EPFL

Thèse n° 9311

Towards Empathetic Open-Domain Chatbots

Présentée le 25 novembre 2022

à la Faculté informatique et communications
Groupe SCI IC PFP
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Yubo XIE

Acceptée sur proposition du jury

Dr R. Boulic, président du jury
Dr P. Pu Faltings, directrice de thèse
Prof. M. Huang, rapporteur
Prof. X. Ma, rapporteuse
Prof. R. West, rapporteur

École polytechnique fédérale de Lausanne

2022

Not to be able to carry one's learning into practice is to be distressed.
—Zhuangzi

To my family and friends...

Acknowledgements

First and foremost, I want to express my sincere gratitude to my supervisor, Dr. Pearl Pu, for her invaluable advice, unwavering support, and constant encouragement on my academic research and daily life. Her guidance made this thesis possible, and also carried me through all the stages from a student who barely knew how to do research, to a Ph.D. graduate who has successfully defended his work. I also want to thank the Swiss National Science Foundation (Grant No. 200021_184602) for providing funding for the research work done in this thesis.

I would like to thank all the jury members, Dr. Ronan Boulic, Prof. Robert West, Prof. Minlie Huang, and Prof. Xiaojuan Ma, for attending my oral exam and providing valuable comments.

I would also like to take this opportunity to thank all the members of the HCI Group, including my colleagues Igor, Ekaterina, Anuradha, and Kavous, who have given me their kindest help and support and made my study and life at EPFL a wonderful time. I also want to thank all the master's students that I have worked with: Francis, Quentin, Ayyoub, Oussama, Yan, Junze, Xiaoyan, Yueran, Zhechen, and Jingrong, from whom I have found a lot of inspiration.

My thanks also go to Didier, who was my manager when I did a six-month internship at Apple last year. He is a wonderful mentor and friend, from whom I have learned vital skills of coding and prototyping, and the boat trips that he took me on Lake Geneva are definitely my unforgettable memories.

Over the past few years, I was fortunate to meet a group of friends with whom I shared a lot of joyful moments of traveling, hiking, hanging out, etc. I want to give my special thanks to Fanyuan, Xiaoxue, Mo, Junze, Yufan, Alexandre, Long, Xunhui, Jiancheng, Jingrong, Luiz, Shaobo, Chenkai, Changyang, and many others. I also want to thank Yuze, who has been my close friend since high school, for sharing countless fun time in video games with me, remotely in New Zealand. Thank you all for giving me wonderful memories throughout my Ph.D. study.

Finally, my deepest gratitude goes to my parents, who have been unconditionally supporting me for the whole of my life. Without their love and encouragement, I would not be able to get such motivation on my journey towards a life filled with happiness. I also want to thank my extended family members for their support during my Ph.D. study, especially Tongyang, my cousin, for being there with me in times of both joy and sadness.

Lausanne, October 25, 2022 Yubo Xie

Abstract

The research community of dialog generation has been interested in incorporating emotional information into the design of open-domain dialog systems ever since neural networks (sequence-to-sequence models in particular) were adopted for modeling dialogs. The major objective is to generate emotionally richer responses or to make the conversational agent sound more empathetic, which entails recognizing and understanding the user's affective states, and then replying with the appropriate emotion. However, there are a number of difficulties encountered when creating such an empathetic chatbot. Some of the existing models explicitly need an emotion label as input in order to produce responses of that particular emotion, which is impractical in real-world scenarios. Others assume manually defined rules such as following or reversing the user's emotion, but psychological literature has not confirmed such rules to be universally appropriate. Moreover, they ignore the subtle emotion exchanges embedded in human-human conversations, where listeners often exhibit certain empathetic intents that are less emotional. To train a chatbot to convey such subtle emotions and intents, we need a large-scale dialog dataset that is properly labeled. Finally, it is also desirable to explicitly represent such emotional interactions found in people's daily conversations (part of so-called social intelligence) using knowledge graphs, in order to facilitate the development of chatbots. In this thesis, we propose novel solutions to these problems. First, we introduce MEED, a multi-turn emotionally engaging dialog model that learns emotion interactions directly from data, without the need of specifying emotion labels or developing heuristic rules. Then, we present MEED2, the second generation of the MEED model, which is more controllable and interpretable, and is capable of generating responses that have finer-grained emotions and empathetic intents. We also curated EDOS, a large-scale dialog dataset labeled with 32 emotions and 8 empathetic response intents, plus the *neutral* category. We adopted a semi-supervised learning framework to grow a seed dataset manually labeled by crowdsourcing workers, while iteratively training an emotion/intent classifier, which was used to label the whole dataset. Finally, we present AFEC, a knowledge graph capturing social intelligence found in casual conversations, which reveals how people communicate with each other in day-to-day social environments. To show the utility of AFEC, we built a retrieval-based dialog model solely based on it, and the experimental results show that the dialog model can produce much more diverse responses, yet still being emotionally appropriate. We conclude the thesis by discussing our findings, lessons learned, and some future directions worth exploration.

