection and content recognition. Table 5.7 illustrates the model's output when new dialogue is composed regarding the COVID-19 case.

As shown by the use case, even though only one dialog is provided as a training

No.	Component	Sample Words
1	Patient's name	Ms. Yuen Mei Ho
2	Location	Resuscitation room, AED
3	Age	65 years old
4	Admission time	This morning, 9:30
5	COVID	COVID-19, COVID infection
6	Diabetes mellitus	Diabetes mellitus
7	Vaccination	One dose in September 2021
…	…	…

Table 5.6: Required components and sample words in the COVID-19 case.

example, the domain knowledge and UDK-DA make our model robust to varying expressions. Despite the fact that the intent sequence in Table 5.7 differs from the sample dialog in Table 5.4, our model can correctly detect the intent. This can largely be attributed to the context expansion method, which generates dialogues with various sequence combinations.

Input sentence	Detected Intent	Recognized Components
This is Dr. Jeffery Lee, resident of AED. Are you ICU Dr. Ng?	I	Doctor's name; Location
I'm calling about Ms. Yuen, who may require ICU care due to respiratory failure caused by COVID infection.	S	Patient's name; Gender; COVID; Respiratory failure
Ms. Yuen Mei Ho is 65 years old and is currently in the resuscitation room.	I	Patient's name; Gender; Age; Location
She was admitted earlier today for a fever and breathing problems.	B	Gender; Fever

Table 5.7: A use case of intent detection and semantic matching on the COVID-19 case.

However, our model did not capture all of the components mentioned in the input

sentences. For instance, the phrase "*earlier today*" indicates "*Admission time*", which the model fails to recognize. Even though "*earlier today*" and our example word "*this morning"* are semantically similar, their word embedding similarity does not surpass the threshold, so it is not considered "*admission time"*. This illustrates that our semantic matching algorithm could still be improved.

5.4 Conclusion

Through this research, we have illustrated that even ordinary users can leverage pre-trained models and data augmentation to build robust AI models with only a handful of sample data. Our experiments within the realm of clinical communication reveal that the proposed IA-BioBERT model, built on the intent-aware mechnism and a pre-trained language model on biomedical corpus, surpasses all other baseline models in terms of intent detection.

Moreover, we introduced UDK-DA, a novel approach that incorporates user-defined knowledge into the conversational data. This is achieved via lexical and sentence-level augmentation methods, namely term replacement and context expansion. For content recognition tasks, we further integrated an n-gram semantic matching strategy. These methodologies noticeably enhance model performance when training data is sparse.

As we look towards the future, we aspire to broaden the range of application scenarios and delve into the feasibility of non-experts constructing AI models applicable to a diverse array of tasks.

Chapter 6

Heallo: Clinical Communication Training System

This chapter presents Heallo, a conversational system designed for clinical communication training, which supports user-defined tasks. Armed with biomedical ontologies, pretrained language models, and data augmentation techniques, Heallo is capable of simulating clinical communication scenarios, delivering timely assessments, and accommodating new tasks with simple editing.

More specifically, Heallo engages with trainees through simulated communication tasks, assuming the role of the receiver. At the conclusion of each session, the system reviews the entire communication history and generates a detailed evaluation report. This assessment adheres to a pre-defined grading rubric, ensuring fair and consistent feedback for all participants. Moreover, Heallo furnishes a platform for trainers to easily modify existing tasks or create new training scenarios. This feature provides great flexibility in designing personalized training programs.

The following sections will provide an in-depth introduction and analysis of Heallo.

First, we will explore the system's architecture, then examine its user interface in detail. Finally, we will discuss its practical application in hospital settings to demonstrate its functionality and relevance in real-world scenarios.

6.1 System Design

The system design of Heallo, as shown in Figure 6.1, is founded upon clinical communication tasks and the proposed structure for autonomous communication training, depicted in Figure 1.1. The design diagram in Figure 6.1 divides the three core modules - the dialogue simulator, evaluator, and task editor - into smaller units, providing a detailed overview of their interactions and data flow.

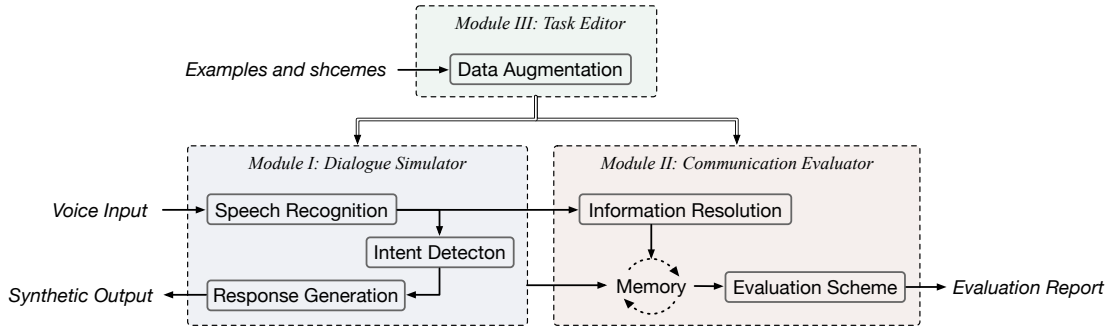

Figure 6.1: Framework of the conversational system for autonomous clinical communication training.

The dialogue simulator module begins by transcribing trainee's audio input into text through a speech recognition unit. This transcribed text is then processed by an intent detection unit, which informs the selection of an appropriate response. The selected response is subsequently converted back into speech and delivered to the trainee.

The communicaiton evaluator module focuses on the ongoing conversation, with the information resolution unit storing all dialogue details and intermediate results in

memory. When the conversation ends, the stored data is used to calculate a final score, following a pre-established evaluation scheme, and generate a detailed feedback report.

