Research on Emotional Conversation based on Deep Learning Approach

She Tianhao

A Thesis submitted to Tokushima University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March, 2022

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
Tokushima University, Japan

Contents

List of Figures

List of Tables

Acknowledgment

This thesis is carried out during my study in University of Tokushima, from April 2019 to March 2022. On the completion of my thesis, I would like to take this opportunity to express my gratitude to everyone who helped and supported me in my Ph.D. course.

First and foremost, I would like to give my sincere gratitude to my supervisor, Prof. Fuji Ren. He guided me along the path of computer science and helped me with my exposure to the research field of affective computing. He provided me with an excellent research environment. With his patient guidance and encouragement, I have the chance to attend international conferences and publish papers in authoritative journals. These research experiences not only enriched my knowledge, but also taught me divergents thinking and allowed me to think comprehensively when solving problems.

I would like to express my sincere gratitude to Professor Kenji Terada and Professor Masami Shishibori for their time and effort in reviewing my thesis and thank for their valuable suggestions.

I would like to thank Shun Nishide and Xin Kang for their practical advice and technical assistance during my Ph.D. course.

Meanwhile, I would like to thank my classmates in the A1 group. Thanks to Qian Zhang, Ziyun Jiao, Yangyang Zhou, and Zheng Liu for their valuable suggestions and shared experience in research. I would also like to thank the graduated seniors, Ning Liu, Duo Feng, Mengjia He, Jiawen Deng, and Siyuan Xue. It is your help that made my research at Tokushima University start so smoothly. Thanks to my classmates in the A1 group for their company and encouragement in life and research.

Last and certainly not least, Thank my parents for their unconditional love and support for me. They supported me in overcoming every obstacle I had met over the past three years.

Abstract

Owning the capability to express specific emotions by a chatbot during a conversation is one of the key parts of artificial intelligence, which has an intuitive and quantifiable impact on the improvement of chatbot’s usability and user satisfaction. Enabling machines to emotion recognition in conversation is challenging, mainly because the information in human dialogue innately conveys emotions by long-term experience, abundant knowledge, context, and the intricate patterns between the affective states. Recently, many studies on neural emotional conversational models have been conducted. However, enabling the chatbot to control what kind of emotion to respond to upon its own characters in conversation is still underexplored. At this stage, people are no longer satisfied with using a dialogue system to solve specific tasks, and are more eager to achieve spiritual communication. In the chat process, if the robot can perceive the user's emotions and can accurately process them, it can greatly enrich the content of the dialogue and make the user empathize.

In the process of emotional dialogue, our ultimate goal is to make the machine understand human emotions and give matching responses. Based on these two points, this thesis explores and in-depth emotion recognition in conversation task and emotional dialogue generation task. In the past few years, although considerable progress has been made in emotional research in dialogue, there are still some difficulties and challenges due to the complex nature of human emotions. The key contributions in this thesis are summarized as below:

(1) Researchers have paid more attention to enhancing natural language models with knowledge graphs these days, since knowledge graph has gained a lot of systematic knowledge. A large number of studies had shown that the introduction of external commonsense knowledge is very helpful to improve the characteristic information. We address the task of emotion recognition in conversations using external knowledge to enhance semantics. In this work, we employ an external knowledge graph ATOMIC to

extract the knowledge sources. We proposed KES model, a new framework that incorporates different elements of external knowledge and conversational semantic role labeling, where build upon them to learn interactions between interlocutors participating in a conversation. The conversation is a sequence of coherent and orderly discourses. For neural networks, the capture of long-range context information is a weakness. We adopt Transformer a structure composed of self-attention and feed forward neural network, instead of the traditional RNN model, aiming at capturing remote context information. We design a self-attention layer specialized for enhanced semantic text features with external commonsense knowledge. Then, two different networks composed of LSTM are responsible for tracking individual internal state and context external state. In addition, the proposed model has experimented on three datasets in emotion detection in conversation. The experimental results show that our model outperforms the state-of-the-art approaches on most of the tested datasets.

(2) We proposed an emotional dialogue model based on Seq2Seq, which is improved from three aspects: model input, encoder structure, and decoder structure, so that the model can generate responses with rich emotions, diversity, and context. In terms of model input, emotional information and location information are added based on word vectors. In terms of the encoder, the proposed model first encodes the current input and sentence sentiment to generate a semantic vector, and additionally encodes the context and sentence sentiment to generate a context vector, adding contextual information while ensuring the independence of the current input. On the decoder side, attention is used to calculate the weights of the two semantic vectors separately and then decode, to fully integrate the local emotional semantic information and the global emotional semantic information. We used seven objective evaluation indicators to evaluate the model's generation results, context similarity, response diversity, and emotional response. Experimental results show that the model can generate diverse responses with rich sentiment, contextual associations.

Chapter 1

Introduction

1.1 Motivation

To have a virtual assistant or a chat system with adequate intelligence has seemed elusive and might only exist in science fiction film movies and novels for a long time. Recently, human-machine dialogue has received more and more attention due to its vast potential and attractive commercial value.

With the development of big data and deep learning technology, the goal of creating an automated human-machine dialogue system as our personal assistant or chat companion is no longer an illusion. Due to the emergence of smart personal devices (such as smartphones, smart speakers, car control terminals, etc.) in the past decade, building a dialogue model for barrier-free communication with humans has become the focus of research. Predominantly, these devices will have a single user who will interact many times over a long period of time, in some sense building a “relationship” with the user. Building machines that are able to vocally communicate with users is driven by the desire to make human-machine interaction as natural and efficient as possible. Due to the remarkable advances in Automatic Speech Recognition and speech synthesis, many personal assistant applications (such as Siri, Google Assistant, Cortana, Alexa, etc.) have been fielded.

Implementing the way humans converse in a machine involves many issues: no more need for traditional interaction devices like the keyboard, hands-free and eye-free interaction, new communication paradigms where a real human-human conversation is simulated. These issues are also emphasized by the new technological trends in the modern world: the Internet of Things (IoT), smart devices, etc. Also, recent advances gave birth to vocal agents both in academia and industry. Nowadays, it is possible to check for restaurants around, check one's account, or send a search query on the web by uttering a few words only. Also, conversational agents are able to understand more and more vocabulary and language variations.

On the one hand, nowadays we can easily access big data for conversations on the Web and we might be able to learn how to respond and what to respond given any inputs, which greatly allows us to build data-driven, open-domain conversation systems between humans and computers. On the other hand, deep learning techniques have been proven to be effective in capturing complex patterns in big data and have powered numerous research fields such as computer vision, natural language processing and recommender systems. Hence, a large body of literature has emerged to leverage a massive amount of data via deep learning to advance dialogue systems in many perspectives. Nevertheless, the ability of these systems to carry out real and natural dialogue is still very limited, which is due to the lack of anthropomorphic emotions. Emotion interaction is a common psychological phenomenon in human's daily life. Accurate emotion recognition is the premise of effective human communication, interaction, and decision-making. Only in the past few years has emotional conversation gained attention from the research community due to the increase of public availability of conversational data. Nevertheless, for this research, it is necessary to continuously develop and optimize calculation methods to achieve more accurate emotion prediction in conversations and generate utterances with richer emotions.

1.2 Significance of Research

The main goal of constructing a dialogue system is to solve users' problems and concerns by simulating the way humans communicate with each other. Since human language is too complex to be regarded as a single goal, dialogue systems must model different aspects of human communication separately. Hence, a large body of literature has emerged to leverage a massive amount of data via deep learning to advance dialogue systems in many perspectives. Recent years have witnessed the emotional conversation models in the context of dialogue systems and, hence, increasing attention from the natural language processing community. Empathy is the capability of projecting the feelings and ideas of the other party to someone's knowledge. Integrating empathy into the design of the dialogue system is essential to improve the user experience in human-computer interaction. Empathy and empathy are necessary factors for the dialogue system to be regarded as a social partner by humans. Emotional support aims at reducing individuals' emotional distress and helping them understand and work through the challenges that they face. Despite the importance and complexity of emotional support, research on data-driven emotional support dialog systems is limited due to a lack of both task design and relevant corpora of conversations that demonstrate diverse emotional support skills in use. Existing research systems that relate to emotional chatting or empathetic responding return messages that are examples of emotion or empathy and are thus limited in functionality, as they are not capable of many other skills that are often used to provide effective emotional support [1]. People are not naturally good at being supportive, so guidelines have been developed to train humans how to be more supportive.

It is also important to analyze the emotions in the dialogue when constructing an emotional support dialogue system. In recent years, due to the development of deep learning, enhanced learning, and the construction of a large dataset for dialogue, the research on emotion recognition in conversation (ERC) has received attention. Text is the most basic element that composes a series of dialogues. It is also a carrier of emotions that can make human-computer interaction more accurate and intelligent. ERC is emerging research, which aims to detect the emotion expressed in discourse in the dialogue between two or more interlocutors. Due to the increase in the openness of dialogue data and the construction of emotional dictionaries, emotion recognition in the conversation has become popular in the field of natural language processing. Emotion recognition in conversation can help analyze conversations in real-time. The task is important to research in several areas, such as affective dialogue systems [2, 3, 4], healthcare [5, 6] recommendation system [7], and so on.

1.3 Research Contents

For a long time, whether there is an emotion or not is one of the most essential to distinguish between human and machine. In other words, whether the machine has emotion is also one of the key factors for the degree of humanization of the machine [8]. As a comprehensive technology, emotional computing is a key step of artificial intelligence's emotionalization, including emotion recognition, expression, and decision [9, 10]. "Recognition" is to let the machine accurately recognize human emotions and eliminate uncertainty and ambiguity; "Expression" means that artificial intelligence expresses emotions with suitable information carriers, such as language, sound, posture, and expression; "Decision-making" mainly studies how to use emotional mechanism to make better decisions [11].

After the barriers of artificial intelligence and other technologies to the application market and speech recognition technology are broken, natural language generation and emotional computing have gradually become the biggest challenges for dialogue and communication between humans and dialogue systems. Natural language generation needs to understand the needs of user input. For the input sentences, the needs of different domain objects are very different, so different processing strategies are needed to accurately understand the meaning of the input and generate accurate responses. On the premise of generating an accurate response, the analysis and addition of emotion in the dialogue is the key step for the dialogue system to inject the soul. Only the accurate classification of emotions can lay the cornerstone for follow-up work.

This thesis revolves around the above challenges. The main research contents mainly include two parts.

(1) ERC with External Knowledge

Emotion Recognition in Conversation (ERC) task is a specific text classification task, which aims to assign all possible sentiment labels to a given dialogue utterance. As a typical natural language processing task, the performance of the ERC model can be guaranteed due to the disclosure of emotional dialogue data and the increase in availability in recent years. Unlike vanilla emotion recognition of sentences/utterances, ERC ideally requires context modeling of the individual utterances. This context can be attributed to the preceding utterances and relies on the temporal sequence of utterances. Compared to the recently published works on ERC [1, 12, 13], both lexicon-based [14, 15] and modern deep learning-based [16, 17] vanilla emotion recognition approaches fail to work well on ERC datasets as these works ignore the conversation specific factors such as the presence of contextual cues, the temporality in speakers' turns, or speaker-specific information.

This thesis proposes utilizing external knowledge to enhance semantics network architecture that incorporates conversational semantic role labeling Information and the commonsense knowledge feature from external corpora for emotion recognition in conversation. A knowledge enhanced language representation layer based on self-attention has been developed for fusion extraction. Based on the utterance representations rich in external knowledge, the contextual external state, and individual internal state are modeled to predict the emotional label of conversation.

(2) Emotional conversation generation

Scholars who study natural language dialogue generation models usually focus on the semantics of generated sentences, including grammaticality, diversity, and topic relevance, but they lack attention to human emotions [18]. Thanks to the rapid advance of deep learning, neural networks make breakthroughs in speech recognition and machine translation and expand to the breakthroughs in conversation generation. The emotional intelligence of conversation generation significantly defines the ability to perceive, understand, express and control emotions, developing the methods for conversation generation problems such as semantics, grammar, smoothness, etc.

Generally, humans have the unique capacity to perceive and express emotions with language in communications, they have unique characters and could control the specific emotion expression in various situations on their own [19]. Prior works have successfully made the conversational agents obtain emotional features in responses. However, the manners of these agents are controlled by users in a conversation, they are only responsible for generating response in manual input manner, which is not the case for human conversation in practice. Hence, the strategy of equipping the model with the ability to control the specific emotional response generation upon its own characters is still in challenge.

This thesis proposes an emotional dialogue model based on Seq2Seq, which is improved from the four aspects of model input, encoder structure, decoder structure, and search algorithm so that the model generates emotional, diverse, and context-related responses. In terms of model input, emotional information and location information are added based on word vectors. In terms of the encoder, this research first encodes the current input and sentence sentiment to generate semantic vectors, and additionally encodes the context and sentence sentiment to generate context vectors and adds contextual information while ensuring the independence of the current input. In the decoder, attention is used to calculate the weights of the two semantic vectors separately and then decode, and at the same time increase the calculation of the emotional distance on the loss function.

1.4 Thesis Organizations

This thesis mainly investigates the background and research status of emotion recognition in conversation task and emotional dialogue generation task, and proposes some specific methods to solve some existing challenges. The organizational structure of this thesis is as follows:

Chapter 1: Introduction

In this chapter, we talk about the motivation and significance of emotional conversation, and introduce the main research contents and organizational structure of this thesis.

Chapter 2: Background

In this chapter, we introduce the background and related works of ERC task and emotional dialogue generation task. This chapter first introduces existing psychological emotion models. Then the research status and progress technology in recent years is reviewed.

Chapter 3: Utilizing External Knowledge to Enhance Semantics in ERC

In this chapter, we give the definition of the ERC task, and proposed a KES model for this task to fuse external commonsense knowledge and SRL information.

Chapter 4: Conversation Generation with Expressed Emotions

In this chapter, we elaborated on the basic concepts of the dialogue model and constructed an emotional dialogue model to generate emotional dialogue responses.

Chapter 5: Conclusion and Future Works

In this chapter, we conclude the main contents of this thesis and give meaningful directions for future work.

Chapter 2

Background

2.1 Psychological Emotion Models

Emotion recognition has been an active research field for many years and has been explored in interdisciplinary fields such as machine learning, signal processing, social and cognitive psychology, etc [20]. In psychology, emotions are divided into basic emotions and complex emotions according to whether it is difficult to sum up with one word. There are two main types of psychological emotion models that describe how humans perceive and classify emotions: discrete emotion model and dimensional emotion model. Emotions are divided into several basic emotions, which are often relatively independent. Although many studies are devoted to the classification of human emotions, there is no consensus on the definition of basic emotions.

Discrete emotion models have been widely used in emotion recognition task because of simplicity and intuitiveness. Emotions are classified into several basic emotions, which are often relatively independent. Although many studies are devoted to classifying human emotions, there is no consensus on the definition of basic emotions. The discrete emotion model uses adjective labels to express emotion. Simply divides discrete emotions into two basic emotions: pain and happiness [21]. In the categorical front, As shown in Figure 2.1, Plutchik [80]’ s wheel of emotions defines eight discrete primary emotion types, each of which has finer related subtypes. Furthermore, Ekman [22] concludes six basic emotions: anger, disgust, fear, happiness, sadness and surprise. Although there are more than 10 emotion description models used within the field, Ekman’s six basic emotions and seven basic emotions added with neutrality are the most used ones [23]. The discrete emotion model is simple, intuitive and widely used, but its description accuracy is not high, its continuity is not good, and the emotions that can be represented by the model are limited. Scholars established dimensional categorization models to overcome these deficiencies instead.

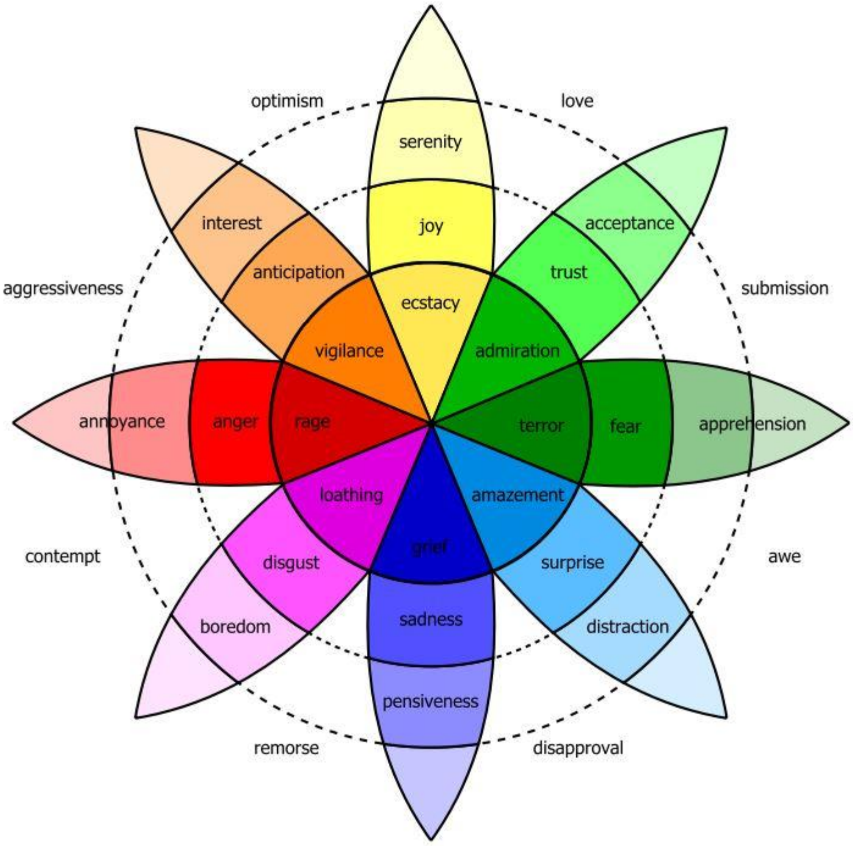

Figure 2.1 Plutchik's wheel of emotion [79].