Keywords: affective computing, emotion, empathy, emotion recognition, dialog systems, chatbots, open-domain dialog systems, dialog generation, neural networks, deep learning, semi-

supervised learning, knowledge graph, information retrieval, human computation, crowdsourcing

Résumé

La communauté de recherche sur la génération de dialogues s'est intéressée à l'incorporation d'informations émotionnelles dans la conception de systèmes de dialogue à domaine ouvert depuis que les réseaux neuronaux (modèles de séquence à séquence en particulier) ont été adoptés pour modéliser des dialogues. L'objectif principal est de produire des réponses plus riches en émotions ou de rendre l'agent conversationnel plus empathique, ce qui implique de reconnaître et de comprendre les états affectifs de l'utilisateur, puis de répondre avec l'émotion appropriée. Cependant, la création d'un tel chatbot empathique est confrontée à plusieurs défis. Certains des modèles existants ont explicitement besoin d'une étiquette d'émotion en entrée afin de produire des réponses de cette émotion particulière, ce qui n'est pas pratique dans les scénarios du monde réel. D'autres supposent des règles définies manuellement, comme suivre ou inverser l'émotion de l'utilisateur, mais la littérature psychologique n'a pas confirmé que de telles règles sont universellement appropriées. De plus, elles ignorent les échanges subtils d'émotions intégrés dans les conversations entre humains, où les auditeurs manifestent souvent certaines intentions empathiques qui sont moins émotionnelles. Pour entraîner un chatbot à transmettre ces émotions et intentions subtiles, nous avons besoin d'un ensemble de données de dialogue à grande échelle correctement étiqueté. Enfin, il est également souhaitable de représenter explicitement ces interactions émotionnelles que l'on trouve dans les conversations quotidiennes des gens (une partie de ce que l'on appelle l'intelligence sociale) en utilisant des graphes de connaissances, afin de faciliter le développement de chatbots. Dans cette thèse, nous proposons des solutions originales à ces problèmes. Tout d'abord, nous présentons MEED, un modèle qui apprend les interactions émotionnelles directement à partir des données, sans avoir besoin de spécifier des étiquettes d'émotion ou de développer des règles heuristiques, et puis génère des dialogues à plusieurs tours qui suscitent des émotions adéquates. Ensuite, nous présentons MEED2, la deuxième génération du modèle MEED, qui est plus contrôlable et interprétable, et est capable de générer des réponses qui ont des émotions plus raffinées et des intentions empathiques. Nous avons également créé EDOS, un ensemble de données de dialogue à grande échelle étiqueté avec 32 émotions et 8 intentions de réponse empathique, plus la catégorie *neutral*. Nous avons adopté un cadre d'apprentissage semi-supervisé pour développer un jeu de données d'origine étiqueté manuellement par des travailleurs du crowdsourcing, tout en formant itérativement un classificateur d'émotions/intents, qui a été utilisé pour étiqueter l'ensemble du jeu de données. Enfin, nous présentons AFEC, un graphe de connaissances capturant l'intelligence sociale trouvée dans les conversations occasionnelles, qui révèle comment les gens commu-

niquent entre eux dans des environnements sociaux quotidiens. Pour montrer l'utilité d'AFEC, nous avons construit un modèle de dialogue basé sur la recherche uniquement sur cette base, et les résultats expérimentaux montrent que le modèle de dialogue peut produire des réponses beaucoup plus diverses, tout en restant émotionnellement appropriées. Nous concluons la thèse en discutant de nos résultats, des leçons apprises, et de certaines directions futures à explorer.