The task editor module allows trainers to design new training tasks. Trainers can provide communication examples, list required components, and define evaluation schemes. The system enhances the input examples and components through Data Augmentation (DA), and these enhanced inputs are then used to train models for subsequent tasks.

6.1.1 Dialogue Simulator

The dialogue simulator simulates clinical communication scenarios, engaging with trainees through conversation. It first transforms voice input into text via speech recognition. Given the ready accessibility and high performance of cloud-based automated speech recognition systems across diverse applications, we have opted to use Google Cloud's speech-to-text application programming interface (API)[1] for our speech recognition component. To improve recognition accuracy within the specific context of clinical scenarios, healthcare-related terms from our corpus are utilized as hot words.

Following speech recognition, the intent detector associates the transcribed sentence with an element from the communication protocol. For this purpose, we employ the IA-BioBERT model, which is outlined in section 5.2.1.

The conversational system primarily acts as an information receiver during the communication training. As such, the system's responses are designed to be relatively simple, primarily aimed at guiding trainees through the conversation. Expert clinicians, acting as trainers, initially establish a set of responses for a variety of conditions within the response pool. Subsequently, the response retriever chooses a fitting response from this pool, based on the predicted intent y_i. In instances where several responses could

[1] https://cloud.google.com/speech-to-text

be suitable, the retriever selects one randomly. This selected response is then converted into an audio clip using Google Cloud's text-to-speech API and played back to the user.

6.1.2 Communication Evaluator

Our evaluation of communication hinges on two aspects: the adherence of the conveyed content to the steps detailed in the protocol, and the inclusion of all predefined critical components within the conversation. In collaboration with clinical experts at Queen Elizabeth Hospital, we formulated three criteria:

1. *Category Number* (*CN*) irepresents the number of intent categories detected in the dialogue. Taking the ISBAR protocol as an example, CN equals four if a dialogue contains categories I, B, A, and R.

2. *Wrong Order* (*WO*) represents the number of incorrectly ordered intents. Similar to the concept of edit distance, WO is determined as the least number of intents that must be removed from a sequence in order to preserve its correct order. For instance, the WO of intent sequence I-S-B-A-B-R-A-R under the ISBAR protocol is 2.

3. *Missed Information* (*MI*) represents the number of required components that have been omitted in the dialogue.

CN and WO can each be calculated using the results of intent detection in *Module I: Dialogue Simulator*, while MI is calculated based on the results of semantic matching 5.2.3.

The trainer can define the evaluation scheme using the evaluation criteria CN, WO, and MI. Table 6.1 provides an example of a categorical evaluation scheme for clinical

handover. Based on this scheme, the system grades the communication performance and generates evaluation reports.

CN	WO	MI	Grade
5	0	≤5	A
≤4	≤1	≤9	B
≤4	≤3	≤13	C
≤3	≤5	≤17	D
	Others		F

Table 6.1: Categorical evaluation scheme for clinical handover following the ISBAR protocol.

6.1.3 Task Editor

The task editor is designed to enable non-IT professionals to generate new clinical communication tasks effortlessly by providing a handful of labeled examples. For task creation, users are asked to supply conversation samples with intent labels (as shown in Table 5.4), a list of components with sample words (as demonstrated in Table 4.1), and an evaluation scheme (as depicted in Table 6.1). DA techniques are then applied to enrich the number and diversity of samples, bolstering the generalizability of downstream models. Alongside well-established DA techniques (i.e., EDA and backtranslation), we also deploy the augmentation methods designed for standardized clinical communication.

6.2 Interface

The interactive interface of Heallo offers two user experiences – one for trainers and the other for trainees. The trainer's interface is illustrated in Figure 6.2, which enables clinical experts to create new communication tasks and monitor trainees' training results. It includes an administration board that enables trainers to manage the user list and track user progress (Figure 6.2a). This section also contains a task editing platform (Figures 6.2b–6.2d), where trainers can establish new communication tasks by providing examples and rules.

The interface depicted in Figure 6.2b enables the uploading of conversation examples and the establishment of context relationships. For a structured clinical handover based on the ISBAR methodology, conversation samples are inputs from the information provider with each sentence labelled in accordance with ISBAR. Taking the COVID-19 case in Table 5.4 as an example, we can divide the resident doctor's inputs into sentences and associate each with its corresponding ISBAR labels to obtain a conversation sample. An example of context relationships of ISBAR can be found in Table 5.1. Figure 6.2c shows the interface for adding critical components and configuring user-defined responses. A component table is similar to Table 4.1, where each component is assigned one or more keywords. Figure 6.2d displays the interface for defining a grading scheme by specifying formulas based on the CN, WO, and MI criteria.

After acquiring data, the system augments conversation samples using EDA, back-translation, term replacement and context expansion. For EDA, we follow the implementation of the original paper[2]. The number of generated augmented sentences per original sentence is set to 9, resulting in tenfold the amount of original data after EDA.

[2]https://github.com/jasonwei20/eda_nlp

Figure 6.2: Interface of Heallo for trainers. (a) Administration board for tracking each user's progress. (b) Task editing interface to enter examples and contextual rules. (c) Task editing interface to enter critical components and conditional answers. (d) Task editing interface to define grading schemes.

The percentage of words in the sentence that will be changed by each operation (i.e., synonym replacement, random insertion, random swap, and random deletion) is set to 10%. For backtranslation, we use the en-zh and zh-en models in Google Cloud's Translation API[3]; each sentence is backtranslated once. Regarding term replacement, each recognized clinical term is substituted with its equivalents or synonyms, resulting in a new sentence each time, resulting in a new sentence each time. For context expansion, we augment each sentence using Algorithm 1.

The augmented conversation samples are then used to fine-tune the IA-BioBERT model for intent detection. During training, we use the Adam optimizer and iterate over the dataset with a maximum sequence length of 32 and a batch size of 16. The learning rate is set to 1e – 5 for parameters in the original BioBERT model and to 1e-3 for parameters in the fully-connected layers. Due to data scarcity, we do not divide the dataset into training, validation, and test sets; rather, we use all augmented data for training. The stopping condition is defined as either the number of training epochs reaching 50 or the accuracy of the training set reaching 100%.