Dimensional emotion model measures emotion states with numerical dimensions. Each emotion is described as a multi-dimensional vector. In each dimension, the value is continuously changed to distinguish the nuances of emotion, and the extremes of two directions mean two polarities. Most dimensional categorization models [24, 25] adopt two dimensions: valence and arousal. Valence represents the degree of emotion, and arousal represents the intensity of emotion. The commonly used ERC dataset IEMOCAP is compatible with classification and dimension models. However, some later ERC datasets like DailyDialogue can only be used for classification models. Most datasets use simple classifications, similar to variations of Ekman. Each emotional utterance in the EmoContext dataset is labeled as one of the following emotions: sadness, happiness and anger. Annotators of the EmoContext dataset are more consistent compare because its emotion classification is very simple. However, the short length of context and simple emotion classification make ERC task less challenging on this dataset.

2.2 Emotion Recognition in Conversation

ERC is a research topic under spotlight in the field of natural language processing in recent years, which has potential significance in many fields, such as healthcare, opinion mining, education, recommendation system, and so on. With the open-source of numerous conversation datasets with acoustic, textual, and visual features, many deep learning methods have been applied to conversational emotion recognition.

Recognition and expression are two key technical links in emotional computing [26]. Emotion recognition can extract the features of emotion signals and get the emotion feature data that can represent human emotions to the maximum extent. Based on this model, the mapping relationship between the external representation data of emotion and the internal emotional state is found out, and then the current internal emotion types of human beings are identified, including voice emotion recognition, facial expression recognition, and physiological signal emotion recognition.

Different from other emotion recognition tasks, conversational emotion recognition is not only for sentences/utterances, but also depends on the context and the state of participants for modeling. Natural dialogues are complex, and they are governed by many variables, which depend on the time sequence of discourse and affect the emotional dynamics of participants. These variables include topic, argumentation logic, intention, interlocutors' personality, and so on [27]. Figure 2.2 depicts how these factors play out in a conversation. The topic of the conversation and the personality of the participants in the conversation always affect the entire conversation, regardless of time.

The current speaker makes a decision on the response based on the context of the conversation, the previous utterance being the most important one since it usually makes the largest change in the joint task model or the speaker's emotional state. Generally, dialogue utterance features (argumentation logic, interlocutor viewpoint, inter-personal relationship and dependency) are all encoded in the individual state. The intention of the speaker is determined based on the previous intention and the status of the speaker, because the interlocutor may change his intention according to the other party's words and the current situation. The speaker would also formulate appropriate emotions for the response based on individual status and intentions. The final dialogue response is based on the speaker's individual state, intentions and emotions. These variables help express the views and discourse structure of the dialogue, thereby improving the semantic understanding of the dialogue, including the expression and recognition of emotions.

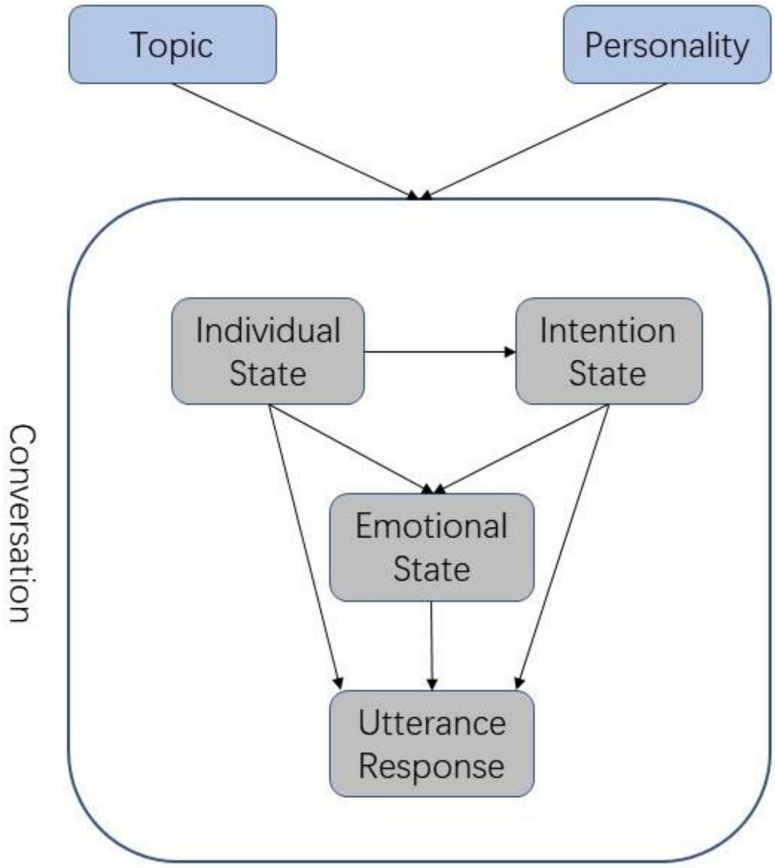

Figure 2.2 The interaction between different variables in the conversation.

Recently, a large number of studies have designed models and algorithms based on deep learning for ERC. Numerous efforts have been devoted to the modeling of conversation context. Basically, they can be divided into two categories: graph-based methods and recurrence-based methods [28]. For graph-based methods, they collect the information of surrounding utterances in a specific window at the same time, while ignoring distant utterances and sequential information. For recurrence-based methods, they consider distant utterances and continuous information by time coding the utterances. However, they tend to use only the relatively limited information from recent utterances to update the status of query utterances, which makes it difficult for them to obtain satisfactory performance.

DialogueRNN [13] mainly solves the problem of emotion recognition in conversation. Propose solutions to the problem of not distinguishing the different parties to the conversation in a meaningful way. For example, it is impossible to determine who is the speaker of a certain utterance in a conversation. DialogueRNN performs individual modeling for each participant of the dialogue, based on the speaker, the context of the current discourse, and the emotion in the context. The model consists of three Gate Recurrent Units (GRU): Global GRU, Party GRU, and Emotion GRU. The Global GRU controls the global state, and updates the current conversation utterance and the individual state of the speaker according to the state of the previous time step. Party GRU models the speakers and listeners participating in the conversation and tracks the individual state of each person. For the individual state of the speaker, the status is updated according to the current conversation utterance, conversation participation state, speaker or listener at the current time step. For the listener, it is only relevant to the conversation when he is speaking. In other words, a person who is silent has no influence on the conversation. Emotion GRU calculates the current emotional state based on the above emotion and the individual state of the speaker.

DialogueRNN adopts an attention mechanism to pool all or part of the dialogue information for each target utterance. However, this pooling mechanism does not consider the information of the utterance and the relative position of the target utterance and other utterances. Inter-speaker dependence or self-dependence helps the model understand the emotional inertia of the individual speaker, that is, the speaker resists the influence of external changes in his emotions. On the other hand, considering the relative position of the target utterance and the contextual utterance determines how the past utterance affects the future utterance. In order to solve this problem, DialogueGCN [29] introduces Graph Involuntary Network (GCN) into ERC task, which models sequential context and speaker-level context through adjacent utterances for each target utterance. DialogueGCN proposes a speaker-level context encoder in the form of a graph network to capture the contextual information related to the speaker in the dialogue. Effective modeling of the speaker level requires capturing the independence between speakers and the speakers themselves. A directed graph is designed in the sequentially encoded utterance to capture the interaction between the participants in the dialogue. In addition, the modified model also proposes a convolution feature transformation process based on local neighbors to create rich speaker-level context encoding features. Each speaker in the conversation would be uniquely affected by each other, so we assume that clearly declaring these relationship edges in the graph would help capture the interdependence and self-dependence between speakers, which would in turn promote the speaker level Context encoding.

Some recent related works have supplemented information by introducing common sense knowledge in the knowledge graph. Zhong et al. [30] propose Knowledge Enriched Transformer (KET), which learns structured conversation representation through layered self-attention and external common knowledge. KET uses Transformer to analyze conversations and detect emotions. Compared with the existing gate RNN and CNN, Transformer's hierarchical self-attention and cross-attention modules enable the model to use contextual information more effectively. KET obtains dynamic situation-aware and emotion-related commonsense knowledge from external knowledge bases and emotion dictionaries to facilitate emotion detection in dialogue. COSMIC [31] proposes a neural network structure framework that introduces external knowledge which introduces external knowledge to improve performance by establishing a huge knowledge base. In a dialogue, the speaker's expression would definitely have an impact on the listener, and these effects have some commonsense information that can help the model to reason. For example, if a person speaks with an angry mood, it would have a negative impact on the person who hears it and may cause the person who hears it to have a negative response.

2.3 Emotional Conversation Generation

2.3.1 Dialogue system

In recent years, deep learning algorithms have provided strong support for the development of dialogue systems. From the perspective of application scenarios, existing dialogue systems can be divided into task-oriented dialogue systems and non-task-oriented dialogue systems. Figure 2.3 shows the framework of the types of dialogue systems.

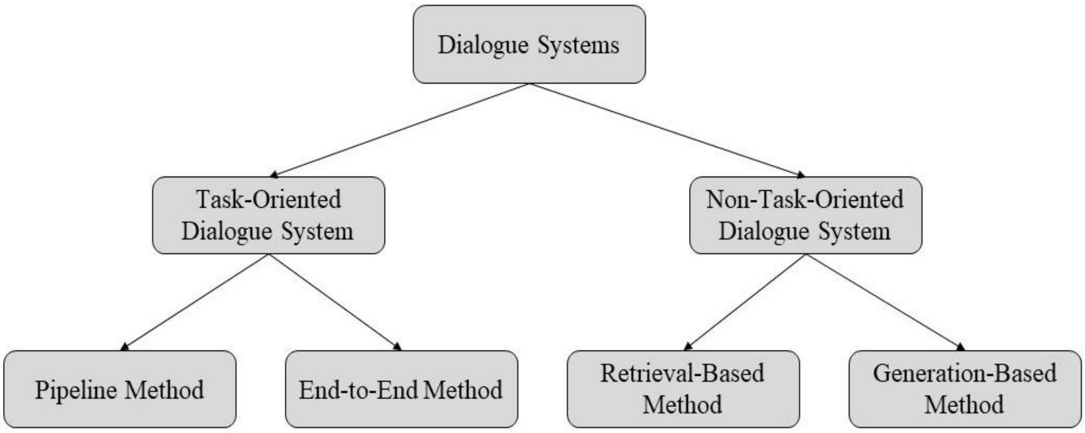

Figure 2.3 Dialogue system classification.

The task-oriented dialogue system is designed to quickly identify the user's appeal and complete the corresponding task in a specific scenario, such as intelligent customer service. The non-task-oriented dialogue system is oriented to the open field, chatting freely with users under the premise of not limiting the subject, and ensuring the consistency and richness of the dialogue content, achieving complete personification. The realization of complete anthropomorphism requires improved interaction with users.

Emotions play an important role in perception and social behavior. Emotional dialogue can empathize with the emotional state of the user, and resonate with the user at the psychological level, thereby enhancing interactivity. The task-oriented dialogue system pursues efficient solution of specific tasks to meet the needs of users. Although whether there is emotion in the dialogue has little influence on the task objectives, if the dialogue system can realize the user's feelings, the whole interaction process can proceed smoothly and happily, and put itself in the user's shoes.

(1) Task-Oriented Dialogue System

The pipeline is a process method, the typical structure of a pipeline-based task-oriented dialogue system is demonstrated in Figure 2.4. It consists of four key components: Natural language understand (NLU), dialogue management (DM) and natural language generation (NLG). Dialogue management can be divided into dialogue state tracking (DST) and dialogue policy learning (DP).

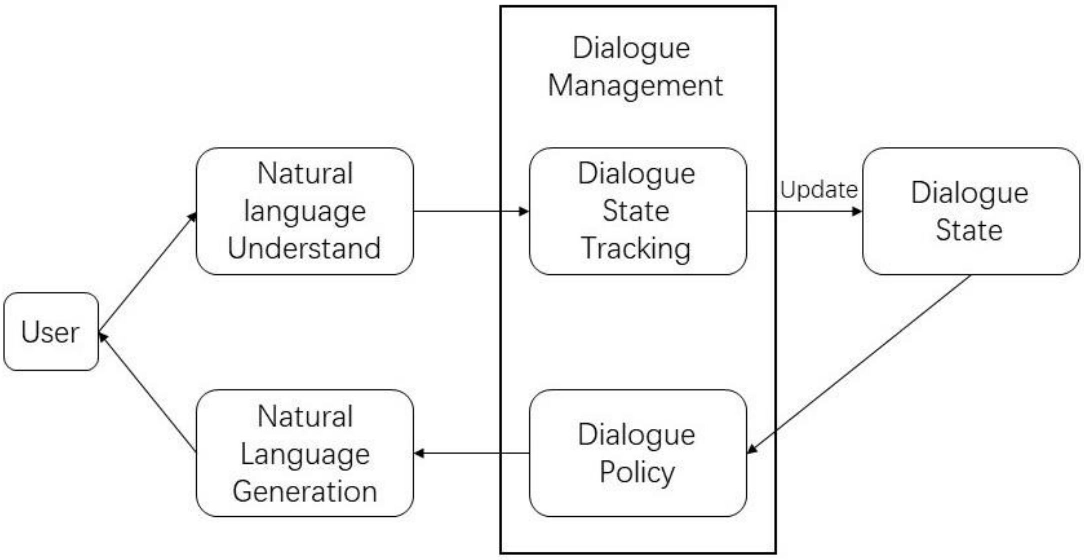

Figure 2.4 Diagram of pipeline method.

Natural language understanding parses the user utterance into predefined semantic slots. It parses the user's intention during the dialogue and forms a state information to pass to the dialogue management. The dialogue management module controls the entire dialogue process, and the dialogue state tracking module uses historical information to fill in the predefined dialogue detail frame. Dialogue state tracker manages the input of each turn along with the dialogue history and updates the current dialogue state. The dialogue strategy learning module determines the optimal strategy for the next action according to the current dialogue state information. Natural language generation maps the selected action to its surface and generates the response. It converts determined strategic behaviors into sentences with reasonable syntax and accurate semantics as output.

Despite a lot of domain-specific handcrafting in traditional task-oriented dialogue systems, which are difficult to adapt to new domains further noted that, the conventional pipeline of task-oriented dialogue systems has two main limitations. One is that the pipeline method needs to perform processing such as feature information extraction in each step, and the information between the previous and subsequent steps is related, which makes the whole process complicated and cumbersome. The input of a component is dependent on the output of another component. When adapting one component to new environment or retrained with new data, all the other components need to be adapted accordingly to ensure a global optimization. Slots and features might change accordingly. This process requires significant human efforts. In order to simplify the dialogue generation process, the above steps can be partially or completely replaced with an end-to-end method.

With the advance of end-to-end neural generative models in recent years, many attempts have been made to construct an end-to-end trainable framework for task-oriented dialogue systems. Instead of the traditional pipeline, the end-to-end model uses a single module and interacts with structured external databases. The dialogue system is trained in a supervised manner. Not only does it require a large amount of training data, but due to the lack of exploration of dialogue control in the training data, it may not be able to find a good strategy robustly.

Task-oriented dialogue systems usually need to query outside knowledge base. The system implements this method by issuing symbolic queries to the knowledge base to retrieve entries based on their attributes. This approach has two drawbacks: (1) the retrieved results do not carry any information about uncertainty in semantic parsing, and (2) the retrieval operation is non-differentiable, and hence the parser and dialog policy are trained separately. This makes online end-to-end learning from user feedback difficult once the system is deployed. The deep learning method can get rid of this cumbersome method. Eric et al. [32] enhances the existing recurrent network architecture and uses a differentiable attention-based key-value retrieval mechanism on the entries of the knowledge base. Combining the retrieval system with deep learning can greatly improve the efficiency and accuracy of retrieval.

(2) Non-task-oriented dialogue systems

Unlike task-oriented dialogue systems, which aim to complete specific tasks for users, non-task-oriented dialogue systems (also known as chatbots) focus on conversing with humans on open domains [33]. According to the different methods of generating dialogue responses, non-task-oriented dialogue systems can be divided into two types: retrieval-based methods and generation-based methods.

Generation-based non-task-oriented dialogue systems aims to learn dialogue patterns from a large-scale corpus. The realization of the generated dialogue system is mainly based on the deep learning model, which is mainly divided into three categories: Sequence-to-sequence model (Seq2Seq), Variational auto-encoder (VAE), and Generative adversarial network (GAN).

In general, chatbots are implemented either by generative methods or retrieval-based methods. Generative models are able to generate more proper responses that could have never appeared in the corpus, while retrieval-based models enjoy the advantage of informative and fluent responses because they select a proper response for the current conversation from a repository with response selection algorithms. However, the use of generative models in non-task-oriented dialogue systems also has certain problems. The basic dialogue model is mainly oriented to the processing of semantic information, and further generates semantically correct and logically fluent responses by extracting the semantic information in the input query. In the process of reply generation, neural networks tend to generate frequently used and strong security utterances, similar to "OK" or "I don't know." The content of these utterances is relatively simple, lacks subjectivity, has no emotional color, and cannot arouse users' interest in the next round of dialogue. The introduction of emotional variables in dialogue generation can enhance the richness and interactivity of content to a certain extent.

Zhou et al. [1] integrated emotional information into the dialogue model based on the generation method and combined the characteristics of emotional decay into the generation process, so that the generated responses have human emotions. Since then, researchers had continued to try to incorporate human emotions into dialogue models. The emotional dialogue model is an extension of the basic dialogue model. The mainstream method is to add an emotional information module to the basic dialogue model, which is mainly used to obtain emotional information under different conditions. In the current emotional dialogue system, there are two main methods for determining the response emotion category: one is to directly specify the response emotion category, and the other is to not specify the response emotion category. The method of specifying the response emotion category requires the emotion information module to convert the discrete or continuous emotion category into an emotion feature vector while considering the change of the emotional state. The method that does not specify the response emotion category thinks that the response emotion information is contained in the context, and the emotion information module needs to effectively extract the emotion information from the context and chat topics. The dialogue system that incorporates emotional information can more arouse the user's active interaction, and the ability of the dialogue system to have emotional cognition is more helpful to promote communication between humans and computers.