Mots-clés : informatique affective, émotion, empathie, reconnaissance des émotions, systèmes de dialogue, chatbots, systèmes de dialogue à domaine ouvert, génération de dialogues, réseaux neuronaux, apprentissage profond, apprentissage semi-supervisé, graphe de connaissances, recherche d'informations, calcul humain, crowdsourcing

Contents

List of Figures

List of Tables

1 Introduction

Conversation, or dialog, is the first language skill that humans acquire as children. It is so unique that when people learn a new language, it is the foremost focus. It lies at the heart of human-human interactions on a daily basis, including social chitchat, ordering food in restaurants, asking customer service for help, etc. In fact, it plays such an important role in the history of human beings that people have had this constant eager to talk to even the inanimate objects that are created in literature, arts, and movies. Since the beginning of the information revolution, people have been imagining the scenario of conducting natural language conversations with machines, which is probably one of the most fascinating pictures that artificial intelligence has depicted. The Turing test (Turing, 1950), perhaps the most influential (at the same time also widely criticized) concept in the field of artificial intelligence, indeed aims at testing the intelligence of a machine by letting a human evaluator conduct natural language conversations with it.

In order to converse with machines, we rely on so-called dialog systems, or chatbots,[1] that can produce responses given users' input. Generally speaking, dialog systems fall into two major categories: task-oriented dialog systems and open-domain dialog systems. Task-oriented dialog systems are often domain specific and are used to help users accomplish certain tasks, e.g., making restaurant reservations, booking flights, providing customer service, etc. Voice assistants on modern electronic devices (e.g., Siri, Google Assistant, Alexa, and Cortana) usually fall into this category. On the contrary, open-domain dialog systems are designed to have extended conversations with humans, often mimicking the open-domain or "chitchat" characteristics of human-human conversations. They can provide effective emotional support for socially excluded individuals (De Gennaro et al., 2020), help cope with loneliness and social isolation during the COVID-19 pandemic (Dosovitsky and Bunge, 2021; Jiang et al., 2022), and provide social companions for the elderly (Vardoulakis et al., 2012; Wanner et al., 2017). Although we limit the scope of this thesis to only open-domain chatbots, the connection between task-oriented dialog systems and open-domain dialog systems goes deeper than one

[1] In this thesis, we interchangeably use the terms *dialog system*, *conversational agent*, and *chatbot*, to refer to the kind of software that is able to conduct natural language conversations with humans. Note that in some literature, *chatbots* specifically mean open-domain dialog systems.

would expect. In fact, open-domain chatbots can also be designed so as to enhance the task-oriented agents and make them more natural and engaging (Zhao et al., 2017a; Lin et al., 2021; Sun et al., 2021; Young et al., 2022). Moreover, recent language models such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) that were trained on massive open-domain data have shown to also perform well in task-specific scenarios. This suggests that, provided with enough training data, open-domain dialog models could also deal with specific tasks.

Compared with task-oriented dialog systems, open-domain chatbots are usually more challenging to develop, due to the fact that the goal to be optimized in open-domain dialogs is open-ended and not clearly defined. Early influential chatbots such as ELIZA (Weizenbaum, 1966) and PARRY (Colby et al., 1971) (though they were also used for practical purposes like psychological study) were based on manually defined patterns/rules that transform an input sentence into a response by matching keywords. With the advent of neural networks and large-scale training, data-driven approaches became more prevalent. Ritter et al. (2011) first considered using statistical machine translation approach to generate responses to Twitter posts. Following work (Vinyals and Le, 2015; Shang et al., 2015; Serban et al., 2016) generalized this idea to an encoder-decoder architecture that encodes the input utterance (possibly along with the dialog context) into a fixed representation, based on which it decodes the response. Due to the pre-trained language models being popular recently, there is also work on building open-domain dialog systems using a language model architecture (Wolf et al., 2019; Lin et al., 2020; Zhang et al., 2020b; Thoppilan et al., 2022). This is usually done by concatenating all the history utterances into one sequence and letting the language model continue to generate the response.