In terms of critical components, we augment sample words with term replacement. Then semantic matching is used to detect the components. In the semantic matching, we use glove.6B.50d [98] for word embedding.

The trainee interface is pictured in Figure 6.3, with Figure 6.3a showing the login screen, Figure 6.3b showing the chat box, Figure 6.3c showing the evaluation report after a conversation, and Figure 6.3d showing the control panel containing historical training records.

Taking clinical handover on the COVID-19 case as an example, once the communication task is created using relevant data, trainees can practice communication via the

[3]https://cloud.google.com/translate

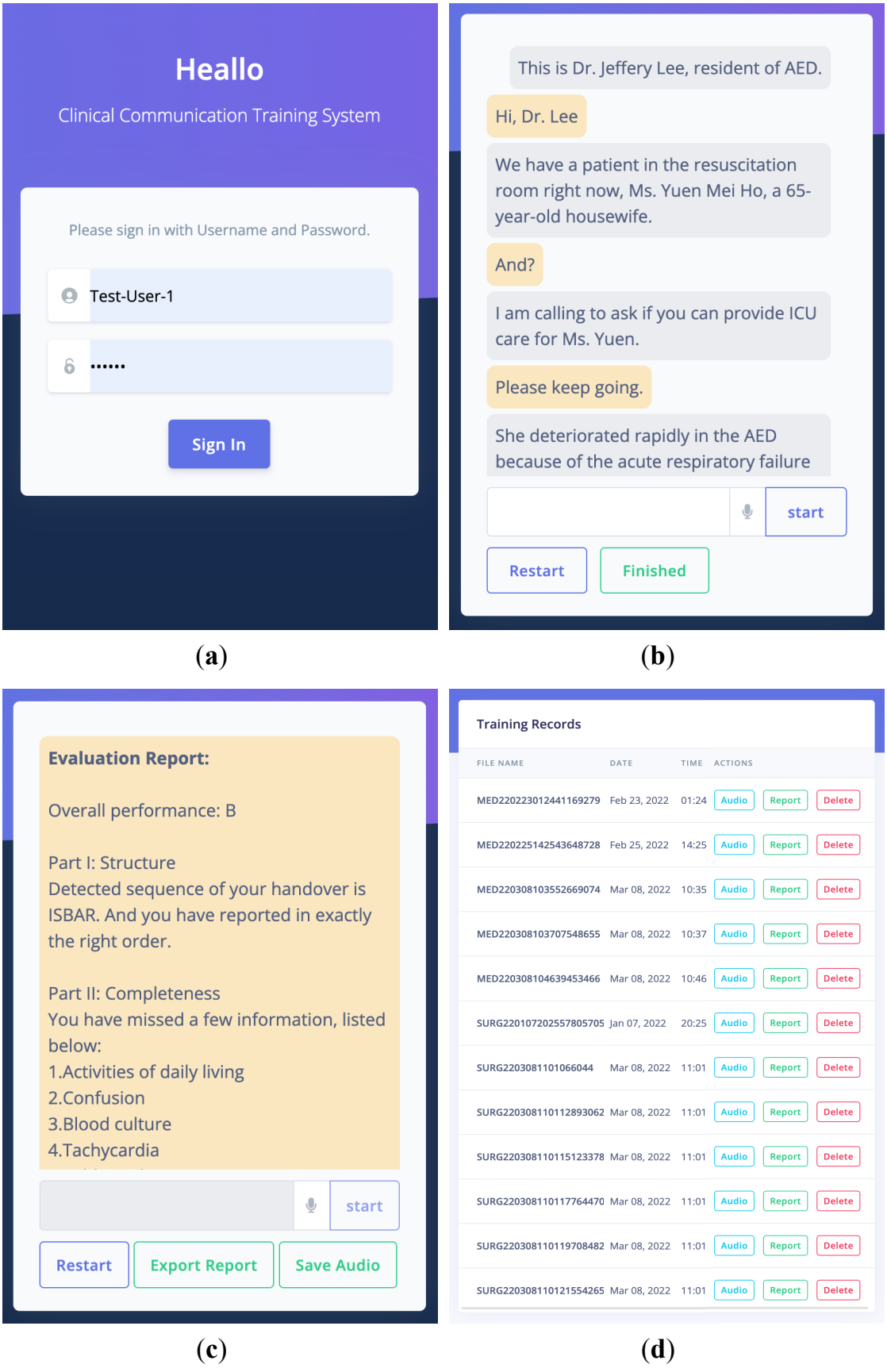

Figure 6.3: Interface of Heallo for trainees. (a) Login screen. (b) Chatting box. (c) Evaluation report. (d) Control panel.

chat interface. Table 6.2 presents a sample interaction between the user and the conversational system. For each input sentence, the system detects its intent and determines the presence of predefined components within it. The system can accurately detect intents and components even when expressions differ from the sample data. Then the system generates an appropriate response based on the detection results. Once the conversation ends, the system calculates the overall score and generates a detailed evaluation report using the results of all processing performed during the conversation and the grading scheme.

6.3 Experiments

Heallo was incorporated into the intern simulation training program at Queen Elizabeth Hospital. A specialized communication task was established by clinical experts, with approximately 30 participants utilizing Heallo for their clinical communication training. We also engaged medical professionals to rate these conversations, comparing the results generated by Heallo with those from human raters. The human raters were guided to adhere to the same grading schemes employed in our system. A tailored questionnaire was also administered to gather information about the user experience and the effectiveness of the training.

We finally obtained 29 training results from medical interns partaking in the intern simulation training program at Queen Elizabeth Hospital. Table 6.3 collates the overall grades attributed to each trainee by Heallo, two independent human raters, and a random scoring system for comparative purposes.