2.3.2 Emotional dialogue system

The primary goal of building a dialogue system is to address users' questions and concerns via emulating the way humans communicate with each other. As human language is too complicated to be considered as a single target, dialogue systems have to model different aspects of human communication separately. Emotion plays an important part in the communication of human beings as it has the potential for enhancing their emotional bond. Integrating empathy into the design of dialogue systems is also crucial to improving the user experience in human-computer interaction. More importantly, empathy is a necessary step for dialogue agents to be regarded as social roles by users. Building an empathetic dialogue system is then premised on the idea that it would result in improved user engagement and, consequently, more effective communication. Research on dialogue system has elaborated on the concept of dialogue system mainly from the perspective of features.

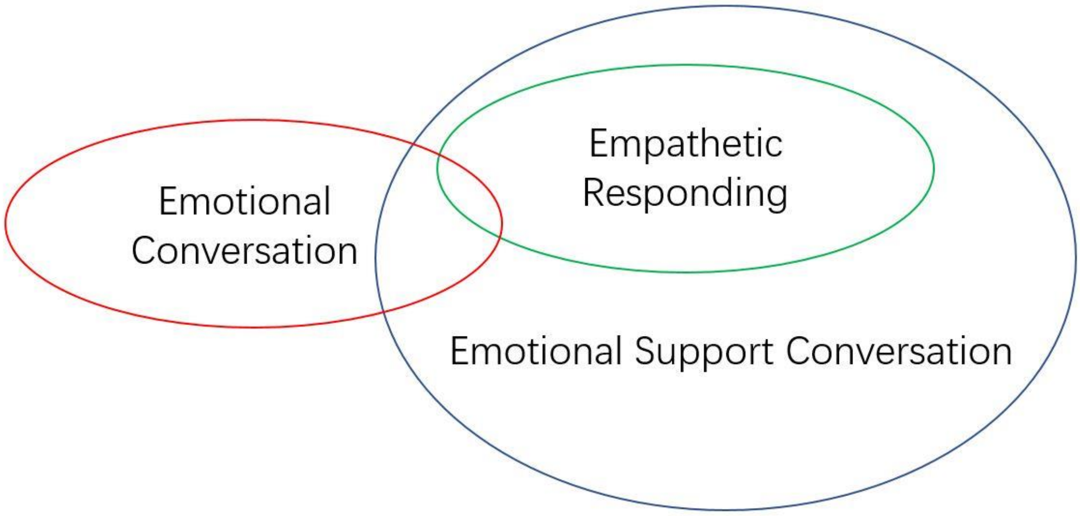

Figure 2.5 Illustrates the relationship between the three tasks.

Figure 2.5 intuitively shows emotional support conversations can include elements of emotional chatting and empathetic responding. Emotional support (ES) aims at reducing individuals' emotional distress and helping them understand and work through the challenges that they face [34]. It is a key ability to embed emotions when the dialogue system interacts with users, particularly for settings that include social interactions, mental health support, and customer service chats. In order to find out the cause of the suffering of the help-seekers, the supporters first explore the problem of the help-seekers. Without exploration, supporters are unlikely to understand the experience and feelings of helpers, so if supporters give irrelevant advice, it may be offensive or even harmful. Supporters can express understanding and empathy when they understand the situation of the help-seeker and use various techniques to alleviate the frustration of the help-seeker. After understanding the help-seekers problem, the supporter may offer suggestions to help the help-seeker cope with the problem. If the supporter-only comforts the help-seeker without any inspiration for action to change, the supporter may not effectively help the help-seekers emotions improve. Although emotional support is very important and complicated, the research on data-driven emotional support dialogue system is limited due to the lack of task design and related dialogue corpus.

Although emotional support may include expressing emotions, such as happiness or sadness, it has a broader goal of reducing users' emotional distress by using appropriate support skills, which is fundamentally different from emotional conversation. The emotional conversation is only the basic quality of the dialogue system, while emotional support is the higher level and more complex ability that the dialogue system should possess. Compared with emotional support, empathy response aims to understand the user's feelings and then respond accordingly. Effective emotional support naturally requires empathy based on the experience and feelings of the seeker, and empathy is only one of the necessary components of emotional support. In addition to empathetic responding, emotional support conversation also needs to explore users' problems and help them cope with difficulties.

In daily dialogue and communication, both parties in the conversation have rich emotions. Therefore, in order to better communicate with users emotionally, a dialogue system with emotional cognitive ability needs to first recognize and judge the user's emotions in the dialogue process, and then generate emotional responses. Therefore, analyzing and expressing emotions are the two major tasks of realizing an emotional dialogue system. The emotional dialogue system can be further divided into two categories, namely, dialogue emotional perception and emotional dialogue generation. The user's emotional state is modeled based on historical dialogue information, and then the emotional state is embedded in the dialogue generation model to improve the algorithm to generate a more natural response with emotion.

Emotional dialogue generation is essentially a type of generative task that generates semantically relevant and emotional responses. Early sentimental responses used rule matching algorithms. The rules were formulated by humans, and the corresponding matching algorithms were used for matching to obtain the final response output. The existing research work on emotional dialogue generation is mostly driven by dialogue data, and the generation of dialogue is realized based on a generative dialogue model.

(1) Rule-based Matching

Early emotional response generation is often based on rule matching templates. Researchers formulate different templates and rules according to different dialogue scenarios in advance, and then use these templates and rules to learn dialogue response strategies.

Chatbot Eliza [35] came out in 1966. This is the world's first chatbot, used to deal with mental illness and other problems. Researchers at Stanford University invented the chat robot Parry [36], which can simulate human emotions and communicate with humans. The replies of these two chatbots rely on hand-made rule bases and do not really understand the semantic information in the dialogue. This makes the generated replies have many problems, and often the reply sentences have semantic discontinuities and unreasonable problems. To solve this universal problem, Keshtkar et al. [37] generate sentences containing emotions according to the user's selection mode and sentence planner. Skowron et al. [38] put forward the idea of emotional monitoring in the dialogue, responding from two levels of content and emotion.

(2) Response generation giving specific emotion

The response generation that generates a specific emotion is to generate a corresponding reply sentence according to the preset emotion category. Emotions are the external manifestations of emotions, so the emotions are first transformed into emotional information, and then the emotions are quantified and then integrated into the model. Emotions can be further divided into discrete emotion models and dimensional emotion models. The discrete emotion model uses labels to represent emotion categories, and the dimensional emotion model maps emotions to a multi-dimensional space and uses continuous values to describe emotions. According to the different types of response emotions, the generation of designated response emotions can be further divided into two categories. The emotional chatting machine (ECM) proposed by Zhou et al. [1] is the first attempt to incorporate discrete emotional information into the generative dialogue model. Discrete emotions are divided into six categories: like, happy, sad, depressed, angry, and others. They introduced three mechanisms on the basis of the Encoder-Decoder model structure: static emotional embedding, dynamic emotional memory and emotional external emotion memory. The model is data-driven, does not rely on any language tools and customized parameters, and can model a variety of emotional interactions between the input post and response. Since then, many studies have made different improvements to the ECM model. For example, Shen et al. [39] believe that the model also needs to consider the issue of emotional consistency when generating responses, rather than just responding based on given emotions. Xie et al. [40] used linguistic inquiry and word count emotional dictionary to process discrete emotions, transformed them into high-dimensional continuous variables through spatial mapping, and then embedded them in the decoder.

(3) Unspecified emotional response generation

It is considered that in the above context, the emotional information required for the response is already contained in the generation of the unspecified reply to emotion, and it is no longer necessary to manually specify the emotional category of the reply. The emotion contained in the sentences generated by this model is flexible and controllable.

Nabiha et al. [41] proposed an emotional dialogue model for English dialogue. This model is different from the ECM model, which needs to specify the response emotional category. They believe that the contextual information contains emotional information and there is no need to specify the emotional category in advance. The model is based on the Seq2Seq model and combines three types of methods for adding emotional variables: emotional word embedding representation, modified loss function, and emotion-based directional search methods. Monomodal text information cannot well capture the user's emotional characteristics. Shi et al. [42] incorporated multi-modal information into Seq2Seq. The model regards the learning of dialogue strategies as a classification task and uses reinforcement learning methods to use user emotional information as real-time rewards. Liang et al. [43] believe that historical emotion information, user facial expressions, speaker personality, etc. are all heterogeneous information nodes, which can be represented by heterogeneous graphs, and then the information is extracted through the heterogeneous graph neural network (HetGNN) and finally integrated into the generation stage. Different from embedding emotional information in the generation stage, Ghosh et al. [18] proposed an emotional language model based on LSTM. The model incorporates emotion labels and emotion strength in the word probability prediction stage, thereby generating reply to sentences with emotion categories and emotion strength.

2.4 Pre-trained Language Model

Word embedding is a type of word representation, and words with similar meanings have similar representations. It is a general term for the method of mapping words to real number vectors. It refers to embedding a high-dimensional space whose dimension is the number of all words into a continuous vector space with a much lower dimension, and each word or phrase is mapped to a vector on the real number field.

With the development of deep learning, the number of model parameters has increased rapidly. The much larger dataset is needed to fully train model parameters and prevent overfitting. However, building large-scale labeled datasets is a great challenge for most NLP tasks due to the extremely expensive annotation costs, especially for syntax and semantically related tasks. In contrast, large-scale unlabeled corpora are relatively easy to construct. To leverage the huge unlabeled text data, we can first learn a good representation from them and then use these representations for other tasks.

One-hot Representation is the most intuitive method in natural language processing. It expresses each vocabulary as a vector whose dimension is the size of the vocabulary. The dimensional vector of the current word is 1 and the remaining positions are 0. One-hot means that it is stored in a sparse manner, which is concise and intuitive, but it is easy to cause dimensional disasters, and any two words are isolated between them, which cannot express the relevance of the semantic level.

Distributed representation describes the scattered embedding of text into another space, generally from high-dimensional space to low-dimensional space. Due to different modeling, it can be divided into three categories: "representation based on matrix distribution", "representation based on clustering distribution" and "representation based on neural network distribution". Word embedding is a process of natural language processing. The by-product of the neural language probability model reflects the semantic information between words by mapping specific words into digital vectors. Word embedding converts the vector component represented by the one-hot form from an integral type to a floating-point type and compresses the previously sparse huge dimension into a smaller dimensional space. Each vocabulary is a vector in the dimension space, and the semantic correlation between vocabularies is obtained by judging the Euclidean distance between the vectors.

Representing words as dense vectors has a long history. word embedding is introduced in pioneer work of neural network language model (NNLM) [44]. Collobert et al. [45] showed that the pre-trained word embedding on the unlabelled data could significantly improve many NLP tasks. To address the computational complexity, they learned word embeddings with pairwise ranking task instead of language modeling. Their work is the first attempt to obtain generic word embeddings useful for other tasks from unlabeled data. Mikolov et al. [46] showed that there is no need for deep neural networks to build good word embeddings. They propose two shallow architectures: Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) models. Despite their simplicity, they can still learn high-quality word embeddings to capture the latent syntactic and semantic similarities among words. Word2vec is one of the most popular implementations of these models and makes the pre-trained word embeddings accessible for different tasks in NLP. Besides, GloVe [47] is also a widely used model for obtaining pre-trained word embeddings, which are computed by global word-word cooccurrence statistics from a large corpus. Although pre-trained word embeddings have been shown effective in NLP tasks, they are context-independent and mostly trained by shallow models. When used on a downstream task, the rest of the whole model still needs to be learned from scratch.

Recently, substantial work has shown that pre-trained models, on the large corpus can learn universal language representations, which are beneficial for downstream NLP tasks and can avoid training a new model from scratch. With the development of computational power, the emergence of the deep models such as Transformer [48], and the constant enhancement of training skills, the architecture of pre-trained models has been advanced from shallow to deep. The first-generation pre-trained models aim to learn good word embeddings. Since these models themselves are no longer needed by downstream tasks, they are usually very shallow for computational efficiencies, such as Skip-Gram [46] and GloVe. Although these pre-trained embeddings can capture semantic meanings of words, they are context-free and fail to capture higher-level concepts in context, such as polysemous disambiguation, syntactic structures, semantic roles, anaphora. The second-generation pre-trained models focus on learning contextual word embeddings, such as CoVe [49], ELMo [50], OpenAI GPT [51] and BERT [52]. These learned encoders are still needed to represent words in context by downstream tasks.

Besides, various pre-training tasks are also proposed to learn pre-trained models for different purposes.

2.5 Knowledge Enhanced Language

Sentiment analysis methods based on machine learning and deep learning often encounter insufficient labeled data and poor generalization ability in the actual application process. In order to make up for this shortcoming, researchers try to introduce external emotional knowledge to provide supervision signals for the model and improve the analysis performance of the model. Despite the great success of pre-trained models, existing pre-training tasks like the masked language model and next sentence prediction neglect to consider the linguistic knowledge. Such knowledge is important for some NLP tasks, particularly for sentiment analysis. Prior knowledge is usually incorporated into deep neural networks as auxiliary information for deeper language understanding, including emotional dictionary resources, commonsense Knowledge, language patterns, emotional semantic rules, and any other emotional-related knowledge. Combining prior knowledge helps to enhance emotional feature representation and achieve more accurate emotion recognition.

General text classification tasks only provide sentence or document-level sentiment labels. The introduction of sentiment dictionary and other prior sentiment knowledge can introduce more fine-grained supervision signals to sentiment text, so that the model can learn feature representations that are more suitable for sentiment analysis tasks. The underlying part of speech and syntax analysis tasks can provide reference information for downstream sentiment classification and extraction tasks. For example, evaluation expressions are usually adjectives or adjective phrases, and evaluation objects are usually nouns. Different sentiment analysis tasks themselves have mutual promotion effects. For example, the distance between the evaluation object and the evaluation word in the sentence is usually relatively close, and joint extraction can improve the performance of both at the same time. Short text comments usually omit a large amount of background knowledge, and it is usually difficult to infer the true sentimentality from the text itself.

Researchers have paid more attention to enhancing natural language models with knowledge graphs these days, since knowledge graph has gained a lot of systematic knowledge. Liu et al. [53] combined the knowledge triple in the knowledge graph with the original text and then modeled it with BERT to get more hidden information. Lin et al. [54] proposed a knowledge-aware graph network model based on a graph convolution network, which has a path-based attention mechanism. Zhang et al. [55] combined entity information with BERT to enhance language representation, which can utilize vocabulary, syntax and knowledge information at the same time. More recently, Bosselut et al. [56] proposed COMmonsEnse Transformers (COMET), which learned to generate commonsense descriptions in natural language by fine-tuning the pre-trained language model in ATOMIC knowledge base. Compared with the extraction method, the fine-tuned language model in the knowledge base has unique advantages of generating knowledge for invisible events, these advantages are very important for tasks that need to combine common knowledge in conversation. Ghosal et al. [31] proposed COmmonSense knowledge for eMotion Identification in Conversations (COSMIC) model based on DialogueRNN structure, which uses external commonsense knowledge generated by COMET and obtains advanced results in ERC. Compared with the extraction method, the fine-tuned language model in the knowledge base has the unique advantage of generating knowledge for invisible events, which is very important for tasks that need to combine common knowledge such as emotion detection in conversation.

2.6 Conversational Semantic Role Labeling

Semantic role labeling (SRL) is one of techniques for shallow semantic parsing of natural language texts that produces predicate-argument structures of sentences. Predicates bear the central meaning of a situation expressed by a text. In most semantic theories, predicates are verbs, verbal nouns, and some other verb forms. Arguments are phrases that fill meaning slots of a situation expressed by a predicate and define its essential details. Semantic role labeling systems aim to recover the predicate-argument structures of sentences – basically to determine “who did what to whom?”, “when?”, and “where?”, etc. It is said that arguments play semantic roles in a situation as roles define meanings of slots. Role meanings and sizes of role inventories vary in different semantic theories and annotated corpora. Converting a text into such shallow semantic structures helps to abstract from syntactic and morphological representations of sentences and is considered to be an important technique for natural language understanding [57].

The whole SRL process can be divided in four steps: predicate identification and identification of its frame (disambiguation), argument extraction (for each predicate), argument classification (or labeling of arguments with semantic roles), and global scoring that deals with linguistic constrains. Predicate-argument structures in some notations can be represented as two-level trees, rooted in predicates, with single tokens (nouns, adjectives, pronouns, proper names) as leaves that denote arguments. We adopt this dependency-based notation and treat the problem of semantic role labeling as constructing such trees.

Traditional SRL often failed analyze conversations since only single utterance can be analyzed by traditional SRL whereas ellipsis and anaphora occur in conversation as well. Conversation semantic role labeling (CSRL) [58] directly models the predicate-argument structure of the whole conversation instead of a single utterance. Most of the discarded or referenced components in the latest conversation can actually be found in the conversation history. CSRL allows arguments to be in different utterances as the predicate, while SRL can only work on each single utterance. Compared with the standard SRL, which needs utterance rewriting or co-referential parsing as the preprocessing step of analyzing dialogues, CSRL can directly deal with conversation and avoid error propagation.

Chapter 3

Utilizing External Knowledge to Enhance Semantics in ERC

3.1 Introduction

In recent years, due to the development of deep learning, enhanced learning, and the construction of a large dataset for dialogue, the research on emotion recognition in conversation (ERC) has received attention. ERC is emerging research, which aims to detect the emotion expressed in discourse in the dialogue between two or more interlocutors. The task is important to research in several areas, such as affective dialogue systems [2, 3, 4], healthcare [5, 6] recommendation systems [7], and so on.

Recent works on ERC use recurrent neural networks (RNNs) to model the utterance [13], [59], which relies on spreading context and order information to utterance. RNNs, such as long short-term memory (LSTM) [60] and gated recurrent unit (GRU) [61], are used to simulate the dependence between utterances, so as to perceive the context at speaker level or situation level. In theory, such a network should be able to transmit long-term contextual information, but in practice, it is not always the case. To mitigate this issue, graph-based methods are introduced [29], which consider distant utterances and sequence information by time coding utterances. However, graph-based methods tend to use information from recent utterances that are relatively limited to update the status of query utterances, which makes it difficult for them to obtain satisfactory performance.