Though the advent of neural networks brought huge advancements to open-domain dialog generation, we still face several key challenges. For example, generative neural-based dialog models are known to suffer from the diversity issue, where the generated responses are often generic and uninformative like "I don't know." There already exists some work that focuses on addressing this diversity issue (Li et al., 2016a,c; Zhao et al., 2017b; Du et al., 2018; Vijayakumar et al., 2018). Some other work attempted at keeping the persona of the chatbot consistent (Li et al., 2016b; Zhang et al., 2018; Qian et al., 2018; Zhang et al., 2019a,b), and others tried to ground the response generation process with external knowledge sources (Ghazvininejad et al., 2018; Zhou et al., 2018b; Dinan et al., 2019; Zhao et al., 2020; Li et al., 2022b). One important aspect when developing open-domain chatbots is to make them empathetic. Empathy is considered to be an innate ability of human beings (Roth-Hanania et al., 2011) and plays an important role in people's social communication (Valente, 2016). The definition of empathy has been debatable since the introduction of the word into the English language (Lanzoni, 2018; Hall and Schwartz, 2019). Nevertheless, it is widely accepted that emotion plays a vital role in empathy, and one major component of empathy is emotional empathy (Mehrabian and Epstein, 1972), which is the ability to respond with an appropriate emotion to another person's mental states (Rogers et al., 2007). It has been shown that integrating empathy into dialog systems could improve user experience for human-computer interaction (Liu and Picard, 2005). Figure 1.1 shows how an empathetic chatbot responds to a user who is in a low mood

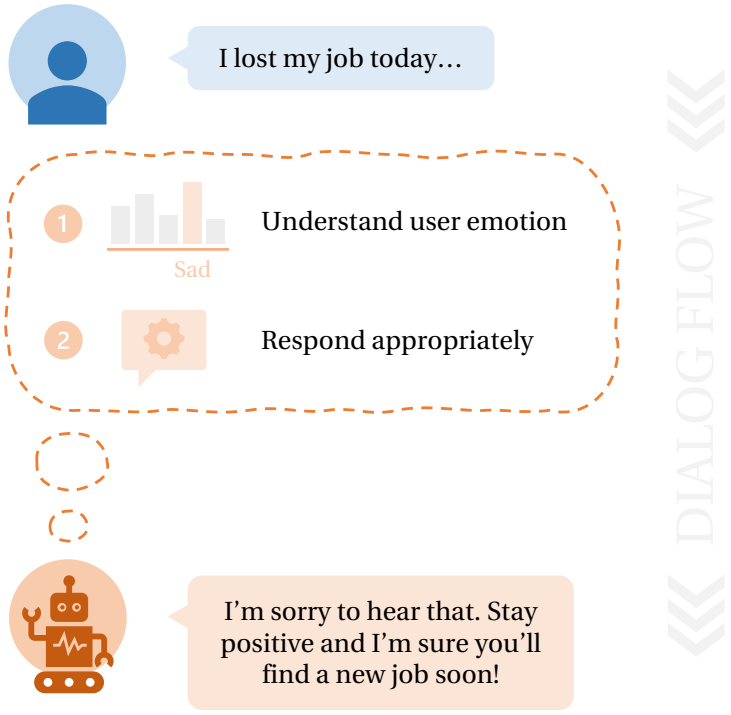


Figure 1.1: How an empathetic chatbot responds to a user in a low mood: (1) the chatbot first tries to understand the user's current mental state (in this case, *sad*); (2) given the user's emotion, the chatbot composes the response accordingly so that it is emotionally appropriate and shows sympathy towards the user.

due to job loss. The process usually consists of two steps: (1) the chatbot tries to understand the user's current mental state (in this case, *sad*); (2) given the user's emotion, the chatbot composes the response accordingly so that it is emotionally appropriate and shows sympathy towards the user. In this thesis, we are mainly interested in investigating the ways to develop open-domain chatbots that are more empathetic, including the exploration of different model architectures, the curation of large-scale emotional dialog datasets, and the evaluation of open-domain empathetic dialog systems.

1.1 Research Motivations

The incorporation of affect information in natural language dialogs has considerable advantages across a wide range of application domains. As Reeves and Nass (1996) have pointed out, people respond to computers in the same way as they do to real human beings. In prior research on human-computer interaction, Klein et al. (2001) discovered that computer-initiated emotional support can reduce users' dissatisfaction caused by a computer system by delivering feedback on emotional content together with sympathy and empathy. Hu et al. (2018) developed a customer support neural chatbot that might potentially replace human customer service representatives on social media platforms by creating dialogs with empathic

and impassioned tones similar to those of people. Participants in a qualitative study (Zamora, 2017) exhibited interest in chatbots that may meet users' emotional needs by listening intently and offering encouragement. Even other participants mentioned that a chatbot is perfect for asking a human about uncomfortable or sensitive questions. Finally, Bickmore and Picard (2005) demonstrated that, even after four weeks of engagement, a relational agent with conscious social-emotional abilities was more respected, liked, and trusted than an equal task-oriented agent.