Table 6.3 reveals some variance in the scores from the raters, predominantly due to individual discrepancies in their understanding and information capture. It's worth not-

User	**Conversational System**	
[*Input*] This is Dr. Jeffery Lee, resident of AED. Are you ICU Dr. Ng?	[*Detected Intent*] [*Detected Components*] [*Output*]	I Doctor's name; Location Hi, Dr. Lee.
[*Input*] I'm calling about Ms. Yuen, who may require ICU care due to respiratory failure caused by COVID infection.	[*Detected Intent*] [*Detected Components*] [*Output*]	S Patient's name; Gender; COVID; Respiratory failure Please go on.
[*Input*] Ms. Yuen Mei Ho is 65 years old and is currently in the resuscitation room.	[*Detected Intent*] [*Detected Components*] [*Output*]	I Patient's name; Gender; Age; Location Ok.
[*Input*] She was admitted this morning at 9:30 a.m. for a fever and breathing problems.	[*Detected Intent*] [*Detected Components*] [*Output*]	B Gender; Admission time; Fever Well, please keep going.
[*Input*] Her condition is deteriorated rapidly with saturation was 90% in 4L/min O2. Now she is confused.	[*Detected Intent*] [*Detected Components*] [*Output*]	A Gender; Deterioration; Saturation; Coma Anything else?
[*Input*] We will intubate the patient. Would you like to come over and assess the patient?	[*Detected Intent*] [*Detected Components*] [*Output*]	R Ok, I will come as soon as possible.
[*Finished*]	[*Output*]	Evaluation Report: Overall performance: B Part I: Structure Detected sequence of your handover is ISIBAR, and the correct sequence should be ISBAR. Part II: Completeness You have missed the following information: 1. Diabetes mellitus 2. Activities of daily living 3. Vital signs 4. Vaccination 5. Chest X-ray

Table 6.2: Interactions between the user and the conversational system.

Sample #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Heallo	B	B	A	B	B	B	B	A	B	B	B	B	C	B	C	B	C	B	B	A	B	A	A	A	C	B	B	A	B
Rater 1	C	C	C	B	C	A	C	A	B	B	B	C	C	B	C	C	C	C	B	A	A	B	B	B	C	B	B	B	B
Rater 2	C	B	A	B	B	A	B	A	B	A	A	B	B	A	C	A	C	A	B	A	A	B	B	A	C	B	C	B	A
Random	F	D	D	C	F	A	B	D	F	F	F	D	F	C	C	F	A	B	A	D	C	A	A	F	F	B	B	A	B

Table 6.3: Overall grades from Heallo and two clinical practitioners.

ing the relatively low incidence of 'A' grades (and the even rarer instance of unanimous 'A' grades from all raters), suggesting ample opportunity for most medical interns to enhance their clinical communication skills through further training.

Comparing the grading groups, the majority of Heallo's predicted grades align with the range between the human raters' results, underlining the system's reliability. To streamline the analysis, we transformed the categorical grades into numerical scores (i.e., A, B, C, D, and F corresponding to 5, 4, 3, 2, and 1, respectively). The average score from Heallo (4.10) resides between that of Rater 1 (3.72) and Rater 2 (4.24), all of which substantially outperform the random scoring average (2.93). And the number of agreements between Heallo and human evaluators (21) matches that between two human evaluators.

To scrutinize the difference between Heallo and human raters further, we computed the grading difference between Heallo and Rater 1 (group 1), Heallo and Rater 2 (group 2), and Rater 1 and Rater 2 (group 3). The average discrepancies for groups 1, 2, and 3 are 0.52, 0.48, and 0.59, respectively, implying marginal variations between the three raters' grades. A t-test was used to determine the significance of these differences ($\alpha = 0.05$). The results indicate no significant differences between groups 1 and 3 (p-value = 0.68), or between groups 2 and 3 (p-value = 0.52), suggesting Heallo's scoring precision is comparable to that of the clinical experts.

Throughout the experiment, Heallo provided a comprehensive evaluation report immediately following each conversation, while human raters were required to listen to recordings repeatedly to discern the intent sequence and the components discussed. Supplanting human raters with Heallo is projected to conserve approximately 0.5 person-hours per training session.

Post-experiment, a follow-up questionnaire was administered to the trainees who were conversant with the communication protocol. Of the 23 respondents, 82.6% acknowledged the necessity of practicing clinical communication in simulated scenarios, and 73.9% affirmed the importance of evaluative feedback to guide their training. These trainees were also asked to rate Heallo using a 5-point Likert scale, with the average user experience score being 4.00 and the average training effectiveness score being 3.83.

Based on the comparative analysis with clinical raters and the feedback from trainees, we confidently assert that Heallo has the potential to deliver effective autonomous clinical communication training on user-defined tasks. Our intention is to continue refining the system in accordance with evolving training needs and user feedback.

6.4 Conclusion

In this research, we have presented Heallo, a pioneering system designed to facilitate autonomous clinical communication training. Heallo is capable of simulate diverse clinical communication scenarios, evaluate clinicians' performance, and accommodate new tasks with simple editing. When applied in a practical setting, Heallo is capable of delivering evaluations comparable to human raters. It has also garnered positive feedback in terms of user experience and the impact on training outcomes.

This exploration of utilizing NLP technologies in clinical communication training

is not merely an academic exercise but an essential step towards enhancing healthcare communication. It's our aspiration that this endeavor will foster progress in the development of sophisticated communication training systems, inspiring further exploration and innovations within the realms of NLP and intelligent healthcare.

Chapter 7

Conclusion

7.1 Summary

Healthcare has received more and more attention with the development of society and the adoption of intelligent healthcare is on the rise. Clinical communication has always been an integral part of healthcare education, and it is also the key to providing safe, high-quality patient care. Faced with a scarcity of medical resources and a high demand for training, this thesis proposes customizable conversational systems for intelligent clinical communication training.