Although the method of the above genres has made well progress, we argue that current techniques which only focus on context modeling at the present limit the ability of language representation. The main limitation of these methods is that they only consider simple contextual features as representation and training objectives, and seldom consider well-defined contextual semantic clues. The common knowledge of conversation plays an indelible role in inferring the potential variables of a conversation. Even though a well-trained language model can express context semantics more or less implicitly, the introduction of a common sense-oriented framework can further enhance this point [31].

Most studies have found that deep learning Frameworks might not really understand the semantics of the natural language [62] and vulnerably suffer from adversarial attacks [63]. Deep learning models often ignore important words and choose relatively safe and unimportant ones. Briefly, semantic role labeling (SRL) [64] aims to restore the predicate-argument structure of sentences, and fundamentally discover change from who did what to whom, when and why did what as the central meaning of sentences, which naturally matches the task goal target ERC. In the conversation task, the conversation content of the participants usually involves various predicate-argument structures. A predicate is a statement or explanation of the subject, pointing out ‘‘what to do’’ or ‘‘how to do,’’ which represents the core of an event, and the question formed with who, what, how, when, and why can be conveniently formalized into the predicate-argument relationship in terms of contextual semantics. Motivated by these, this chapter attempts to integrate extra SRL knowledge into the pre-trained model, and model the context and the emotional state of the participants to understand the conversation context.

In this chapter, we propose a KES for emotion recognition in conversation. By introducing SRL information and clear context semantic clues into the pre-training language model, the algorithm enriches the sentence context semantics in multiple predicate-specific argument sequences. The proposed model incorporates the concept of external knowledge and SRL information, at the same time, it learns representation in a fine-grained manner on plain context representation and explicit semantics in order to achieve deeper meaning representation. Through an individual internal state encoder, our model tracks and predicts the speaker's continuous emotional self-dependence. The information reflecting the contextual state of context and the speaker's influence and dependence on others is encoded and processed by the global state, so that the proposed model can understand the contextual information and the emotional transfer among the participants in the conversation. We conduct extensive experiments on three different conversation corpus and comparisons with several baseline models. The results show that the proposed method achieves comparable performance with the state-of-the-art models.

To sum up, the main contributions in this research are summarized as below:

(1) We introduce SRL information to enrich the semantic structure in conversation, and obtain commonsense knowledge from external knowledge graph to promote emotion detection in conversation.

(2) We design the attention mechanism to integrate external commonsense knowledge and conversation level SRL information, and utilize a transformer structure to replace the recurrent attention neural networks commonly used for emotion detection in conversation.

(3) We conduct extensive experiments, which prove that SRL information and commonsense knowledge are beneficial to the performance of emotion detection. The proposed model KES is superior to the state-of-the-art models in most test datasets.

The rest of the chapter is organized as follows: Section 3.2 defines the task objectives. Section 3.3 shows an overview of the proposed method. Section 3.4 introduces the relevant situation of the experiment. Section 3.5 provides experimental results and analysis of Section 3.4. Section 3.6 presents our conclusions and future work.

3.2 Task Definition

ERC task aims to recognize the emotion of each utterance from several predefined emotions within/among the provided conversation records and participant information of each utterance in a conversation emotion recognition task. In a conversation between two people, each utterance is marked by latent emotion. Formally, let there be M participant/parties $\{p_1, p_2, \ldots, p_M\}$ in a conversation, which is defined as a sequence of utterances $\{u_t = u_1, u_2, \ldots, u_N\}$, where N is the number of utterances. The task is to predict the emotion labels (happy, sad, frustrated, excited, angry, and neutral) of the constituent utterances u_t, where utterance u_t is uttered by participant $p_{s(u_t)}$, where s representing the mapping between utterance and index of its corresponding participant. In related research, the traditional method is to first produce context-independent representations by pre-trained language model and then perform context modeling to classify each of the constituting utterances into its appropriate emotion category.

3.3 Methodology

This section describes in detail the overall methods of external knowledge acquisition and neural networks. Commonsense knowledge is extracted from the external knowledge graph, which contains the relationship between different entities. In order to better combine SRL and commonsense knowledge, a transformer-based neural network is proposed to integrate deep features.

3.3.1 Utterance Feature Extraction

He et al. [64] presented a deep highway BiLSTM architecture with constrained decoding, which is simple and effective. In the practice of data preprocessing, each utterance is annotated into several semantic sequences by the pre-trained semantic annotator.

The original utterance sequence and semantic role tag sequences are expressed as embedding vectors to feed a pre-trained BERT. The input utterance is a sequence of words of length n, which is first labeled as a word fragment. Then, the transformer captures the context information of each token through self-attention and generates a sequence of context embeddings. For m SRL label sequences associated with each predicate, we have $\{t_1, t_2, \dots, t_m\}$ where t_i contains n labels denoted as $\{\mathrm{label}_1^i, \mathrm{label}_2^i \dots, \mathrm{label}_n^i\}$. We employ SemBert [65] to extract semantically enhanced language representation by SRL semantic sequences, and fine-tune SemBert large-scale emotional label prediction model from the transcript of the utterances. It can be trained jointly with KES, so its gradient would be updated in the whole building training process. It can also be trained as an individual task of discourse classification with emotional labels. However, related experiments show that although the traditional SRL system (even with the help of common reference parsing or rewriting) does not perform well in analyzing conversation, modeling the conversation history and participants is of great help to the performance, which indicates that adapting SRL to conversations is very promising for general conversation understanding. Therefore, we introduce the concept of external knowledge to try to further extract the missing features in the conversation.

3.3.2 Fusion of Semantic Role Labeling

In SemBert, the words in the input sequence are passed to the semantic role labeling to obtain multiple predicate derivation structures of explicit semantics, and the corresponding embeddings are aggregated after the linear layer to form the final semantic embedding. At the same time, the input sequence is segmented into sub-words by the BERT word segment tagger, and then the sub-word representation is converted back to the word level through the convolutional layer to obtain the contextual word representation. Finally, the word representation and semantic embedding are connected to form a joint representation of downstream tasks.

We mark the input sentence of length n as word fragments. Then, the transformer encoder captures the contextual information of each tag through self-attention, and generates a series of contextual embeddings. We regard the semantic signals as embeddings and use a lookup table to map these labels to vectors $\{v_1^i, v_2^i, \dots, v_n^i\}$ and feed a BiGRU layer to obtain the label representations for m label sequences in latent space:

$$h(t_i) = BiGRU(v_1^i, v_2^i, \dots, v_n^i) \tag{3.1}$$

where $0 < i \leq m$. For m label sequences, let L_i denote the label sequences for token u_i, we have $h(L_i) = \{h(t_1), h(t_2), \dots, h(t_m)\}$. We concatenate the m sequences of label representation and feed them to a fully connected layer to obtain the refined joint representa-tion h_n:

$$h(L_i) = \boldsymbol{W}_1[h(t_1), h(t_2), \dots, h(t_m)] + \boldsymbol{b}_1 \tag{3.2}$$

$$h_n = \{h(L_1), h(L_2), \dots, h(L_n)\} \tag{3.3}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ are trainable parameters.

This integrated module combines lexical text embedding and label representation. Since the original pre-training, Bert is based on a series of subwords, and the semantic role labeling we introduce is on words, we need to align these sequences of different sizes. Therefore, we group the sub-words of each word and use a convolutional neural network (CNN) with maximum pooling to obtain a word-level representation. CNN is faster than recurrent neural network (RNN), and CNN is more conducive to sub-word derivation modeling tasks in capturing local features.

3.3.3 Knowledge Feature Extraction

In this work, we employ an external knowledge graph ATOMIC [66] to extract the knowledge sources. ATOMIC is a collection of if-then common knowledge that describes the daily reasoning of an organization through text. It is composed of nine different types of if-then relationships to distinguish between agents and themes, causes and effects, voluntary and non-voluntary events, and actions and mental states. Due to the expressiveness of events and the improved relationship type, ATOMIC is used in the If-Then reasoning task to achieve the result competing with human evaluation. Items with weights less than the threshold or containing words that are not in the selected vocabulary would be removed from the knowledge graphs. Items are triples with the form {subject, relation, object}. Given an event in which the speaker participates, the 9 relation types are inferred as follows: intent of speaker, need of speaker, attribute of speaker, effect on speaker, wanted by speaker, reaction of speaker, effect on others, wanted by others and reaction of others. For example, given an event or topic phrase: ‘‘PersonX puts PersonY in touch’’ from ATOMIC’s inference of relation phrases, PersonX’s intention and reaction of others would be ‘‘PersonX want to keep the relationship’’ and ‘‘others want to express gratitude’’ respectively. There are a total of 9 relation types, of which four are used: the intent of speaker (denoted as XI), effect on speaker (denoted as XE), the reaction of speaker (denoted as XR), and reaction of others (YR).

Given an utterance u_t, we can compare it with each node in the knowledge graph and retrieve the most similar one. Each utterance u_t is annotated with a part-of-speech (POS) tag by NLTK [67]. Usually, nouns, adjectives, and verbs with parts of speech contain more information than other tokens. Therefore, the items related to them are searched preferentially in the knowledge map. In all the chosen items, we extract the top K events, and obtain their intentions and reactions. We employ BERT [52] calculation to capture the causes between two sequences, and the last hidden state is taken as the output, which is denoted as $c_t = \{c_{nk}^{XI}, c_{nk}^{XE}, c_{nk}^{XR}, c_{nk}^{YR}\},\ k = 1,\ 2\ \ldots,\ K$.

3.3.4 Model

It is crucial to consider contextual information when classifying discourse in a sequence, since other discourses in the sequence have a great influence on the emotion of the current discourse. In other words, the dependence between speakers is important for the emotional dynamics in a conversation. For example, the current speaker's emotions can be changed by the other's words, and it is crucial to consider context information for simulating the emotional dynamics in a conversation.

The conversation is a sequence of coherent and orderly discourses. For neural networks, the capture of long-range context information is a weakness. We adopt Transformer [48] a structure composed of self-attention and feed forward neural network, instead of the traditional RNN model, aiming at capturing remote context information.

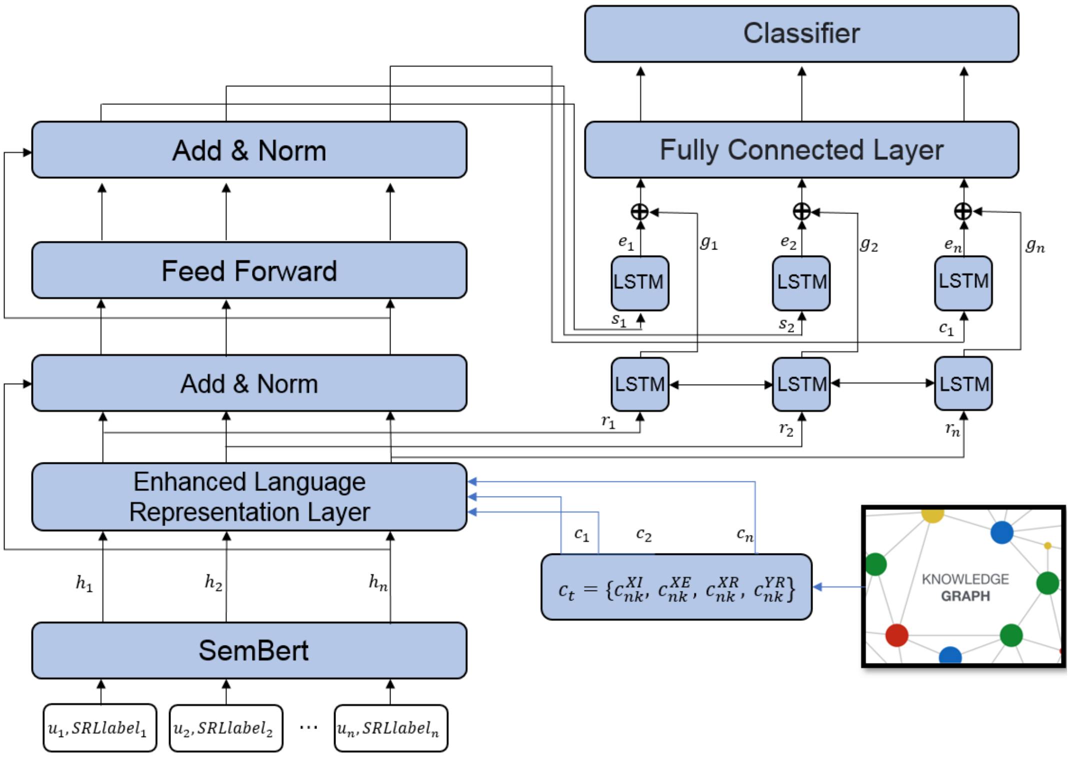

Figure 3.1 Overview of KES.

Enhanced Language Representation Layer takes the textual feature u_t of the tth utterance and its SRL labels in conversation as input and generates code the enhanced semantic feature h_t, which is input to the transformer structure network. We design the first layer self-attention mechanism of the transformer structure to integrate the candidate knowledge, h_t and external commonsense knowledge c_t as the input of this attention mechanism to generate r_t. The transformer structure network generates encoding s_t and inputs it into the individual internal encoder LSTM to generate e_t. Also, rt is fed into the individual context encoder BiLSTM to outputs another encoding g_t. Each e_t and g_t to obtain the final prediction for each utterance in the conversation.

3.3.4.1 Knowledge Enhanced Language Representation Layer

With the knowledge source extracted, the commonsense features from the knowledge graph are obtained. We design an attention mechanism to integrate the candidate knowledge in Transformer. We modified the structure of the transformer so that it separately encodes the internal state information of the individual in the conversation and the contextual global state information. The Attention mechanism in the encoder is used to fuse different common sense and knowledge information, and integrate and filter effective information into the connected layer. A conversation consists of N utterances $\{u_1, u_2, \dots, u_n\}$, in which M distinct participants $\{p_1, p_2, \dots, p_m\}$ take part. For every $t \in \{1,2 \dots, N\}$, we apply the enhanced semantic features ht obtained by fusing Bert feature vector and SRL information to generate new knowledge representation. The attention mechanism is used to refine the result of each knowledge source and aggregate the representation of each feature as follows:

$$v_t = \tanh([c_t, h_t]\mathbf{W}_\alpha) \tag{3.4}$$

$$\alpha_t = \frac{\exp\,(v_t v_t^T)}{\sum_{k=1}^{K} \exp\,(v_t v_t^T)} \tag{3.5}$$

$$r_t = \sum_{k=1}^{K} \alpha_t c_t \tag{3.6}$$

in which $\mathbf{W}_{\alpha}$, indicate the model parameters.

We further integrate the knowledge feature through the self-attention mechanism and the final event representation is denoted r_t.

The context encoder takes enhanced textual features and knowledge features of utterances as input and applies multi-head self-attention attention operation to it by a feed-forward layer which is completely concatenated point by point, so as to generate the contextualized cultural vectors of utterances:

$$s_t = FFN(r_t) \tag{3.7}$$

3.3.4.2 Context State

It is essential to consider the contextual information when classifying the emotions of a conversation discourse in a sequence since other discourse information in the sequence has a great influence on the emotions of the current discourse, and the contextual state stores and transmits the information of the whole utterance-level along the sequence of the conversation flow. In other words, the current speaker's mood would be forced to change by the other's words. This fact reflects the dependence between speakers, which is closely related to the tendency of speakers to imitate each other in conversation [54] and is important for simulating the emotional dynamics in a conversation. Our work is not only to extract the emotional influence information between speakers from the knowledge graph but also to encode the context state of the current discourse to obtain the dynamic emotional changes [68]. In view of the sequence of conversation, we use BiLSTM to capture the contextualized cultural vectors. The upper and lower cultural vectors generated by the Transformer structure are fed to the BiLSTM layer, and the BiLSTM layer fuses remote sequential contextual information to generate context coding.

Finally, the contextualized feature representation is input into BiLSTM, and context feature is obtained:

$$g_t = \mathrm{BiLSTM}_t(r_t) \tag{3.8}$$

3.3.4.3 Individual Internal State

The individual internal state tracks each utterance in the conversation, which reflects the speaker’s emotional influence on himself in the conversation. The individual internal state of participants depends on their feelings and the effects they feel from other participants. Participants may not always express their feelings or opinions clearly through external positions or reactions. This state can also be considered to include aspects that participants actively try not to express or features that are considered common sense and do not need explicit communication. Under the influence of emotional inertia, every speaker in the conversation tends to keep a stable emotional state until the other person causes changes.

We model the individual internal state of the participants using LSTM, which is the internal encoder to output all speaker states for timestep t. It exploits the currently integrated knowledge discourse representation to update the state of the corresponding speaker:

$$e_t = \mathrm{LSTM}_t(s_t) \tag{3.9}$$

In time-step t, the output of each LSTM corresponds to the speaker and is updated by the knowledge discourse representation r_t of the current utterance u_t.

3.3.4.3 Emotion Classification

Finally, we connect the global feature vectors generated by the context state and the internal feature vectors generated by the individual internal state, calculate the probabilities of six emotion-class and select the most possible emotion class.

$$P_t = \text{softmax}(W_{smax}(g_t \oplus e_t) + b_{smax}) \tag{3.10}$$

$$\hat{y}_t = \underset{i}{\text{argmax}}\, P_t\,[i] \tag{3.11}$$

3.4 Experiments Setup

For ease of comparison with state-of-the-art methods, we evaluate our model on three benchmark datasets: IEMOCAP [69], DailyDialogue [70], and MELD [71], and mention their properties. Further, our report summarizes the experimental results of conversational emotion recognition from the text information of all three benchmark data sets.

3.4.1 Datasets

Information about the datasets is shown in Table 3.1.

Table 3.1 The Statistics of Three Datasets.

Dataset	# Conversations			# Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
DailyDialogue	11118	1000	1000	87170	8069	7740
MELD	1038	114	280	9989	1109	2610

IEMOCAP is a multimodal ERC dataset, which contains videos of two-way conversations of ten unique speakers. The trainset conversations come from the first eight speakers, whereas the testset conversations are from the last two. Each video in IEMOCAP contains a single dyadic conversation from the performance based on a script by two actors. Each discourse has 2476 annotations, with one of the following six emotions: happiness, sadness, neutrality, anger, excitement and depression.