The open-domain dialog generation community is quite excited about recent advancements in neural language modeling. The success of sequence-to-sequence (seq2seq) learning (Sutskever et al., 2014; Cho et al., 2014) in machine translation has encouraged researchers to use the recurrent neural network (RNN) encoder-decoder structure for response generation (Vinyals and Le, 2015). Additionally, researchers are beginning to include affect information in neural dialog models. While enhancing the emotional richness of the responses appears to be a fundamental subject, existing approaches largely go in two directions. In one, the machine expressly needs an emotion label as input in order to produce sentences with that specific emotion label or kind (Zhou et al., 2018a; Huang et al., 2018; Zhou and Wang, 2018; Colombo et al., 2019; Song et al., 2019; Shen and Feng, 2020). The major goal of a different group of work is to create manual rules, either explicit or implicit, to decide the emotion state for the response to be generated (Asghar et al., 2018; Li and Sun, 2018; Zhong et al., 2019; Lin et al., 2019; Li et al., 2022a). Both methods call for an emotion label as input (either provided or manually created), which may not be feasible in real dialog scenarios. Furthermore, to the best of our knowledge, there are no definitive guidelines for emotional interactions in the psychology and social science literature.

The fact that present emotional dialog systems miss the delicate exchanges caught in human conversations, where the listener frequently displays empathetic intents that are more neutral, is another issue. Welivita and Pu (2020) found that listeners are far more inclined to ask about someone else's sadness or anger than they are to convey their own comparable or opposing emotions. As a result, it is essential to explicitly include these additional empathetic response intents in dialog system design. Existing neural dialog systems adopt an empathetic dialog dataset that either has no *neutral* category (Rashkin et al., 2019), or the *neutral* category is a conglomerate of intents that cannot be clearly defined. This is why this category is often called *other*, which shows it is not sufficiently treated (Chatterjee et al., 2019; Li et al., 2017). While Xu et al. (2018) proposed to model open-domain dialog generation as the selection of dialog acts that control the generation of responses, they did not specifically focus on the generation of empathetic dialogs.

In order to train neural chatbots to generate emotionally appropriate responses, it is also desirable to curate a large-scale emotion and intent labeled dataset. Annotating such a large-scale dataset requires expensive human labor, and given the fine-grained emotion and intent labels, human labeling is more difficult and error-prone than the coarser-grained *angry-happy-sad* emotion categories. These two factors make curating such a dataset technically challenging.

As a result, existing manually labeled emotional dialog datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and DailyDialog (Li et al., 2017) are smaller in scale and contain only a limited set of emotions (emotions derived from basic emotion models such as the Ekman's (Ekman, 1992)). Most importantly, existing datasets fail to distinguish between *neutral* and *questioning*, or any of the other eight empathetic response intents proposed by Welivita and Pu (2020). When the utterance is not emotional, they lump everything together under the broad label *neutral* or *other*. However, there are key components to constructing empathic dialogs, such as *questioning*, *agreeing*, *acknowledging*, *sympathizing*, *encouraging*, *consoling*, *suggesting*, and *wishing*. These eight response intents, which we refer to as the plus categories in our work, are novel and contribute to the model's ability to learn important response patterns in the data.

Last but not least, empathy is seen as a crucial component of social intelligence, one type of commonsense knowledge. According to Daniel (2006), social intelligence is made up of two components: *social awareness*, which is the capacity to comprehend the complexity of social situations and to understand the feelings of others, and *social facility*, which is the capacity to carry out smooth and effective interactions based on social awareness. Even for humans, let alone AI systems, having desirable social intelligence is not always simple or obvious. Previous work has attempted to explicitly express various commonsense knowledge in knowledge graphs so that AI systems can use them as external resources to perform various commonsense reasoning tasks. This ranges from knowledge about common concepts (Speer et al., 2017; Tandon et al., 2017) to inferences over common events (Sap et al., 2019; Zhang et al., 2020a). However, to the best of our knowledge, no work has yet been done to build a knowledge graph which captures the social intelligence that people display when conducting social conversations with each other in day-to-day social environments. By examining the knowledge graph, one could learn how to appropriately respond to another person with desired emotions and intents. The knowledge graph can also serve as an external resource to facilitate the development of both open-domain and task-oriented dialog systems.