In Chapter 1, we bring up the framework for an autonomous, low-cost, customizable clinical communication training system and formulate three research problems from it.

In Chapter 2, the applications and general techniques of conversational systems are reviewed. Then, we present our preliminary work on personal chatbot customization.

Chapter 3 focuses on sentence-level intent detection in standardized clinical communication. We first collect a standard clinical handover dataset, CLINIC-ISBAR, of real-world cases in collaboration with QEH practitioners. We propose the novel Intent-

aware Long Short-term Memory (IA-LSTM) model referring to the context of clinical communication standards. Extensive experiments and comparisons on CLINIC-ISBAR have validated effectiveness, generalizability and robustness of our intent-aware design. The collected dataset and proposed algorithm lay the groundwork for the implementation of clinical communication training systems.

Chapter 4 investigates the task of content recognition in clinical conversation by incorporating explicit knowledge from biomedical ontology and implicit knowledge from pretrained language models. Due to the variability in expression habits among doctors and the existence of multiple forms for medical terms, we constructed a knowledge graph utilizing biomedical ontology. We proposed Knowledge-infused Prompt Learning (KIPL) to incorporate external knowledge into prompts as cues and hints for the pre-trained language model. Our experimental results demonstrate that KIPL achieves superior performance, particularly exhibiting significant advantages in situations with limited data or complex components.

Chapter 5 investigates the problem of customizable conversational system with insufficient training data. We propose Data Augmentation with User-Defined Knowledge (UDK-DA), which increases the robustness and generalizability of ML models by injecting user-defined lexical knowledge and context knowledge into training samples. UDK-DA enables non-IT users to design tasks using a small number of samples and self-defined knowledge. Experiments demonstrate that when training data is insufficient, UDK-DA can significantly improve the performance of learning models, allowing non-professionals to design new communication tasks with minimal editing.

In Chapter 6, we present the conversational system, Heallo, for autonomous, low-cost and customizable clinical communication training. Heallo is able to simulate clinical communication scenarios, evaluate the performance of clinicians, and be customized

to new tasks with minimal editing. In real-world practice, Heallo is able to provide expert-level evaluations and receives positive feedback regarding the user experience and training outcomes.

We hope that this attempt to apply NLP technologies to clinical communication training will promote the development of intelligent healthcare and motivate NLP researchers.

7.2 Future Work

Heallo can be viewed as an initial attempt at a customizable conversational system for clinical communication training; There are still several areas that can be improved.

- In designing algorithms and experiments, we focus primarily on the clinical handover task. In the future, we will collect data for additional task scenarios in an effort to enhance the model.
- In this paper, we discuss three important parts for designing a communication training system. There are other modules within the conversational system that can be enhanced. For instance, the speech recognition module is not optimized for clinical communication, and our response retriever does not provide a realistic interactive experience. In the future, we will attempt to refine the system further.
- Currently, we only pilot in the Queen Elizabeth Hospital for the real-world usage, and the experiment's personnel and duration are limited. In the future, we will promote it to hospitals in Hong Kong and other regions in order to support clinical communication training and intelligent healthcare. In the interim, additional data and feedback can be gathered to further enhance the system.

- Our system currently only supports English. Future objectives also include the creation of algorithms that support multiple languages, which will benefit more practitioners.

References

[1] Sameera A Abdul-Kader and JC Woods. “Survey on chatbot design techniques in speech conversation systems”. In: *International Journal of Advanced Computer Science and Applications* 6.7 (2015).

[2] Bayan AbuShawar and Eric Atwell. “ALICE chatbot: Trials and outputs”. In: *Computación y Sistemas* 19.4 (2015), pp. 625–632.

[3] Eleni Adamopoulou and Lefteris Moussiades. “Chatbots: History, technology, and applications”. In: *Machine Learning with Applications* 2 (2020), p. 100006.

[4] Fatima Alshehri and Ghulam Muhammad. “A comprehensive survey of the Internet of Things (IoT) and AI-based smart healthcare”. In: *IEEE Access* 9 (2020), pp. 3660–3678.

[5] Zahra Ashktorab et al. “Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 254.

[6] Anthony Back, James A Tulsky, and Robert M Arnold. *Communication skills in the age of COVID-19*. 2020.

[7] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *3rd International Conference on Learning Representations, ICLR 2015*. 2015.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *International Conference on Learning Representations*. 2015.

[9] Bobby Batacharia et al. “CONVERSE: a conversational companion”. In: *Machine conversations*. Springer, 1999, pp. 205–215.

[10] Aditya Bhargava et al. “Easy contextual intent prediction and slot detection”. In: *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8337–8341. ISBN: 1479903566.

[11] Hemanthage S Bhathiya and Uthayasanker Thayasivam. “Meta learning for few-shot joint intent detection and slot-filling”. In: *Proceedings of the 2020 5th International Conference on Machine Learning Technologies*. 2020, pp. 86–92.

[12] Abdullah Bin Sawad et al. “A systematic review on healthcare artificial intelligent conversational agents for chronic conditions”. In: *Sensors* 22.7 (2022), p. 2625.

[13] Petter Bae Brandtzaeg and Asbjørn Følstad. “Why people use chatbots”. In: *International Conference on Internet Science*. Springer. 2017, pp. 377–392.

[14] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[15] Sharon Buckley et al. “Tools for structured team communication in pre-registration health professions education: a Best Evidence Medical Education (BEME) review: BEME Guide No. 41”. In: *Medical teacher* 38.10 (2016), pp. 966–980.

[16] Leonardo Campillos-Llanos et al. “Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation”. In: *Natural Language Engineering* 26.2 (2020), pp. 183–220.

[17] Heloisa Candello, Claudio Pinhanez, and Flavio Figueiredo. “Typefaces and the perception of humanness in natural language chatbots”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 3476–3487.