DailyDialogue is a human-written dyadic conversation dataset from daily communications. DailyDialogue takes the Ekman’s six emotion types [22] as the annotation protocol and reflects daily communication way and covers various topics about human daily life. The emotion can belong to one of the following seven labels: anger, disgust, fear, joy, neutral, sadness, and surprise. The dataset contains more than 83% neutral emotion labels, which were excluded during the evaluation of Micro F1.

MELD is a multi-modal ERC dataset extended from Emotionlines dataset [12]. MELD is constructed from the script of the urban life TV series Friends, which contains more than 1400 dialogues and 13000 words of contains textual, acoustic, and visual information. Each utterance has seven emotional labels, including neutrality, happiness, surprise, sadness, anger, disgust and fear

3.4.2 Baselines

To comprehensively evaluate the proposed model KES, we compare our model with the following baselines:

(1) CNN [72]

CNN is a convolutional neural network model, which is trained on the basis of pre-trained GloVe [47]. It is the only baseline model without modeling contextual information.

(2) CNN + cLSTM [59]

CNN is used to extract Textual features, and LSTM is used to model context information based on CNN.

(3) DialogueRNN [13]

An RNN-based method uses the speaker, context, and emotion information from adjacent utterances to model the emotion of utterance in conversation.

In this model, CNN is used to extract text features, and independent GRU networks are used to model speaker state and contextual information respectively.

(4) DialogueGCN [29]

DialogueGCN creates a graph based on the interaction to take into account the conversation structure between the participants of speakers. A Graph Convolutional Network (GCN) is employed to encode the speakers. Contextual features and speaker-level features are connected, and attention mechanism based on similarity is used to obtain the final classified discourse representation.

(5) KET [30]

KET is the first model which integrates the external commonsense knowledge and emotion information in emotion lexicon into conversation text feature. The model uses the Transformer decoder to predict the emotional label of the target utterance.

(6) COSMIC [31]

In this model, COMET [53] is used to retrieve commonsense knowledge of event eccentricity from ATOMIC [66]. Based on DialogueRNN structure, this model is applied to ERC tasks with common sense knowledge and has achieved advanced results.

3.5 Results and Analysis

3.5.1 Comparison with Baselines

We compare the performance of the proposed model KES with the baseline method, and the experimental results are shown in Table 3.2 and Table 3.3. The overall results of all the compared methods on the three datasets are reported. We can note from the results that the proposed KES model competitive performances across the three datasets and reaches a new stateof-the-art on the IEMOCAP, DailyDialogue, and MELD datasets.

Table 3.2 Coverall performance on IEMOCAP. Best performances are highlighted in bold.

Model	IEMOCAP						
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-Avg F1
CNN	35.34	53.66	51.61	62.17	50.66	55.56	51.28
CNN+cLSTM	33.90	69.76	48.40	57.55	62.37	57.64	56.04
DialogueRNN	37.94	78.08	58.95	**64.86**	68.11	58.85	62.26
DialogueGCN	42.75	84.54	63.54	64.19	63.08	**66.99**	64.18
KET	-	-	-	-	-	-	59.56
COSMIC	-	-	-	-	-	-	65.28
KES	**47.74**	**84.63**	**64.25**	62.48	**73.26**	63.48	**66.32**

(1) IEMOCAP

IEMOCAP datasets contain binary conversations with natural and coherent discourses. Since the six emotional tags in IEMOCAP are unbalanced, the F1 score of a single label is also reported. The new and most advanced F1 score of KES model is IEMOCAP with 66.32. We observe that the proposed model is around 4% better than DialogueRNN, 2% better than DialogueGCN. As for the models which also adopt the method of quoting the concept of external commonsense knowledge, the proposed model is around 7% better than KET, 1% better than COSMIC. For the model based on CNN and LSTM, there is a performance gap of more than 10%. One of the main reasons of this large performance gap is that some models, such as CNN, CNN + cLSTM and KET, ignore the speaker-level information modeling, which makes the models treat different speakers equally, resulting in a certain loss of performance.

Considering that the average utterance length in IEMOCAP is more than 50, and even the maximum session length is more than 100, the Transformer can capture remote dependencies better than RNNs-based context encoders. Moreover, the conversational semantic role labeling information clarifies the utterance structure and concentrates the conversation information. Besides, External knowledge also enriches semantic information, makes the conversation context more closely related, and expresses the emotional influence of the speaker on himself and other conversation participants.

Table 3.3 Coverall performance on DailyDialogue and MELD. Best performances are highlighted in bold.

Model	DailyDialogue	MELD
	Macro F1	W-Avg F1
CNN	36.87	55.02
CNN+cLSTM	51.84	56.87
DialogueRNN	41.80	57.03
DialogueGCN	49.95	58.10
KET	-	58.18
COSMIC	51.05	65.21
KES	**52.28**	**66.46**

(2) DailyDialogue

In the DailyDialogue dataset, neutral emotion accounts for more than 80% of the test dataset. Because of the unbalanced data distribution, we use the macro F1 score excluding the neutral class as the evaluation index. DailyDialogue dataset contains many short utterances with an average length of 8. In this case, using a speaker encoder to model speaker-level information can release more capabilities to improve performance. According to Xu et al. [30], conversational semantic role labeling information is more sensitive to the information of single sentence dialogue, and it is easier to find the central semantic information fundamentally.

(3) MELD

MELD dataset consists of multiparty conversations, and We follow the same metrics used on the IEMOCAP dataset. Utterances in MELD are much shorter compares to other datasets, and rarely contain emotion-specific expressions, which means that emotion modeling is highly dependent on context. Many dialog scenes contain conversations of more than five speakers, but the average conversation length is only 10 and the minimum length is only 1, which means that emotion modeling is highly dependent on context. Most participants in MELD conversation have only a few words. It is difficult to build a self-reliance model for short conversations. The advantages of transformer over RNNs in capturing the dependence between long-distance speakers are not obvious. Additionally, the discourse in MELD lacks specific emotional expression, which further increases the difficulty of emotional modeling. Nevertheless, the proposed model achieves better results than other baselines, because of the screening and fusion of commonsense knowledge, the utterances in the conversation show more emotional connection in the dialogue context.

3.5.2 The Role of External Knowledge

In Table 3.4, we also report the results of ablation studies by removal of various components from the proposed model KES.

Table 3.4 Ablation Results on Three Datasets.

Method	IEMOCAP	DailyDialogue	MELD
KES	66.32	52.28	66.46
w/o SRL	64.86	51.03	65.78
w/o KG	61.54	50.24	63.96

It can be observed that the performance of KES continues to decline in all datasets. In IEMOCAP dataset, compared with all other datasets, we observed a more severe performance decline without using an external knowledge graph (KG), and the weighted macro F1 dropped by nearly 5%. This might be attributed to considering that the average conversation length in IEMOCAP is at least 50, it is difficult to grasp the core meaning that affects emotion in the absence of commonsense knowledge. It confirms the importance of using commonsense knowledge to identify conversation emotions. Comparing two different knowledge features extracted based on BERT, using ATOMIC knowledge graph or using conversational semantic role labeling, we observe mixed results and can prove the effectiveness of the self-attention-based fusion method.

3.5.3 Case Study

We illustrate a case study on a conversation instance from the IEMOCAP dataset in Table 3.5.

We introduce four key commonsense knowledge relations: intent of speaker (denoted as XI), effect on speaker (denoted as XE), reaction of speaker (denoted as XR), and reaction of others (denoted as YR). The whole conversation transited from negative emotion to neutral utterance, but then the situation quickly turned to negative emotion and ended with neutral utterance. When there is a sudden mood change, it is difficult to find this scene by traditional methods. These models can cause confusion when analyzing similar emotions, such as Frustrated and sad. As can be seen from Table 3.5, due to the intervention of common knowledge, the model is easier to deal with sudden emotional transition and has better sensitivity to similar emotions. The commonsense knowledge model not only predicts the emotional type of the next utterance from the current emotional state of itself, but also predicts the emotional state of the listener. When the conversational utterance is neutral, the commonsense knowledge model predicts that the reaction of others is a negative emotion, which plays a huge role in determining the contextual emotional state of the conversation.

Table 3.5 Case study from the IEMOCAP dataset.

Utterances	Commonsense knowledge	Emotion
A: I’m just so tired all the time.	XI: rest XE: gets hurt XR: exhausted YR: -	Sad
B: Have you been trying to get a job?	XI: to be supportive XE: gets yelled at XR: proud YR: becomes annoyed with B	Neutral
A: I’ve been looking for like eight months.	XI: - XE: gives up XR: worried YR: regret	Frustrated
B: It’s really hard to find a job.	XI: to be better XE: accomplish XR: worried YR: regret	Frustrated
A: I’m tired of the same excuses.	XI: to escape from responsibility XE: lies XR: stupid YR: disappointed	Frustrated
B: Well, okay.	XI: - XE: okay XR: relieved YR: -	Neutral

3.5.4 Impact of Commonsense Relation Type

We investigate the impact of commonsense relation types on the performance of our proposed model KES. Considering that five of the nine relation types of ATOMIC are used in the COSMIC model, that is, the intent of speaker, effect on speaker, reaction of speaker, Effect of others and reaction of others. Intentions and effects on the speaker and others can be divided into psychological states, and their reactions are events. Intention is also a causal variable, and the rest is effect. There are other relation types, which determine the preconditions and post-conditions of a given event and describe how the subject is perceived by others. We expand the relation set to five relation types and all nine relation types, respectively. We calculate the F1 scores of KES with these two categories of relation types added step by step.

From Table 3.6 we can conclude that the inclusion of two extra relation types or all relation types degrades the F1 scores on almost all datasets. We find that too many types of If-Then commonsense relationships would not bring any benefits to the enhancement of knowledge. Although the extra event description enriches the commonsense information in a certain sense, for our model, in the integration and screening of knowledge, the model would miss some important information and focus on the commonsense information that is not critical to the emotional relation.

Table 3.6 Impact of Relation Types on KES.

Dataset	Relation Type		
	4 Relation Types	5 Relation Types	All
IEMOCAP	66.32	66.23	65.68
DailyDialogue	52.28	51.47	50.96
MELD	66.46	65.92	65.14

3.5.5 Analysis of Context Length

The context encoder of KES model is used to process the context of the conversation. Conversation length would seriously affect the performance of the model. In order to compare and verify the performance of our model, we evaluated the influence of KES on conversations with different lengths. On the IEMOCAP dataset, conversations are grouped by length and fed into two models: our semantic enhanced global context encoder and the contrast model using only LSTM global context encoder.

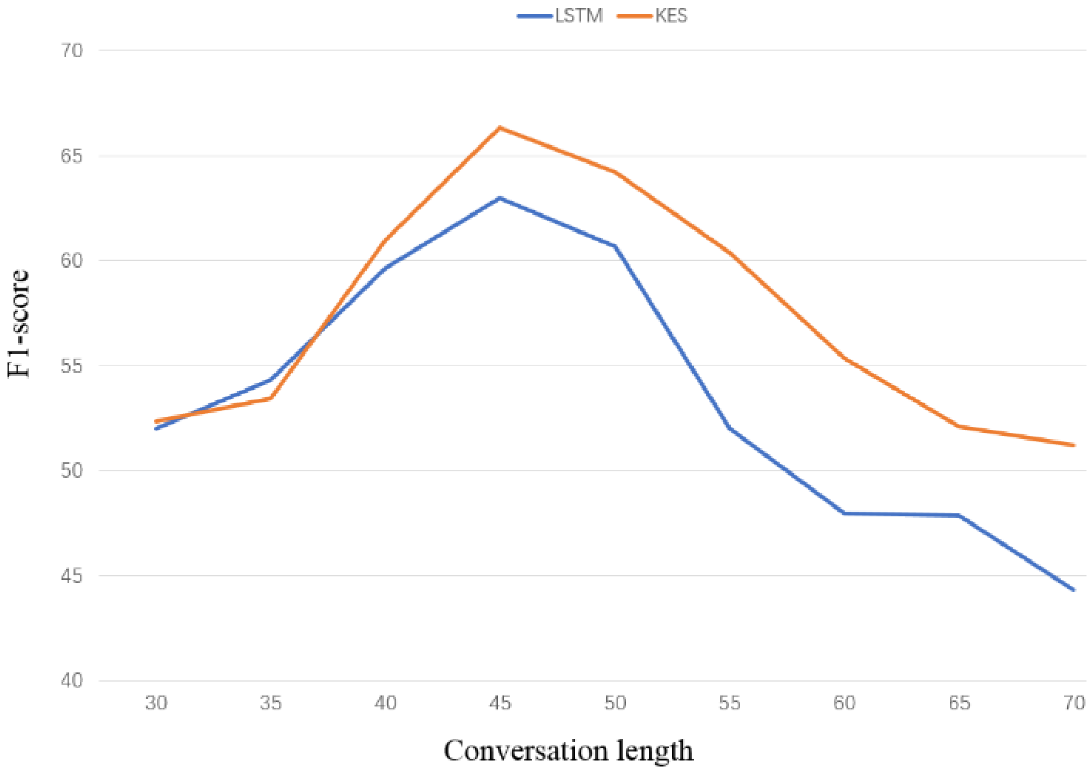

Figure 3.2 Performance of two models in different context lengths.

The F1 scores of different lengths in the two models are shown in Figure 3.2. It is clear that incorporating context into KES improves performance on all datasets. The two context encoders have similar effects on relatively short conversations. However, as the conversation length exceeds 36, KES has more obvious advantages, which proves the contribution of the enhanced semantic encoder based on Transformer to remote context information modeling.

3.6 Summary

This chapter proposes utilizing external knowledge to enhance semantics network architecture that incorporates conversational semantic role labeling Information and the commonsense knowledge feature from ATOMIC for emotion recognition in conversation. A knowledge enhanced language representation layer based on self-attention has been developed for fusion extraction. Based on the utterance representations rich in external knowledge, the contextual external state, and individual internal state are modeled to predict the emotional label of conversation. We have done a lot of experiments on three benchmark data sets. KES has made new state-of-the-art advanced achievements, which proves the effectiveness of the proposed model in external knowledge integration.

Future work will focus on integrating more diversified external knowledge. We also plan to incorporate multimodal information into KES and evaluate it on more natural conversation datasets.

Chapter 4

Conversation Generation with Expressed Emotions

4.1 Introduction

The human-machine dialogue system is an important and challenging task in the field of artificial intelligence, enabling computers to interact with humans through human languages. In recent years, with the rapid growth of social data on the Internet, data-driven open-domain dialogue systems have gradually become the focus of attention in the academic community, and the human-machine dialogue has gradually changed from the role of service to the role of emotional partner [73]. Thanks to the rapid advance of deep learning, neural networks make breakthroughs in speech recognition and machine translation, and expand to the breakthroughs in conversation generation. The emotional intelligence of conversation generation significantly defines the ability to perceive, understand, express and control emotions [74], developing the methods for conversation generation problems such as semantics, grammar, smoothness, etc.

Conversational modeling is an important task in natural language understanding and machine intelligence. Advances in end-to-end training of neural networks have led to remarkable progress in many domains such as speech recognition, computer vision, and language processing. In the expression of the conversation system, talking with users is primarily conducted through language information, but non-language information, such as facial expressions and gestures, is also very important.

Many pieces of research have focused on adding emotional characteristics to conversation generation. For instance, Ghosh et al. [75] proposed a language model called Affect-LM, the Long Short-Term Memory (LSTM) that was used to extend the language model to generate natural emotional expressions. Inspired by Affect-LM, Zhou et al. [1] developed a model called an emotional chatting machine (ECM) that was equipped with new mechanisms for emotional conversation generation. The model has an internal emotion state for balancing grammatical and emotional dynamically, and an external emotion memory to help generate more explicit and unambiguous emotional expressions. Nabiha et al. [41] elaborated a new method that combined the advantages of Affect-LM and ECM. They introduced an effective dictionary for emotional vector embedding, and improved the model on the basis of Seq2Seq in terms of word vector, objective function and decoding, respectively. To cope with the problem in the low quality of response, Peng et al. [77] have grown an interest in combining topic and emotion for response generation, they proposed a topic-enhanced emotional conversation generation model that incorporated emotional factors and topic information. Based on the structure of the Seq2Seq model, Asghar et al. [41] proposed a combination of emotional word embedding representation, modified loss function, and sentiment-based directional search methods to generate dialogue. Shi et al. [42] extracted emotional information from multi-modal dialogue datasets, and used the reinforcement learning method to directly use the user's emotional information as real-time rewards in the reinforcement learning algorithm. Shantala et al. [78] also used Seq2Seq as the basic model and added the trained emotion vector in the decoding stage to generate the reply sentence. Based on multimodal dialogue data, Liang et al. [43] proposed to treat historical emotion information, user facial expressions, and speaker personality as node information in heterogeneous graphs, and use heterogeneous graph neural networks to extract the information.

In this chapter, we propose a dynamic emotional session generation model based on Seq2Seq and attention mechanism in response to the problem of emotional response in open-domain dialogue systems.

The rest of the chapter is organized as follows: Section 4.2 defines the task objectives. Section 4.3 introduces the basic model Seq2Seq and the attention mechanism in detail. Section 4.4 shows an overview of the proposed method. Section 4.5 introduces the relevant situation of the experiment. Section 4.6 provides experimental results and analysis of Section 4.5. Section 4.7 presents our conclusions and future work.

4.2 Task Definition

Open-domain dialogue generation is an essential task in Natural Language Understanding (NLG). With the recent development in Deep Learning, neural conversational models have attracd increasing attention in the past years [81, 82, 83].

Besides evaluating the probability of a sentence, a Language Model can also estimate the probability of the next word given a sequence of words. Formally, let $\{\omega = \omega_1, \omega_2, \dots, \omega_n\}$ be a sentence of length n. In the case of prediction the word ω_t given $\{\omega_1, \omega_2, \dots, \omega_{t-1}\}$, a language model that yields the following joint probability gives the following:

$$\mathcal{P}(\omega = \omega_1, \omega_2, \dots, \omega_n) = \mathcal{P}(\omega_t | \omega_1, \omega_2, \dots, \omega_{t-1}) \mathcal{P}(\omega_1, \omega_2, \dots, \omega_{t-1}) \quad (4.1)$$

After applying the chain rule of probability, it gives the following:

$$\mathcal{P}(\omega) = \prod_{t=1}^{n} \mathcal{P}(\omega_t | \omega_1, \omega_2, \dots, \omega_{t-1}) \quad (4.2)$$

where $\mathcal{P}(\omega_1 | \omega_0) = \mathcal{P}(\omega_1)$, since we can regard w0 as a constant value that indicates the signal of starting.