1.2 Research Agenda

The goal of this thesis is to study how to build an open-domain dialog system that is more empathetic. Figure 1.2 illustrates the key components or steps when building such a chatbot, i.e., dialog data, emotion model, generation model, and automatic/human evaluation (aligned with the studied research problems and the corresponding thesis chapters). Several challenges arise from the process of developing an empathetic chatbot, which we describe in detail:

- **Learning emotion interactions directly from data.** Understanding other people's feelings or mental states and responding in an emotionally appropriate manner is one of the key elements of empathy. The ability to enable seamless and appropriate emotion interactions between the user and the machine is therefore a key component of a successful empathetic chatbot. In order to describe the emotional state of the response

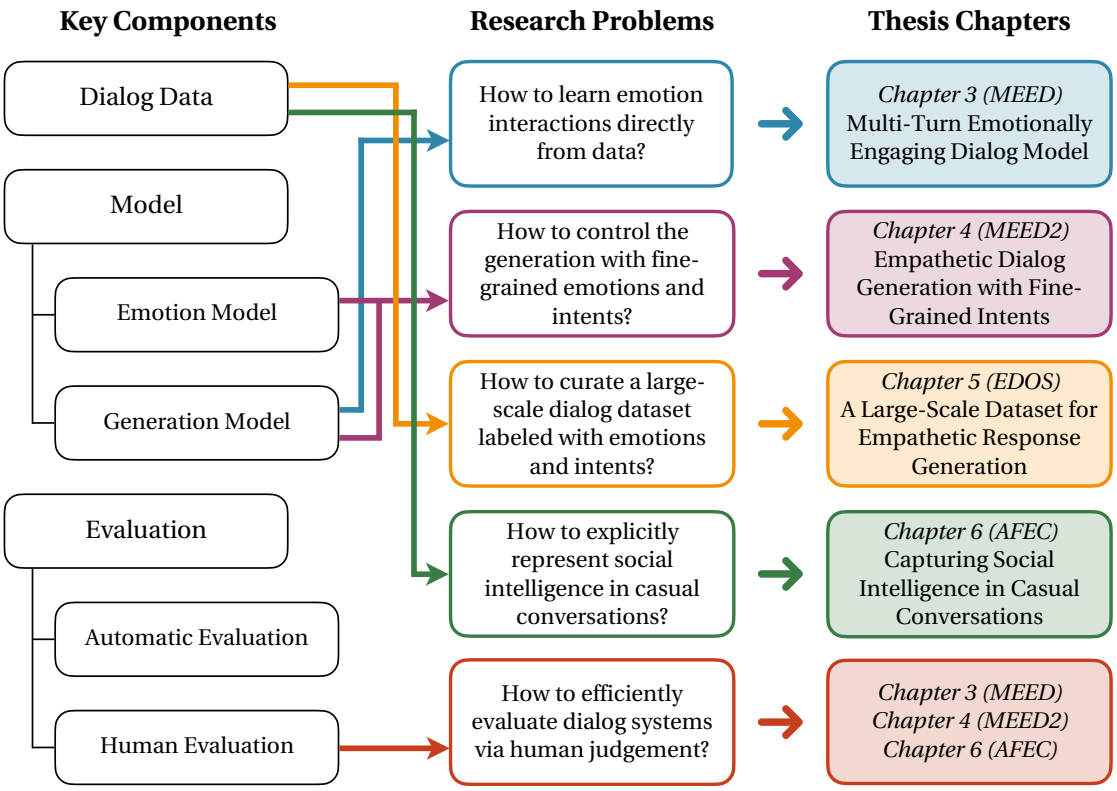


Figure 1.2: An overview of the key components of building an empathetic dialog system aligned with the investigated research problems and the corresponding thesis chapters.

to be generated, several existing emotional dialog systems ask the user to specify an emotion label, which is impractical in real-world situations. Others develop manual rules for how the model exchanges emotions with the user (e.g., following/reversing the user emotion, maximizing/minimizing the emotion information in the response, etc.), but these rules have not been supported by psychological literature. Therefore, we suggest a data-driven approach that learns these emotion interaction patterns directly from human-human conversations. We are interested in the following research questions: How to encode the emotion information of the user's input along with the history utterances? How to effectively combine the encoded emotion information with the textual input to produce the response? How to adequately validate that the model has learned useful embeddings to distinguish different emotions?