[18] Lorainne Tudor Car et al. “Conversational agents in health care: scoping review and conceptual analysis”. In: *Journal of medical Internet research* 22.8 (2020), e17158.

[19] Sonali Chandel et al. “Chatbot: efficient and utility-based platform”. In: *Science and Information Conference*. Springer. 2018, pp. 109–122.

[20] Bonnie Chantarotwong. “The learning chatbot”. In: *Final year project.[Online]: http://courses. ischool. berkeley. edu/i256/f06/projects/bonniejc. pdf* (2006).

[21] Xiang Chen et al. “Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction”. In: *Proceedings of the ACM Web conference 2022*. 2022, pp. 2778–2788.

[22] Peng Cheng and Utz Roedig. “Personal voice assistant security and privacy—a survey”. In: *Proceedings of the IEEE* 110.4 (2022), pp. 476–507.

[23] Kenneth Mark Colby. *Artificial paranoia: a computer simulation of paranoid process*. Pergamon Press, 1975.

[24] Antonio Coronato et al. "Reinforcement learning for intelligent healthcare applications: A survey". In: *Artificial Intelligence in Medicine* 109 (2020), p. 101964.

[25] CSSEGISandData. *CSSEGISANDDATA/covid-19: Novel coronavirus (COVID-19) cases, provided by JHU CSSE*. URL: https://github.com/CSSEGISandData/COVID-19.

[26] Robert Dale. "The return of the chatbots". In: *Natural Language Engineering* 22.5 (2016), pp. 811–817.

[27] Thomas Davenport and Ravi Kalakota. "The potential for artificial intelligence in healthcare". In: *Future healthcare journal* 6.2 (2019), p. 94.

[28] David DeVault et al. "SimSensei Kiosk: A virtual human interviewer for healthcare decision support". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2014, pp. 1061–1068.

[29] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186.

[30] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[31] Kevin Donnelly et al. “SNOMED-CT: The advanced terminology and coding system for eHealth”. In: *Studies in health technology and informatics* 121 (2006), p. 279.

[32] John Dowding et al. “Gemini: A natural language system for spoken-language understanding”. In: *arXiv preprint cmp-lg/9407007* (1994).

[33] Lisa Ehrlinger and Wolfram Wöß. “Towards a definition of knowledge graphs.” In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48.1-4 (2016), p. 2.

[34] Aysu Ezen-Can. “A Comparison of LSTM and BERT for Small Corpus”. In: *arXiv preprint arXiv:2009.05451* (2020).

[35] Ahmed Fadhil and Silvia Gabrielli. “Addressing challenges in promoting healthy lifestyles: the al-chatbot approach”. In: *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare*. 2017, pp. 261–265.

[36] Steven Y Feng et al. “A Survey of Data Augmentation Approaches for NLP”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, pp. 968–988.

[37] Arnstein Finset et al. “Effective health communication–a key factor in fighting the COVID-19 pandemic”. In: *Patient education and counseling* 103.5 (2020), p. 873.

[38] Cynthia Foronda, Brent MacWilliams, and Erin McArthur. “Interprofessional communication in healthcare: An integrative review”. In: *Nurse education in practice* 19 (2016), pp. 36–40.

[39] Adriana Foster et al. “Using virtual patients to teach empathy: a randomized controlled study to enhance medical students’ empathic communication”. In: *Simulation in Healthcare* 11.3 (2016), pp. 181–189.

[40] Debasis Ganguly et al. “Word embedding based generalized language model for information retrieval”. In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 2015, pp. 795–798.

[41] Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. “Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system”. In: *Stud Health Technol Inform* 252 (2018), pp. 51–56.

[42] *Global Smart Healthcare Market Size, share report, 2030*. URL: https://www.grandviewresearch.com/industry-analysis/smart-healthcare-market.

[43] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. In: *IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649. ISBN: 1479903566.

[44] David Griol and Zoraida Callejas. “Mobile conversational agents for context-aware care applications”. In: *Cognitive Computation* 8.2 (2016), pp. 336–356.

[45] Yuxian Gu et al. “PPT: Pre-trained Prompt Tuning for Few-shot Learning”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 8410–8423.

[46] Patrick Haffner, Gokhan Tur, and Jerry H Wright. "Optimizing SVMs for complex call classification". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. Vol. 1. IEEE. 2003, pp. I–I.

[47] Xu Han et al. "Ptr: Prompt tuning with rules for text classification". In: *AI Open* 3 (2022), pp. 182–192.

[48] Braden Hancock et al. "Learning from Dialogue after Deployment: Feed Yourself, Chatbot!" In: *arXiv preprint arXiv:1901.05415* (2019), pp. 3667–3684.

[49] Nicolas Heist et al. "Knowledge Graphs on the Web–An Overview". In: *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges* (2020), pp. 3–22.

[50] Ryuichiro Higashinaka et al. "Towards an open-domain conversational system fully based on natural language processing". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 928–939.

[51] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015).

[52] Sepp Hochreiter, J urgen Schmidhuber, and Corso Elvezia. "LONG SHORT-TERM MEMORY". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[53] Oasis Hu and Bingcun Li. *Hong Kong boosts COVID-19 efforts*. Mar. 2022. URL: https://www.chinadailyhk.com/article/262900.

[54] Shengding Hu et al. “Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 2225–2240.

[55] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. “Evorus: A crowd-powered conversational assistant built to automate itself over time”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 295.

[56] Bernd Huber et al. “Emotional dialogue generation using image-grounded language models”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 277.

[57] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. “A survey on conversational agents/chatbots classification and design techniques”. In: *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer. 2019, pp. 946–956.

[58] Mohit Jain et al. “Convey: Exploring the use of a context view for chatbots”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 468.

[59] Armand Joulin et al. “Fasttext. zip: Compressing text classification models”. In: *arXiv preprint arXiv:1612.03651* (2016).

[60] Young Juhn and Hongfang Liu. “Artificial intelligence approaches using natural language processing to advance EHR-based clinical research”. In: *Journal of Allergy and Clinical Immunology* 145.2 (2020), pp. 463–469.