Compared to the standard language model, a conditional language model assigns probabilities to a sequence of words ω, given some conditioning context x, the conditional probability as follows:

$$\mathcal{P}(\omega|x) = \prod_{t=1}^{n} \mathcal{P}(\omega_t|x, [\omega_1, \omega_2, \dots, \omega_{t-1}]) \tag{4.3}$$

where $\mathcal{P}(\omega_1|\omega_0) = \mathcal{P}(\omega_1)$. Conditional language models have many real-word applications such as machine translation [84, 85, 86, 94, 95], question answering [87, 88], dialogue generation [83, 89], etc.

Open-domain dialogue generation tasks itself can be regarded as a problem of conditional language modeling, the context x is the history of the whole conversation session or just a single input from the user and the generated sequence ω is the response by the dialogue generation model given the context x.

There have been several types of unique deep learning network frameworks that are able to handle the task of conditional language modeling, for example, Sequence-to-sequence (Seq2Seq) [86, 83, 94, 95], Conditional Variational Autoencoders (CVAE) [90, 91], Sequence Generative Adversarial Nets (SeqGAN) [92, 93], etc. In theory, any conditional language model can be used for the task of open-domain dialogue generation. The model generates sequence Y when the user gives input X. Therefore, in this study we chose Seq2Seq as the backbone of the dialogue generation model.

After determining the basic conditional language model, we need to make it generate more emotional responses. Let e be the emotion we want to express, following the above annotation, we need to train a conditional language model that uses both the context X and emotion e as conditions, the conditional probability distribution becomes $\mathcal{P}(Y|X, e)$, where context X, depending on the task, maybe the entire conversation context, or just a single query.

4.3 Seq2Seq with Attention

Our approach was based on recent work which proposed the use of neural networks to map Sequences-to-sequences. This framework has been applied to neural machine translation and has been improved in English–French and English–German translation tasks in the WMT14 data set [96,97]. It has also been used for other tasks such as parsing [98] and image subtitles [99]. The sequences-to-sequences model is a deep learning method based on a recurrent neural network (RNN). In general, feedforward neural networks are unidirectional network structures formed from the input layer to the output layer, while RNNs are closed networks with its own output as its own input is formed. The canonical sequence-to-sequence model is an encoder–decoder structure. The encoder and decoder have the same architecture; the encoder takes the input sequence and maps it on to an encoded representation of the sequence. The RNN of the encoder compresses the input sequence into a vector and then transmits the vector to the decoder to generate an output sequence. If the data are text based, a network containing the previous words are created by inputting the words into the article one by one.

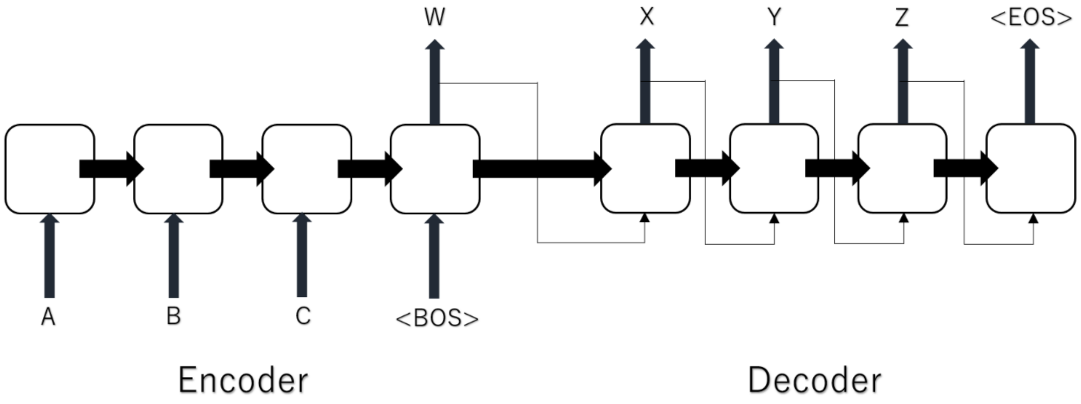

Figure 4.1 Structure of Sequence-to-sequence model.

It is well known that vanilla RNNs suffer from vanishing gradients, and most researchers use variants of long short-term memory (LSTM) recurrent neural networks [100] and the gated recurrent unit (GRU) [101]. Usually, we used embedding before we entered data into the model. The embedding layer is a type of word embedding that is learned jointly with the neural network model in specific natural language processing tasks. We must first compile a “vocabulary” list containing all the words that our model should be able to use or read. The model's input must be a tensor containing the identification of words in the sequence.

On the whole, the sequences-to-sequences model has a very wide range of application scenarios, and the effect is very strong. Because the sequences-to-sequences model is an end-to-end model, it reduces many manual processing and rule-making steps. Although the Encoder-Decoder framework is very classic, it is also very limited. The only connection between the encoder and the decoder is a fixed-length semantic vector, and the encoder needs to compress the entire sequence of information into a fixed-length vector. The semantic vector may not completely represent the information of the entire sequence, and the information contained in the content input to the network first would be overwritten by the information input later. This phenomenon becomes more serious when the input sequence is longer. This situation makes the decoder not obtain enough information of the input sequence at the beginning when decoding. Bahdanau et al. [95] introduced an attention mechanism based on the Encoder-Decoder structure, which made the depth of the methods more prominent in every task.

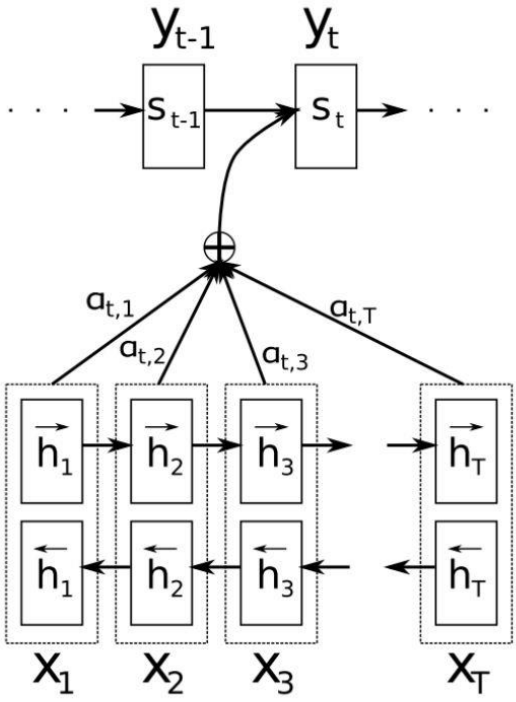

Figure 4.2 Structure of Attention mechanism [95].

In the decoding process, the correlation degree between the hidden state $\{h_1, h_2, .., h_T\}$ of each encoder and the hidden state s_{t-1} of the decoder is calculated and the softmax normalization operation is performed to obtain the weight a_{ij} of each hidden layer vector. The calculation process as follows:

$$e_{ij} = v_a^T \tan h\,(W_a s_{i-1} + U_a h_j) \tag{4.4}$$

$$a_{ij} = \frac{\exp\,(e_{ij})}{\sum_{k=1}^{T} \exp\,(e_{ik})} \tag{4.5}$$

where e_{ij} represents the correlation between the previous hidden layer state s_{i-1} of the ith output and the jth input hidden layer vector h_j, a_{ij} represents the normalized weight value of e_{ij} after softmax operation.

The decoder performs a weighted summation on $\{h_1, h_2, .., h_T\}$ to obtain the encoding vector c_i of the input sequence $\{x_1, x_2, .., x_T\}$ corresponding to this decoding. The calculation formula is as follows:

$$c_i = \sum_{j=1}^{T} a_{ij} h_j \tag{4.6}$$

Then calculate the hidden state s_i of the decoder at time i according to the code vector c_i, and finally get the output y_i of the decoder.

The most important step of the Attention mechanism is how to generate a different language encoding vector ci at each moment, which indicates which parts of the input sequence should be focused on when outputting, and then generate the next output according to the area of interest. Compared with the original Encoder-Decoder model, the biggest difference after adding the Attention mechanism is that it does not require the encoder to encode all input information into a fixed-length vector. Instead, the encoder needs to encode the input into a sequence of vectors, and each step of the decoding process would selectively select a subset from the sequence of vectors for further processing.

4.4 Methodology

4.4.1 Overview

We assume that given a pretreatment corpus $\{sen_1, sen_2, \dots, sen_i, \dots, sen_n\}$, sen_i is corpus sentences, words and sentences by the plurality of sequences $\{sen_i = w_1, w_2, \dots, w_i, \dots, w_n\}$. At the same time, there is an emotion dictionary dic, which consists of a positive emotion vocabulary dic_{pos} and a negative emotion vocabulary dic_{neg} . We divide the emotion category z into six emotions $\{Anger, Disgust, Satisfaction, Happiness, Sadness, Neutral\}$, aiming to generate a response that includes a specific emotion category $\{Y = y_1, y_2, \dots, y_N\}$. We use a semi-supervised method to construct the required sentiment dictionary [102], which is constructed from a corpus of sentences with sentiment categories and annotations. Given w_j look up the corresponding emotion polarity pol in the emotion dictionary dic, and according to the vector mapped from the position of w_j in the sentence sen_i, the emotion word vector is obtained by splicing the word vector with the vector. At the same time, get the corresponding sentence sen_i according to w_j, and map the emotional polarity corresponding to each word in the sentence to the same vector X. The word vector and the vector X are spliced together to obtain the emotional sentence vector.

The emotional vocabulary generated by the model is often related to the word order in the corpus. When each sentence contains an emotional dialogue corpus, the effect may be good. However, in practical applications, large-scale corpora do not have too much emotional information annotation, and emotional information is not Intensive, and there is not a lot of manpower and material resources to clean the corpus specifically for emotional information. The emotional response generated in this way is not stable, nor can it output a specified and reasonable emotional response.

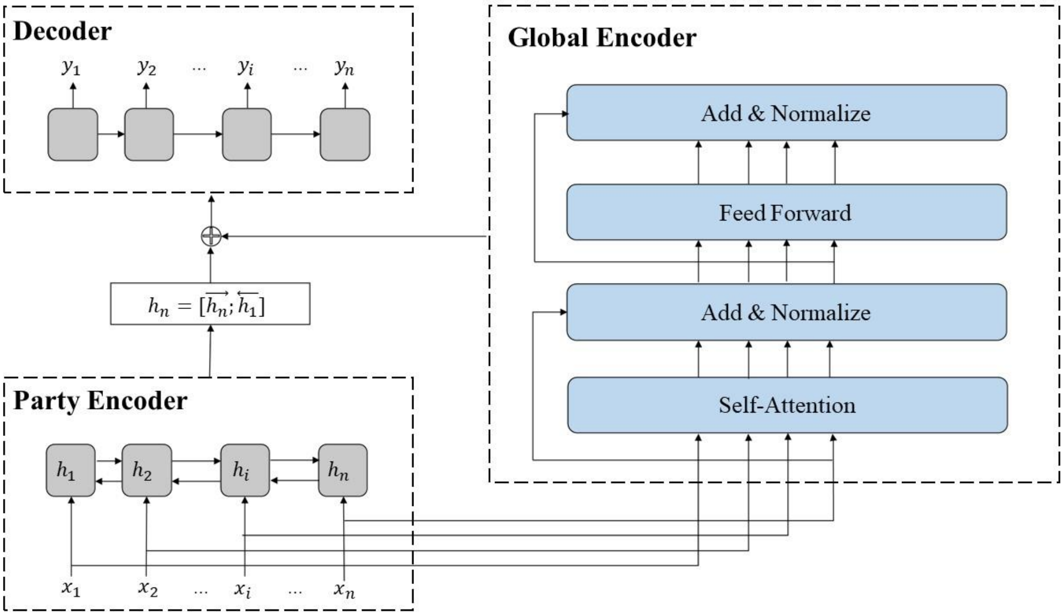

Figure 4.3 Overview of the proposed emotional dialogue model.

The proposed model contains party encoder, global encoder, and decoder, as shown in Figure 4.3. In general, our model is different from the traditional Sequence to sequence models. As illustrated in Figure 4.3, the party encoder encodes the individual sentences and emotional information of the dialogue. Here, we adopt the bidirectional LSTM structure to ensure that the information before and after the dialogue would not be ignored. At the same time, the global encoder of transformer structure would encode the context and emotional information of the entire dialogue to ensure that the model would combine the context and consider the cause and effect when generating the dialogue.

4.4.2 Embedding of emotional information

Because in computer vision and natural language processing, the development of neural network model requires a large amount of computing and time resources, and the technical span is also relatively large; thus, the pre-trained model is usually reused as the starting point of computer vision and natural language processing tasks [76]. We proposed to divide the embedding of emotional information into two parts: word-level emotional embedding and sentence-level emotional embedding, using emotional word vectors and emotional sentence vectors, respectively. Word-level emotion embedding is based on word vectors. In the input part of the model, the current input and context are combined to obtain the emotion word vectors in all inputs. Sentence-level emotions are embedded in the encoder part of the model, mainly to capture the co-occurrence and dependency relationships between emotional words.

The word vector does not have the ability to obtain emotional information, so we need to improve the traditional word vector. The emotional content of the dialogue system is mainly based on emotional vocabulary, so choosing the emotional dictionary as the source of the model's emotional information is the most feasible choice. The emotional dialogue model combines traditional word vectors word2vec with emotional information to form emotional word vectors according to the emotional dictionary during input, so that the model can obtain word-level emotional information. Word2Vec is an open-source tool based on deep learning. Due to its high accuracy and relatively low computational cost in analyzing the semantic similarity between two words, it has become more and more popular recently. It has two modes: Continuous Bag-of-Words (CBOW) and Skip-Gram, which can be used to quickly learn word embedding from the original text, and use the constructed neural network model to capture word relationships. We adopt the Skip-Gram model for word embedding training. After the text preprocessing step, the word vector representation of all words in the document is learned through the Skip-Gram model, and the vector representation of each word encountered in the input is extracted from the model. Then, the representation of each sentence can be received by averaging the vectors of all its constituent words. The vector representation of the new document is derived in the same way in the sentiment prediction stage.

After the text preprocessing step, each word in each sentence of the dialogue is converted into a vector. Then according to their cosine similarity, the obtained vector is compared with the vector of the emotional word. Select the emotional word vectors with high similarity for combination. Due to the existence of the emotional dictionary, the feature extraction method based on the dictionary is very promising. Generally, the sentiment dictionary consists of a set of language-specific words, including information about the sentiment category, polarity strength, etc. to which they belong. Compared with a simple sentiment dictionary, a fine-grained multi-sentiment polarity dictionary can improve the classification accuracy in sentiment analysis tasks [104].

A set of documents and sentiment dictionary are given to build a model and generate sentiment word vectors. A set of text $D = \{d_1, d_2, \dots, d_n\}$ and the vocabulary $T = \{t_1, t_2, \dots, t_n\}$ which are unique terms extracted from the corpus. The word representation of the terms t_i are mapped from the Word2Vec model, and then a set of word representations of all words in the vocabulary is derived, i.e., $V = \{v_1, v_2, \dots, v_m\}$.

In order to construct the emotional word vector, the embedding model uses all the emotional words in the dictionary as emotional vocabulary $E = \{E_1, E_2, \dots, E_k\}$, and obtains their word representations $V_E = \{V_E, V_{E_2}, \dots, V_{E_k}\}$. Due to the scale of the word embedding model of the training corpus, there is a low coverage problem, that is, some words are in the emotional vocabulary but not in the vector space learned from Word2Vec. Therefore, these words can be ignored and deleted in subsequent processing.

The emotion vector embedded in this way not only contains the current input emotion information, but also contains the position information of the emotion vocabulary, which belongs to a kind of position embedding [48, 105]. It shows that the positive emotion vocabulary and the negative vocabulary are in different positions in the sequence, so that the neural network can learn the relationship between position and emotion. This serial number relationship is the absolute position in the word sequence. In addition to the absolute position, the relative position is also very important, because the length of the sentence itself is inconsistent, so in order to add the relative distance information, it is convenient to calculate in parallel. We set the emotion vector part of the emotion word vector to the maximum length of the sentence. When the dimension of the emotion vector is determined, the absolute position and relative position of each word would be the same, so that position embedding can embed both absolute position and relative position information at the same time.

After the emotion vector is obtained, it is spliced with the word vector to obtain the emotion word vector. Compared with the original word vector, the emotional word vector increases the emotional information in the emotional dictionary, and at the same time embeds the information of the absolute position and the relative position.

In this work, we construct an emotion-aware word embedding to represent each word by connecting the dictionary-based emotional word vector and the semantic word vector based on the Word2Vec model. The embedding model can capture the emotional orientation of words, and is a word representation method strictly based on word embedding. In addition, it makes full use of the context and semantic expression capabilities of the word embedding model. Emotion perception word embedding combines emotional word vectors and semantic word embedding to simplify the combination function, which shows that the advantages of the two models could be combined in a hybrid representation.

4.4.3 Model

4.4.3.1 Party Encoder

The encoder part of the Seq2Seq framework uses a bidirectional LSTM [17]. The current input information is the sentence entered by the user in this round of dialogue. After word segmentation and preprocessing, it is transformed into a word2vec word vector and an emotion vector, and the two are spliced to obtain an emotion word vector.

We assume that given an input sequence $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$, an output sequence $Y = \{y_1, y_2, \dots, y_i, \dots, y_n\}$. At a certain time node t , the encoder's hidden layer state h_t at the current time t is calculated by the hidden layer state $h_t - 1$ of the encoder at the time $t - 1$ and the current time input x_t, the following formula can be expressed:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}) \tag{4.7}$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t-1}}) \tag{4.8}$$

$$h_t = \overrightarrow{h_t} + \overleftarrow{h_t} \tag{4.9}$$

Where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the hidden layer state of the encoder t of the forward LSTM and the backward LSTM respectively, h_t is the hidden layer state of the encoder at the current moment.