[61] Takeshi Kamita et al. “A chatbot system for mental healthcare based on SAT counseling method”. In: *Mobile Information Systems* 2019 (2019).

[62] Tian Kang et al. “UMLS-based data augmentation for natural language processing of clinical research literature”. In: *Journal of the American Medical Informatics Association* 28.4 (2021), pp. 812–823.

[63] Young-Bum Kim, Sungjin Lee, and Karl Stratos. “Onenet: Joint domain, intent, slot prediction for spoken language understanding”. In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2017, pp. 547–553.

[64] Ahmad R Kirmani. “Artificial intelligence-enabled science poetry”. In: *ACS Energy Letters* 8.1 (2022), pp. 574–576.

[65] Sebastian Köhler et al. “The human phenotype ontology in 2021”. In: *Nucleic acids research* 49.D1 (2021), pp. D1207–D1217.

[66] Kamran Kowsari et al. “Text classification algorithms: A survey”. In: *Information* 10.4 (2019), p. 150.

[67] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. “Data augmentation using pre-trained transformer models”. In: *arXiv preprint arXiv:2003.02245* (2020).

[68] Siwei Lai et al. “Recurrent convolutional neural networks for text classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 2015. ISBN: 2374-3468.

[69] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[70] Michael Leonard, Suzanne Graham, and Doug Bonacum. “The human factor: the critical importance of effective teamwork and communication in providing safe care”. In: *BMJ Quality & Safety* 13.suppl 1 (2004), pp. i85–i90. ISSN: 2044-5415.

[71] Jiwei Li et al. “Dialogue learning with human-in-the-loop”. In: *arXiv preprint arXiv:1611.09823* (2016).

[72] Jiwei Li et al. “Learning through dialogue interactions by asking questions”. In: *arXiv preprint arXiv:1612.04936* (2016).

[73] Kam Cheong Li et al. “Evaluation of mobile learning for the clinical practicum in nursing education: application of the FRAME model”. In: *Journal of Computing in Higher Education* 31 (2019), pp. 290–310.

[74] Linfeng Li et al. “Real-world data medical knowledge graph: construction and applications”. In: *Artificial intelligence in medicine* 103 (2020), p. 101817.

[75] Xiang Lisa Li and Percy Liang. “Prefix-Tuning: Optimizing Continuous Prompts for Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 4582–4597.

[76] Bing Liu et al. “Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems”. In: *arXiv preprint arXiv:1804.06512* (2018).

[77] Jiao Liu, Yanling Li, and Min Lin. “Review of intent detection methods in the human-machine dialogue system”. In: *Journal of Physics: Conference Series*. Vol. 1267. 1. IOP Publishing. 2019, p. 012059.

[78] Pengfei Liu et al. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. In: *ACM Computing Surveys* 55.9 (2023), pp. 1–35.

[79] Xiao Liu et al. “GPT understands, too”. In: *arXiv preprint arXiv:2103.10385* (2021).

[80] Ewa Luger and Abigail Sellen. “Like having a really bad PA: the gulf between user expectation and experience of conversational agents”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 5286–5297.

[81] Brady D Lund and Ting Wang. “Chatting about ChatGPT: how may AI and GPT impact academia and libraries?” In: *Library Hi Tech News* 40.3 (2023), pp. 26–29.

[82] Computing Machinery. “Computing machinery and intelligence-AM Turing”. In: *Mind* 59.236 (1950), p. 433.

[83] James Malone et al. “Modeling sample variables with an Experimental Factor Ontology”. In: *Bioinformatics* 26.8 (2010), pp. 1112–1118.

[84] James Manyika. *An overview of Bard: an early experiment with generative AI*. https://ai.google/static/documents/google-about-bard.pdf. Accessed: 2023-07-12. 2023.

[85] Stuart Marshall, J Harrison, and B Flanagan. “The teaching of a structured tool improves the clarity and content of interprofessional clinical communication”. In: *BMJ Quality & Safety* 18.2 (2009), pp. 137–140. ISSN: 2044-5415.

[86] Yoichi Matsuyama et al. “Socially-aware animated intelligent personal assistant agent”. In: *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*. 2016, pp. 224–227.

[87] Andrew McCallum, Kamal Nigam, et al. “A comparison of event models for naive bayes text classification”. In: *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1. Citeseer. 1998, pp. 41–48.

[88] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[89] Tomáš Mikolov et al. “Advances in Pre-Training Distributed Word Representations”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[90] Zachary Munn et al. *Developing guidelines before, during, and after the COVID-19 pandemic*. 2020.

[91] Tom Nadarzynski et al. “Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study”. In: *Digital health* 5 (2019), p. 2055207619871808.

[92] Hoang Long Nguyen, Dang Thinh Vu, and Jason J Jung. “Knowledge graph fusion for smart systems: A survey”. In: *Information Fusion* 61 (2020), pp. 56–70.

[93] Xiaolei Niu and Yuexian Hou. “Hierarchical Attention BLSTM for Modeling Sentences and Documents”. In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 167–177.

[94] OpenAI. *ChatGPT*. May 23 Version. https://chat.openai.com/. 2023.

[95] Ankur P Parikh et al. "A decomposable attention model for natural language inference". In: *arXiv preprint arXiv:1606.01933* (2016), pp. 2249–2255.

[96] Soya Park et al. "Facilitating knowledge sharing from domain experts to data scientists for building nlp models". In: *26th International Conference on Intelligent User Interfaces*. 2021, pp. 585–596.

[97] Heiko Paulheim. "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic web* 8.3 (2017), pp. 489–508.

[98] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[99] Juanan Pereira and Óscar Díaz. "Using health chatbots for behavior change: a mapping study". In: *Journal of medical systems* 43 (2019), pp. 1–13.

[100] Rob Price. "Microsoft is deleting its AI chatbot's incredibly racist tweets". In: *Business Insider* (2016).

[101] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[102] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[103] Abhinav Rastogi et al. "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 8689–8696. ISBN: 2374-3468.

[104] Haran Ratna. “The importance of effective communication in healthcare practice”. In: *Harvard Public Health Review* 23 (2019), pp. 1–6.

[105] Baidu Research. *ERNIE Bot: Baidu’s Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology*. http://research.baidu.com/Blog/index-view?id=183. Accessed: 2023-07-12. 2023.

[106] Hyekyun Rhee et al. “Mobile phone-based asthma self-management aid for adolescents (mASMAA): a feasibility study”. In: *Patient preference and adherence* 8 (2014), p. 63.

[107] Verena Rieser and Oliver Lemon. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media, 2011.

[108] Stephen E Robertson and Steve Walker. “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval”. In: *SIGIR’94*. Springer. 1994, pp. 232–241.

[109] WHO Patient Safety and World Health Organization. “Patient safety curriculum guide: Multi-professional edition”. In: (2011). ISSN: 8555268508.

[110] Stefan Schulz et al. “Strengths and limitations of formal ontologies in the biomedical domain”. In: *Revista electronica de comunicacao, informacao & inovacao em saude: RECIIS* 3.1 (2009), p. 31.

[111] Joseph Seering et al. “Beyond Dyadic Interactions: Considering Chatbots as Community Members”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 450.

[112] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Improving neural machine translation models with monolingual data”. In: *arXiv preprint arXiv:1511.06709* (2015).

[113] Mohammed Yousef Shaheen. “Applications of Artificial Intelligence (AI) in healthcare: A review”. In: *ScienceOpen Preprints* (2021).

[114] Lifeng Shang, Zhengdong Lu, and Hang Li. “Neural responding machine for short-text conversation”. In: *arXiv preprint arXiv:1503.02364* (2015), pp. 1577–1586.

[115] Bayan Abu Shawar and Eric Atwell. “Chatbots: are they really useful?” In: *Ldv forum*. Vol. 22. 1. 2007, pp. 29–49.

[116] Ying Shen et al. “CBN: Constructing a clinical Bayesian network based on data from the electronic medical record”. In: *Journal of biomedical informatics* 88 (2018), pp. 1–10.

[117] Taylor Shin et al. “AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 4222–4235.

[118] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1 (2019), pp. 1–48.

[119] Shijing Si et al. “Students Need More Attention: BERT-based Attention Model for Small Data with Application to Automatic Patient Message Triage”. In: *Machine Learning for Healthcare Conference*. PMLR. 2020, pp. 436–456.

[120] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[121] Rincy Mariam Thomas et al. "Survey on Artificially Intelligent Chatbot". In: *Journal of Applied Science and Computations* 6.1 (2019), pp. 85–94.

[122] Ashish Vaswani et al. "Attention is All you Need". In: *NIPS*. 2017.

[123] Wei Wang et al. "Context-aware intent identification in email conversations". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 585–594.

[124] Jason Wei and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". 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)*. 2019, pp. 6382–6388.

[125] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[126] HENRY Weld et al. "A survey of joint intent detection and slot-filling models in natural language understanding". In: *arXiv preprint arXiv:2101.08091* (2021).

[127] Jason E Weston. "Dialog-based language learning". In: *Advances in Neural Information Processing Systems*. 2016, pp. 829–837.

[128] Thomas Wolf et al. “Transformers: State-of-the-art natural language processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45.

[129] Zhiyuan Wu et al. “Rumor detection based on propagation graph neural network with attention mechanism”. In: *Expert systems with applications* 158 (2020), p. 113595.

[130] Rui Yan. “” Chitty-Chitty-Chat Bot”: Deep Learning for Conversational AI.” In: *IJCAI*. Vol. 18. 2018, pp. 5520–5526.

[131] Rui Yan, Yiping Song, and Hua Wu. “Learning to respond with deep neural networks for retrieval-based human-computer conversation system”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016, pp. 55–64.

[132] Rui Yan et al. “Shall i be your chat companion?: Towards an online human-computer conversation system”. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM. 2016, pp. 649–658.

[133] Zi Yin, Keng-hao Chang, and Ruofei Zhang. “Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 2131–2139.

[134] Yue Yuan, Yanli Wang, and Kan Liu. “Perceiving more truth: A dilated-block-based convolutional network for rumor identification”. In: *Information Sciences* 569 (2021), pp. 746–765.

[135] Aohan Zeng et al. “GLM-130B: An Open Bilingual Pre-trained Model”. In: *The Eleventh International Conference on Learning Representations*. 2022.

[136] Haichao Zhang, Haonan Yu, and Wei Xu. “Listen, interact and talk: Learning to speak via interaction”. In: *arXiv preprint arXiv:1705.09906* (2017).

[137] Hongyi Zhang et al. “mixup: Beyond Empirical Risk Minimization”. In: *International Conference on Learning Representations*. 2018.

[138] Saizheng Zhang et al. “Personalizing Dialogue Agents: I have a dog, do you have pets too?” In: *arXiv preprint arXiv:1801.07243* (2018), pp. 2204–2213.

[139] Ye Zhang and Byron C Wallace. “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2017, pp. 253–263.

[140] Chao Zhao et al. “EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning”. In: *Artificial intelligence in medicine* 87 (2018), pp. 49–59.

[141] Hao Zhou et al. “Emotional chatting machine: Emotional conversation generation with internal and external memory”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[142] Sijia Zhou and Xin Li. “Feature engineering vs. deep learning for paper section identification: Toward applications in Chinese medical literature”. In: *Information Processing & Management* 57.3 (2020), p. 102206.

[143] Xiaohan Zou. “A survey on application of knowledge graph”. In: *Journal of Physics: Conference Series*. Vol. 1487. 1. IOP Publishing. 2020, p. 012016.