Song et al. found that there are at least two ways to make sentences contain emotional information. One is to clearly use strong emotional words to describe emotional states (such as "Happy", "Sad", "Angry", "Excited", "Frustrated", etc.). The other is to increase the intensity of emotional experience, not through words in the emotional vocabulary, but through the implicit combination of neutral vocabulary in different ways to express emotion. The proposed emotional dialogue model does not forcibly insert strong emotional expression words into the generated utterances. We believe that this would cause the generated sentences to deviate from the original meaning, and the naturally generated replies based on the query would be more in line with the central meaning of the chat.

4.4.3.2 Global Encoder

Context information is one of the important components of emotional dialogue. Context can provide context information for prediction. At the same time, context may also have certain emotions. By encoding context information, it can also affect some interrelated emotions. Vocabulary is conducive to the model's learning of emotional changes, and it can also learn some hidden emotional information that may exist.

As the number of dialogue turns increases, the context length would also become longer. When the length exceeds a certain distance, RNNs cannot capture long-distance dependence. However, the length of each round of dialogue is relatively short, so the use of a transformer would be very effective. We propose to adopt transformer as the basis of the global encoder structure to process the encoding information of the context.

The method proposed conversation as a set of dialog sentences sequence composition $D = \{sen_1, sen_2, \dots, sen_n\}$, and each statement as a series combination of words sen_i , the hidden node state information at the tail node is the semantic encoding of all contexts. Sentence-level LSTM obtains contextual information by iteratively processing the vector generated by each utterance sequence.

The proposed method unfolds all tokens in the dialogue into $D = \{w_1, w_2, \dots, w_m\}$. Where m is the number of tokens entire conversation. An end-of-turn delimiter is inserted between each two turns. Then use Transformer to encode the expanded token sequence. We connect all the markers in the dialogue as input, and hope that the Transformer can learn rudimentary coreference information through the self-attention mechanism. For each token w_i, the input embedding is the sum of its word embedding, position embedding and turn embedding:

$$\boldsymbol{I}(w_i) = \boldsymbol{WE}(w_i) + \boldsymbol{PE}(w_i) + \boldsymbol{TE}(w_i) \tag{4.10}$$

Where the word embedding $\boldsymbol{WE}(w_i)$ and the position embedding $\boldsymbol{PE}(w_i)$ are the same as in normal Transformer architectures [48]. According to the work of Su et al. [106], we added the turn embedding $\boldsymbol{TE}(w_i)$ to indicate which turn each token belongs to. Tokens from the same turn would share the same turn embedding. Tokens from the same round would share the same round embedding. Then the input embedding is forwarded to l stacked encoders to obtain the final encoding representation. Each encoder contains a self-attention layer followed by a feedforward neural network.

$$\boldsymbol{E}_0 = (\boldsymbol{E}_1, \boldsymbol{E}_2, \dots, \boldsymbol{E}_m) \tag{4.11}$$

$$\boldsymbol{E}_l = TransformerBlock(\boldsymbol{E}_{l-1}) \tag{4.12}$$

As shown in Figure 4.3, Each self-attention and feedforward component has a residual connection and layer normalization step, and the final encoding is the output of the lth encoder $\boldsymbol{E}_l$.

4.4.3.3 Decoder

The decoder is the opposite of the encoder. Our model focuses on modeling party sentiment information and global sentiment information when generating responses based on posts. Therefore, the decoder should combine party and global vectors for decoding. We adopt LSTM and Transformer to respectively encode the dialogue to obtain the hidden state of the model, and then the model obtains the attention distribution in the attention mechanism, and then connects them. As shown in equation 4.13, given h_t and $\boldsymbol{E}_l$, the decoder generates the response p:

$$p = g([h_t\, \boldsymbol{E}_l]) \tag{4.12}$$

Where g is a nonlinear multilayer neural network to calculate the probability of each word. In the decoding process, the decoder with LSTM is used to predict the response by token.

4.5 Experiments Setup

4.5.1 Dataset

As shown in Table 3.1 in Chapter 3 of this thesis, this experiment also used the dialogue model proposed by the MELD and DailyDialogue emotional dialogue datasets.

MELD It is a multi-party and multimodal emotional dialogue dataset from the Friends TV series, which contains textual, acoustic, video, and speakers information. Each utterance in every dialogue is labeled with one of the seven emotion categories.

DailyDialogue To verify the generalizability of our model across datasets, we conduct experiments on DailyDialogue, a larger scale dataset only containing textual utterances with the same emotion categories as MELD.

4.5.2 Baselines

To comprehensively evaluate the proposed emotional dialogue model, we compare our model with the following baselines:

(1) Seq2Seq + Attention [86]

A standard Seq2Seq model with attention method is widely used as a baseline in conversation generation tasks. This model proposes a general end-to-end sequence learning method. In this method, the encoder maps the input sequence to a fixed-dimensional vector. The fixed-dimensional vector is regarded as the semantic expression of the input sequence. The decoder generates the output sequence according to the semantic expression of the input sequence.

(2) ECM [1]

A classic model of emotional conversation generation. This method put forward an emotional dialogue system based on memory neural network. Based on the traditional sequence-to-sequence model, with the static emotion vector expression, a dynamic emotional state memory network, and an external memory mechanism of emotion words, so that the ECM can output responses corresponding to the emotions according to the designated emotion category. Because the ECM model cannot automatically select the appropriate emotional label to respond, the ECM model is not suitable for direct use as a baseline. Therefore, we manually specify the most frequent response sentiment for ECM for fair comparison.

4.5.3 Metrics

In the evaluation of the non-task-oriented dialogue system, the accuracy of utterance selection and utterance generation were evaluated. In other words, we compare the utterance generated by the model with the response in the test data to verify the accuracy of the generated utterance. Another method is to consider the effect of the response sentence through the meaning of each word and to judge the relevance of the test data's response. The word vector is the basis of this evaluation method. The advantage of using a word vector is that it can increase the diversity of answers to a certain extent because most of them are characterized by word similarity, which is much lower than the restriction of requiring identical words in word overlap.

To measure the effectiveness of our approach, we take the following metrics:

(1) BLEU [107]

Measuring the consistency of emotional conversation generation without losing the syntax performance can effectively highlight the effect of conversation generation. As for objective syntax evaluation, BLEU, a syntax measure to compute n-gram overlaps between the generated response and the reference response, was also used to measure the syntax of the responses. The n-gram is used to compare the similar proportions of n groups of words between an utterance and a reference. A result of 1-gram represents the number of words in the text that were translated separately, so it reflects the fidelity of the translation. When we calculated more than 2-gram, more often the results reflected the fluency of the translation, and the higher the value, the better the readability of the article. BLEU uses the following formula to calculate the similarity between the generated response utterance and the reference of the test data based on the number of n-gram matches between two utterances.

$$BLEU = BP \cdot \exp\,(\sum_{n=1}^{N} \omega_n \log p_n) \tag{4.13}$$

p_n compares the generated response sentence with the reference utterance of the entire test data set and calculates the n-gram matching rate. The score is calculated by calculating the geometric mean from 1-gram to N-gram. The BLEU score is represented by a real number from 0 to 1. The higher the value, the better the response generated.

(2) ROUGE [108]

We use the ROUGE to assess the recall rate of the generated responses. It is different from the BLEU metrics. The BLEU calculates the accuracy rate, and the ROUGE calculates the recall rate. The ROUGE is a common metric for quantitative comparison to measure the actual role of responses in a generation.

(3) PPL

We use natural language processing commonly used quantitative metrics to assess the quality of the generated responses. PPL is a common metric for quantitative comparison to measure the ability of language in express.

(4) Greedy Matching [109]

The greedy matching method is a matrix matching method based on word level. For each word of the real response, the word with the highest similarity in the generated response is found, and the cosine similarity matching is added and averaged to the maximum extent. The same is performed for the generated response again and the average of the two is taken.

(5) Embedding Average Cosine Similarity [110]

The embedding average method directly uses sentence vectors to calculate the similarity between real response and generated response, while sentence vectors are obtained by the weighted average of each word vector, and then it uses cosine similarity to calculate the similarity between two sentence vectors.

(6) Vector Extrema Cosine Similarity [111]

The utterance vector is calculated by the word vector, and the cosine similarity between the utterance vectors is used to indicate the similarity between the two. The calculation of the speech vector is slightly different; here, the calculation method for the similarity of the speech vector is “Vector Extrema”

(7) Skip-Thought Cosine Similarity [112]

Skip-thoughts vectors is the name given to simple neural network models for learning fixed-length representations of sentences in any natural language without any labeled data or supervised learning. With Word2vec, words can be displayed in distributed expressions, and the processing that takes into account the meaning of words can be performed. Kiros et al. let the model learn distributed expressions by learning to predict the words before and after a word in a sentence. This model is called the Skip-gram model of Word2vec. The model uses large novel texts as a training data set, and the encoder part of the model obtained with the help of the Seq2Seq framework is used as a feature extractor, which can generate vectors of arbitrary sentences. The trained Skip-Thoughts model encodes similar sentences that are close to each other in the embedding vector space. In this experiment, the cosine similarity was calculated based on the vector generated using the Skip-Thoughts model.

When comparing the frequency of words in different texts, the words to be compared are mostly predetermined by the analyst. Such a comparison method can statistically verify the hypothesis put forward by the analyst. However, in actual statistical analysis, it is not certain which word should be paid attention to in advance. Generally, the frequency of all words that appear in the text are compared, and the words with a large frequency difference between the texts are searched. This method does not consider the similarity of words according to their linguistic meanings but is based on the distribution of words appearing around the text set in space. With reference to the word frequency in the test data set, we analyzed the utterances generated by the proposed dialogue model and baselines and calculated their spatial similarity.

4.6 Results and Analysis

4.6.1 Automatic Evaluation Results

The results are given in Table 4.1, numbers in bold mean that improvement from the model on that metric is statistically significant over the baseline methods. It can be seen from the experimental results that the proposed conversation model was stronger than the other models for all evaluation indicators. Seq2Seq performed rather poorly on nearly all emotion metrics, primarily because it did not consider any effect factor and tended to generate generic responses.

Table 4.1 Automatic evaluation results.

Model	BLEU	ROUGE	PPL	Greedy	Average	Extrema	S-Tho.
Seq2Seq	0.23	0.16	85.65	0.42	0.59	0.36	0.18
ECM	0.25	0.18	78.83	0.46	0.62	0.43	0.35
Ours	**0.26**	**0.19**	**76.51**	**0.48**	**0.63**	**0.48**	**0.42**

Table 4.1 shows that our model consistently surpasses other competitors on most evaluation metrics. This suggests that modeling the global context has a positive effect on response generation, and demonstrates the superiority of our model on understanding content and emotion, which yields new state-of-the-art performances.

From the results in Table 4.1, we see that compared with Seq2Seq and ECM, the proposed model had improved BLEU evaluation indicators of 0.03 and 0.01 from the results in Table 4.1, respectively. Especially in 1-gram, the results of the proposed dialogue model were much better than the results of the other two models. Two-gram was not much greater than the other two models. In the field of machine translation, 1-gram becomes an indicator of the correctness of word translation, and high-order n-gram is an indicator of translation fluency. Since n-grams only have the same words in the utterance, even synonyms would be regarded as different, thereby reducing the results. The proposed model had the best performance in the generation of the same single word, but compared to multiple words, it was not as good as ECM. Among all the results, the BLEU result was the lowest value overall. The response space in the dialogue system often diverged. BLEU did not care about grammar, only the distribution of content, which is suitable for measuring the performance of the data set, and it could not play a good role at the sentence level. In terms of evaluating non-task-oriented dialogue systems, it is difficult to say that BLEU was the best evaluation index. Therefore, it is of great significance to study appropriate evaluation indicators.

We observe that the proposed model is better than Seq2Seq by about 0.03, 9.14, 0.06, 0.04, 0.12, and 0.24 on ROUGE, PPL, greedy matching, embedding average cosine similarity, vector extrema cosine similarity, and skip-Thought cosine similarity, respectively. Our model outperforms the best ECM by about 0.01, 2.32, 0.02, 0.01, 0.05, and 0.07 on ROUGE, PPL, greedy matching, embedding average cosine similarity, vector extrema cosine similarity, and skip-Thought cosine similarity, respectively. These results show that the proposed model indeed generates more fluent, diverse, coherent, and emotional responses than baselines. The word vector evaluation index focuses on comparing the semantic similarity between the generated sentence and the actual sample, but it is difficult to capture long-distance semantic information based only on the word vector. Intuitively, words with special meaning in the text should have a higher priority than the commonly used expressions. Since most texts show tendencies to a greater or lesser extent, if this method calculates the similarity in the vector space, higher-order general sentences would be output first. It can be seen from the results that our model's choice of key information was due to the other two models, and our model could generate utterances corresponding to the key information of the input utterance.

Furthermore, ECM adopts random emotional words to decode the response. When the opposite emotional information is added, the decoder may suffer the situation that the response generation does not conform to the syntax, and may reduce the quality of response. Different metrics demonstrate that our model allows response to transition naturally and conform to grammar.

Since there are not any standard automatic metrics to evaluate the response generated by the dialogue model, human judgment evaluation is a good index strategy at this stage. Unfortunately, it suffers from time-consuming and laborious.

4.6.2 Case Study

To give an insight on whether the emotion of the generated response is expressed appropriately, Table 4.2 lists some examples generated by the Seq2Seq, ECC, and the proposed model. We select the query of different emotion categories for consistency check. It shows that the proposed model can generate responses with suitable emotions.

Table 4.2 Sample responses generated by all models.

Query	Model	Response	Emotion
Why no one thinks that I look handsome in this photo.	Seq2Seq	I don't know!	Frustrated
	ECM	I'm sorry for that.	
	Ours	I'm sorry that I did not get a good picture for you.	
Your words are really hurtful.	Seq2Seq	Okay, really?	Angry
	ECM	I do not care what people think.	
	Ours	I'm sorry if you're hurt.	
Warmest congratulations on your achievement.	Seq2Seq	Okay, thank you.	Happy
	ECM	I'm happy that I did.	
	Ours	I appreciate your congratulation.	
I was dumped one months ago, and I'm still heartbroken.	Seq2Seq	Uh-huh. I'm sorry.	Sad
	ECM	What happened? Why are you still depressed?	
	Ours	Everything will be okay. I'm with you.	

From Table 4, we can see that the Seq2Seq model tends to generate safe responses such as "Okay", "Thank you" and "I don't know". The Seq2Seq model not only fails to grasp the user's emotions well, but also fails to produce the quality of empathy as an emotional support dialog system. Due to the limitations of the corpus, a lot of meaningless words would be generated in the conversation generation. But the proposed model largely avoids the generation of words like "I don't know" compared to Seq2Seq.

In contrast, the replies generated by the proposed model and ECM produced a certain degree of emotion. ECM needs to specify the sentiment of the response, and there would be more related sentiment words in the generated response. For example, "I do not care" is generated when the specified emotion is angry, "happy" is generated when the specified emotion is happy, and "depressed" is generated when the specified emotion is sad. The emotions generated by our emotional dialogue model are more obscure, and would not appear abrupt and straightforward. In terms of dialogue skills, our model appears to be more professional. For the response in the “Anger” line, our model does not have a tendency to continue the quarrel like ECM, but cleverly chooses to resolve conflicts. Such skills can be seen in the example conversation that we showed. For the response in the “Sad” line, while learning about the help-seekers situation, the proposed model express understanding and empathy to relieve the help-seekers frustration by using dialogue skills.

Due to the architecture of the proposed model, the context could be learned, and we believe that the proposed model can predict the before and after responses. Furthermore, we observed that the generated response text had positive emotions. Although the grammar of the response sentence generated by the ECM model was fluent, the response sentence generated by this model was not strongly related to the expected context when the input sequence was short. For the Seq2Seq model, large security response texts were generated, and the model generated large sentences that lacked fluency and sentences irrelevant to the query. For “Anger” categories, Seq2Seq generates responses that are not only emotionally inconsistent, but also have nothing to do with the post. Instead, most examples in experimental results illustrate that our model is efficient in consistence when we give the post to generate the response.

4.7 Summary

In this chapter, we proposed a party encoder and a global encoder to generate emotional responses for dialogue generation. Emotional intelligence is the ability to monitor an interlocutor's emotions and in turn, appropriately express emotions in response. We separately model and monitor local emotional dialogue information and global emotional dialogue information. Extensive experimental results showed that our emotional dialogue model performed favorably on both content coherence and user satisfaction against other models.

Future work will focus on integrating more diversified external emotional knowledge. We also plan to incorporate multimodal information into the proposed model and evaluate it on more natural large-scale emotional dialogue datasets. In terms of evaluation experiments, human judgments evaluation experiments are carried out in order to obtain accurate and comprehensive results.

Chapter 5

Conclusion and Future Works

Human beings are rich in emotions, and various emotions, such as happiness, anger, sorrow, and joy are easily expressed in dialogue, so emotional dialogue is one of the important components of human dialogue. Dialogues that lack emotion are often empty and weak and hard to resonate with people. With the development of deep learning, enhanced learning, and the construction of large dialogue data sets, research on emotion recognition tasks in dialogue has attracted attention. In many real-world applications, the dialogue model focuses on emotion recognition and emotional dialogue generation, which aims at reducing individuals' emotional distress and helping them understand and work through the challenges that they face. In this thesis, we focus on the emotional conversation recognition model and the emotional dialogue generation model. We will summarize the whole thesis and then give our future work in this section.

5.1 Conclusion

This thesis focuses on the research of emotional conversation recognition and generation, and proposed some related methods for perusing more accurate emotional prediction and generation in dialogue. In the task of dialogue emotion recognition, our work mainly revolves around some remaining challenges in this field: (1) How to achieve more accurate modeling of dialogue text, (2) How to introduce external commonsense knowledge to enrich emotional information, (3) How to integrate different external knowledge information to improve the performance of emotional model. In the task of generating emotional dialogue, our work mainly revolves around the emotional accuracy of the generated response, the diversity of the response sentence, and the relevance between the response and the context. The concrete summaries and contributions are displayed as follows:

- We proposed KES model, a new framework that incorporates different elements of conversational semantic role labeling Information and the commonsense knowledge feature from ATOMIC, where build upon them to learn interactions between interlocutors participating in a conversation. Based on the discourse representation rich in external knowledge, two different networks composed of LSTM are responsible for tracking the individual internal state and the contextual external state to predict the emotional label of the dialogue. In the processing of external commonsense knowledge, we do not extract all the knowledge, but select the four types of knowledge information that have the greatest impact on the dialogue for modeling and fusion. Experiments show that too much or too little knowledge would degrade the performance of the model.
- We proposed an emotional dialogue model based on the encoder-decoder structure of the Seq2Seq model, which aims to generate empathetic responses to reduce users' emotional problems. In terms of emotion coding, for the defect that the word vector does not contain emotional information, the emotional word vector is formed by adding emotional information and location information according to the emotional dictionary on the basis of the word vector. We proposed a party encoder and a global encoder to generate emotional responses for dialogue generation. We separately model and monitor local emotional dialogue information and global emotional dialogue information. We separately model and monitor local emotional dialogue information and global emotional dialogue information, making it easier for the proposed model to generate emotional vocabulary sentences and maximize emotions.

5.2 Future Work

Future work will try to explore more effective methods to more accurately identify weak emotions. Abundant resources are the basis of neural network training, and external emotional resources can be used as prior knowledge for segmentation to enhance emotional feature representation. In future work, more emotional resources can be incorporated to better understand emotions. With the open-source of the diverse emotional dialogue information corpus, we are interested in incorporating human facial expressions, voices, and other multi-modal information into the model, and extracting features in all aspects to enrich the emotions in the dialogue. The development of intelligent dialogue emotion recognition systems and emotional chat robots deserve further attention in future work. Related subtasks such as multi-party emotional interaction, personality modeling, dynamic emotion tracking, emotional chatting, and emotional support conversation can form new research directions.

Bibliography

[1] H. Zhou, M. Huang, T. Zhang, X.. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory", In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[2] N. Lubis, S. Sakti, K. Yoshino and S. Nakamura, "Positive Emotion Elicitation in Chat-Based Dialogue Systems", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 866-877, 2019.

[3] Fuji Ren, Yu Wang, Changqin Quan, “TFSM-based dialogue management model framework for affective dialogue systems,” IEEJ Transactions on Electrical and Electronic Engineering, pp. 404-410, 2015.

[4] Fuji Ren, Yu Wang, Changqin Quan, “A novel factored POMDP model for affective dialogue management,” Journal of Intelligent and Fuzzy Systems, Vol. 31, pp. 127-136, 2016.

[5] Tim Althoff, Kevin Clark, and Jure Leskovec, “Large-scale analysis of counseling conversations: An application of natural language processing to mental health,” Transactions of the Association for Computational Linguistics, vol. 4, pp. 463-476, 2016.

[6] Fuji Ren, Xin Kang, “Changqin Quan, Examining Accumulated Emotional Traits in Suicide Blogs with an Emotion Topic Model,” IEEE Journal of Biomedical and Health Informatics, Vol. 20, No. 5, pp. 1384-1396, 2016.

[7] Tae-Yeun Kim, Hoon Ko, Sung-Hwan Kim, Ho-Da Kim, “Modeling of Recommendation System Based on Emotional Information and Collaborative Filtering,” in Sensors, 2021.

[8] Robinson P, El Kaliouby R, “Computation of emotions in man and machines,” Philos Trans R Soc Lond B, pp. 3441–3447, 2019.

[9] Jiawen Deng, Fuji Ren, “A Survey of Textual Emotion Recognition and Its Challenges,” IEEE Transactions on Affective Computing, 2021.

[10] Fuji Ren, “From Cloud Computing to Language Engineering, Affective Computing and Advanced Intelligence,” International Journal of Advanced Intelligence, Vol. 2, No. 1, pp. 1-14, 2010.

[11] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” in Proceedings of Interspeech, pp. 320–323, 2009.

[12] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku et al., “Emotionlines: An emotion corpus of multi-party conversations,” arXiv preprint arXiv:1802.08379, 2018.

[13] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “Dialoguernn: An attentive RNN for emotion detection in conversations,” arXiv preprint arXiv:1811.00405, 2018.

[14] S. M. Mohammad and P. D. Turney, “Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon,” in Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics, pp. 26–34, 2010.

[15] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni, “Emotion recognition from text based on automatically generated rules,” in 2014 IEEE International Conference on Data Mining Workshop. IEEE, pp. 383–392, 2014.

[16] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger, “Decision support with text-based emotion recognition: Deep learning for affective computing,” arXiv preprint arXiv:1803.06397, 2018.

[17] [17] N. Colneric and J. Demsar, “Emotion recognition on twitter: Comparative study and training a unison model,” IEEE Transactions on Affective Computing, 2018.

[18] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, Stefan Scherer, "Affect-lm: A neural language model for customizable affective text generation," arXiv preprint arXiv:1704.06851, 2017.

[19] R. Zhang, Z. Wang, and D. Mai, “Building emotional conversation systems using multi-task seq2seq learning,” In National CCF Conference on Natural Language Processing and Chinese Computing, pp. 612-621. Springer, Cham, 2017.

[20] Rosalind W Picard, “Affective computing: from laughter to ieee,” IEEE Transactions on Affective Computing, pp. 11–17, 2010.

[21] Andrew Ortony and Terence J. Turner, “What’s basic about basic emotions?,” Psychological review, pp. 315-331, 1990.

[22] P Ekman, W V Friesen, M O'Sullivan, A Chan, I Diacoyanni-Tarlatzis, K Heider, R Krause, W A LeCompte, T Pitcairn, P E Ricci-Bitti, et al, “Universals and cultural differences in the judgments of facial expressions of emotion,” Journal of personality and social psy-chology, pp. 712-717, 1987.

[23] Schuller B W, “Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends,” Communications of the ACM, pp. 90- 99, 2018.

[24] J. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, pp. 1161– 1178, 12 1980.

[25] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” Current Psychology, vol. 14, no. 4, pp. 261–292, Dec 1996.

[26] Heise, D. R, “Enculturating Agents with Expressive Role Behavior,” In R. Trappl & S. Payr (Eds), Agent Culture: Designing Human-Agent Interaction in a Multicultural World, pp. 127-142, 2004.

[27] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances,” IEEE Access, vol. 1, no. 1, pp. 100 943 – 100 953, 2019.

[28] Weizhou Shen, Siyue Wu, Yunyi Yang and Xiaojun Quan, “Directed Acyclic Graph Network for Conversational Emotion Recognition,” arXiv preprint arXiv: 2105.12907, 2021.

[29] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh, “Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation,” arXiv preprint arXiv:1908.11540, 2019.

[30] Peixiang Zhong, Di Wang, and Chunyan Miao, “Knowledge-enriched transformer for emotion detection in textual conversations,” In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 165–176, 2019.

[31] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, “COSMIC: COmmonSense knowledge for eMotion identification in conversations,” In Findings of the Association for Computational Linguistics: EMNLP 2020, Online, pp. 2470–2481, 2020.

[32] M. Eric and C. D. Manning, "Key-value retrieval networks for task-oriented dialogue," arXiv preprint arXiv:1705.05414, 2017.

[33] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011.

[34] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang and Minlie Huang, "Towards emotional support dialog systems," arXiv preprint arXiv:2106.01144, 2021.

[35] N. Lubis, S. Sakti, K. Yoshino and S. Nakamura, "Positive Emotion Elicitation in Chat-Based Dialogue Systems", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 866-877, 2019.

[36] K. Colby, "Modeling a paranoid mind", Behavioral and Brain Sciences, vol. 4, no. 4, pp. 515-534, 1981.

[37] Fazel Keshtkar, and Diana Inkpen, "A pattern-based model for generating text to express emotion," International Conference on Affective Computing and Intelligent Interaction, Springer, Berlin, Heidelberg, 2011.

[38] M. Skowron, "Affect listeners: Acquisition of affective states by means of conversational systems," Development of Multimodal Interfaces: Active Listening and Synchrony, Springer, Berlin, Heidelberg, pp. 169-181, 2010.

[39] Lei Shen, and Yang Feng, "CDL: Curriculum dual learning for emotion-controllable response generation," arXiv preprint arXiv:2005.00329, 2020.

[40] Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu, "A multi-turn emotionally engaging dialog model," arXiv preprint arXiv:1908.07816, 2019.

[41] Asghar N, Poupart P, Hoey J, et al, "Affective neural response generation," European Conference on Information Retrieval, Springer, Cham, 2018.

[42] Weiyan Shi, and Yu Zhou, "Sentiment adaptive end-to-end dialog systems," arXiv preprint arXiv:1804.10731, 2018.

[43] Yunlong Liang, Fandong Meng, Ying Zhang, Jinan Xu, Yufeng Chen, Jie Zhou, "Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation," arXiv preprint arXiv:2012.04882, 2020.

[44] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin, "A neural probabilistic language model," Journal of machine learning research, 3, pp.1137-1155, 2003.

[45] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, "Natural language processing (almost) from scratch," Journal of machine learning research 12.ARTICLE, pp. 2493-2537, 2011.

[46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, 2013.

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

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," In Advances in neural information processing systems, pp. 5998-6008, 2017.

[49] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher, "Learned in translation: Contextualized word vectors," arXiv preprint arXiv:1708.00107, 2017.

[50] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, "Deep contextualized word representations," arXiv preprint arXiv:1802.05365, 2018.

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," 2018.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," Proceedings of NAACL-HLT, 2019.

[53] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, ‘‘K-BERT: Enabling language representation with knowledge graph,’’ in Proc. 34th AAAI Conf. Artif. Intell., pp. 2901–2908, 2019.

[54] B. Y. Lin, X. Chen, J. Chen, and X. Ren, ‘‘KagNet: Knowledge-aware graph networks for commonsense reasoning,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), pp. 2822–2832, 2019.

[55] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, ‘‘ERNIE: Enhanced language representation with informative entities,’’ in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, pp. 1441–1451, 2019.

[56] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, ‘‘COMET: Commonsense transformers for automatic knowledge graph construction,’’ arXiv:1906.05317, 2019.

[57] Daniil Larionov, Artem Shelmanov, Elena Chistova, Ivan Smirnov, "Semantic role labeling with pretrained language models for known and unknown predicates," Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019.

[58] K. Xu, H. Wu, L. Song, H. Zhang, L. Song, and D. Yu, ‘‘Conversational semantic role labeling,’’ IEEE/ACM Trans. Audio, Speech, Language Process., vol. 29, pp. 2465–2475, 2021.

[59] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. Morency, "Context-dependent sentiment analysis in user-generated videos," Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017.

[60] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, pp. 1735–1780, Nov. 1997.

[61] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, ‘‘Empirical evaluation of gated recurrent neural networks on sequence modeling,’’ arXiv:1412.3555, 2014.

[62] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, ‘‘Did the model understand the question?’’ arXiv:1805.05492, 2018.

[63] R. Jia and P. Liang, ‘‘Adversarial examples for evaluating reading comprehension systems,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., pp. 1–11, 2017.

[64] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, ‘‘Deep semantic role labeling: What works and what’s next,’’ in Proc. 55th Annual Meeting Assoc. Comput. Linguistics, pp. 473–483, 2017.

[65] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, ‘‘Semantics-aware BERT for language understanding,’’ in Proc. AAAI Conf. Artif. Intell., pp. 9628–9635, 2020.

[66] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, ‘‘Atomic: An atlas of machine commonsense for if-then reasoning,’’ in Proc. AAAI Conf. Artif. Intell., vol. 33, pp. 3027–3035, 2019.

[67] S. Bird, ‘‘NLTK: The natural language toolkit,’’ in Proc. COLING/ACL Interact. Presentation Sessions, pp. 63–70, 2006.

[68] C. Navarretta, ‘‘Mirroring facial expressions and emotions in dyadic conversations,’’ in Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC), pp. 469–474, 2016.

[69] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, ‘‘IEMOCAP: Interactive emotional dyadic motion capture database,’’ Lang. Resour. Eval., vol. 42, no. 4, p. 335, Dec. 2008.

[70] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, ‘‘DailyDialog: A manually labelled multi-turn dialogue dataset,’’ arXiv:1710.03957, 2017.

[71] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, ‘‘MELD: A multimodal multi-party dataset for emotion recognition in conversations,’’ arXiv:1810.02508, 2018.

[72] Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), pp. 1746–1751, 2014.

[73] Serban I V, Lowe R, Charlin L, ‘‘Generative deep neural networks for dialogue: A short review.’’ arXiv preprint arXiv: 1611.06216, 2016.

[74] John D. Mayer, Peter Salovey, David R. Caruso, and Lillia Cherkasskiy, "What is emotional intelligence," Emotional development and emotional intelligence: Educational implications, 3: 31, 1997.

[75] Prendinger H, Mori J, Ishizuka M, “Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game,” International journal of human-computer studies, no. 2, pp. 231-245, 2005.

[76] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering, 22,no. 10, pp. 1345-1359, 2009.

[77] Peng Y, Fang Y, Xie Z, et al, "Topic-enhanced emotional conversation generation with attention mechanism," Knowledge-Based Systems ,163, pp. 429-437, 2019.

[78] Shantala, Roman, Gennadiv Kyselov, and Anna Kyselova, "Neural dialogue system with emotion embeddings," 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC), IEEE, 2018.

[79] Daniel Jurafsky, and James H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," 2nd Edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009.

[80] Robert Plutchik, “Emotions and psychotherapy: A psychoevolutionary perspective,” Emotion, psychopathology, and psychotherapy, Elsevier, pp. 3–41, 1990.

[81] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, Dan Jurafsky, “Deep Reinforcement Learning for Dialogue Generation,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas: Association for Computational Linguistics, pp. 1192–1202, Nov. 2016.

[82] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, Dan Jurafsky, “Adversarial Learning for Neural Dialogue Generation,” In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark: Association for Computational Linguistics, pp. 2157–2169, Sept. 2017.

[83] Oriol Vinyals and Quoc V. Le, “A Neural Conversational Model,” CoRR abs/1506.05869, arXiv: 1506.05869, 2015.

[84] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation,” Computational Linguistics 19.2, pp. 263–311, 1993.

[85] Alexander M. Rush, Sumit Chopra, and Jason Weston, “A Neural Attention Model for Abstractive Sentence Summarization,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, pp. 379–389, September 2015.

[86] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, “Sequence to Sequence Learning with Neural Networks,” Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, pp. 3104–3112, December 2014.

[87] Shuohang Wang and Jing Jiang, “Machine Comprehension Using MatchLSTM and Answer Pointer,” 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings, April 2017.

[88] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, Xiaoming Li, “Neural Generative Question Answering,” Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, pp. 2972–2978, July 2016.

[89] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, Bill Dolan, “A Persona-Based Neural Conversation Model,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Berlin, Germany, Volume 1: Long Papers, August 2016.

[90] Xiaoyu Shen, Hui Su, Shuzi Niu, Vera Demberg, “Improving Variational Encoder-Decoders in Dialogue Generation,” Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, pp. 5456–5463, February 2018.

[91] Tiancheng Zhao, Ran Zhao, Maxine Eskenazi, “Learning Discourselevel Diversity for Neural Dialog Models using Conditional Variational Autoencoders,” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, Volume 1: Long Papers, pp. 654–664, 2017.

[92] Lantao Yu, Weinan Zhang, Jun Wang, Yong Yu, “SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient,” Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, pp. 2852–2858, February 2017.

[93] Yi-Lin Tuan and Hung-yi Lee, “Improving Conditional Sequence Generative Adversarial Networks by Stepwise Evaluation,” IEEE/ACM Trans. Audio, Speech & Language Processing 27.4, pp. 788–798, 2019.

[94] Nal Kalchbrenner, and Phil Blunsom, "Recurrent continuous translation models," Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.

[95] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[96] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba, "Addressing the rare word problem in neural machine translation," arXiv preprint arXiv:1410.8206, 2014.

[97] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio, "On using very large target vocabulary for neural machine translation," arXiv preprint arXiv:1412.2007, 2014.

[98] Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton, "Grammar as a foreign language," Advances in neural information processing systems 28, pp. 2773-2781, 2015.

[99] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "Show and tell: A neural image caption generator," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164, 2015.

[100] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, pp. 1735–1780, Nov. 1997.

[101] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[102]Zhenqiao Song, Xiaoqing Zheng, Lu Liu Mu Xu and Xuanjing Huan, “Generating Responses with a Specific Emotion in Dialog,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. (ACL), Florence, Italy, pp. 3685-3695, 2019.

[103]Qiang Lu, Zhenfang Zhu, Fuyong Xu, Qiangqiang Guo, “Research on Bi-LSTM Chinese sentiment classification method based on grammar rules,” Data Analysis and Knowledge Discovery, pp. 3(11), 99-107, 2019.

[104]J. Carrillo-de-Albornoz and L. Plaza, "An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification", Journal of the American Society for Information Science and Technology, vol. 64, no. 8, pp. 1618-1633, 2013.

[105]Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin,. "Convolutional sequence to sequence learning," International Conference on Machine Learning, PMLR, 2017.

[106]Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou, "Improving multi-turn dialogue modelling with utterance rewriter," arXiv preprint arXiv:1906.07004, 2019.

[107]Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002.

[108]Lin, Chin-Yew, "Rouge: A package for automatic evaluation of summaries," Text summarization branches out, 2004.

[109]Rus, Vasile, and Mihai Lintean, "An optimal assessment of natural language student input using word-to-word similarity metrics," International Conference on Intelligent Tutoring Systems, Springer, Berlin, Heidelberg, 2012.

[110]John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu, "Towards universal paraphrastic sentence embeddings," arXiv preprint arXiv:1511.08198, 2015.

[111]Gabriel Forgues, Joelle Pineau, Jean-Marie Larcheveque, Réal Tremblay, "Bootstrapping dialog systems with word embeddings," Nips, modern machine learning and natural language processing workshop, Vol. 2, 2014.

[112]Kiros, Ryan, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, "Skip-thought vectors," In Advances in neural information processing systems, pp. 3294-3302, 2015.