- **Controlling the response generation process with fine-grained emotions and intents.** Most of the existing emotion labeled dialog datasets have emotion categories that are just coarse-grained, and they do not have the *neutral* category, or the *neutral* category is just a jumble of intents that are not clearly defined. Thus, the nuanced emotion interactions in human conversations, where the listener frequently exhibits empathetic intents that are more neutral, are missed by the current emotional dialog systems. Furthermore, it is also desirable to precisely control the emotion state of the generated

response, so that the model has more interpretability. We are interested in the following research questions: How to explicitly learn the fine-grained emotion interactions from the dialog data so that the response generation process is more interpretable and can be controlled precisely by a specific emotion? How to obtain sufficient dialog data that are labeled with fine-grained emotions as well as empathetic response intents?

- **Curating a large-scale dialog dataset labeled with emotions and intents.** It is expensive to curate a large-scale dialog dataset with accurate emotion labeling. Existing datasets that were manually created by humans are often limited in size (Rashkin et al., 2019). Given the fine-grained emotion and intent labels, human labeling is more difficult and error-prone than the coarser-grained emotion categories, which is why annotating such a large-scale dataset requires expensive human labor. On the other hand, automatic labeling using a trained classifier is fast and low-cost, but the accuracy cannot be guaranteed. Therefore, we suggest a hybrid approach that combines the advantages of both human labeling and automatic labeling. We are interested in the following research questions: How to obtain a seed dataset with accurate emotion and intent labels? How to properly grow the seed dataset and increase the accuracy of a trained classifier? How to find similar dialogs to label when growing the seed dataset?

- **Explicitly representing social intelligence in casual conversations.** Empathy is also an important component of social intelligence. It has been attempted in earlier work to explicitly represent different commonsense knowledge in knowledge graphs so that AI systems can use them as external resources to carry out different commonsense reasoning tasks. However, a knowledge graph that represents the social intelligence that individuals exhibit when engaging in social conversations with one another in regular social settings has not yet been developed. Therefore, we suggest curating such a knowledge graph of social intelligence from casual conversations found online. We are interested in the following research questions: How to find reliable sources of casual conversations that cover a wide range of topics in daily social environments? How to preprocess the dialog data and filter out unstructured utterances? How to calculate the semantic similarity between utterances and cluster them into nodes? How to label the nodes with proper emotions and intents?

- **Efficiently evaluating dialog systems via human judgement.** The evaluation of dialog model is still an unsolved issue. Most existing work still adopts automatic evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), which were originally designed for tasks like machine translation and text summarization. However, Liu et al. (2016) found that these metrics tend to correlate poorly with human judgement. To this end, the community of dialog generation still rely on human evaluation to effectively evaluate the dialog models, and a fundamental desire is to obtain high quality ratings in a short period of time while maintaining a relatively low cost at the same time. While evaluating our chatbots, we pay close attention to the following questions: How to make sure that the evaluators are well

instructed, the dialogs are easy to read, and the tasks are straightforward to accomplish? How to effectively screen the candidate workers and filter out robots and free-riders? How to maintain a relatively low cost while having enough number of dialogs evaluated?

1.3 Main Contributions

Our main contributions in this thesis are:

- We present MEED, a novel emotion-tracking dialog generation model that learns the emotional interactions directly from the data. This approach is free of human-defined heuristic rules, and hence, is more robust and fundamental than those described in existing work. We compare MEED with the generic seq2seq model and the hierarchical model of multi-turn dialogs, HRAN (Xing et al., 2018). Offline experiments show that our model outperforms both seq2seq and HRAN by a significant amount. Further experiments with human evaluation show our model produces emotionally more appropriate responses than both baselines, while also improving the language fluency. We also illustrate a human-evaluation procedure for judging machine produced emotional dialogs. We consider factors such as the balance of positive and negative emotions in test dialogs, a well-chosen range of topics, and dialogs that our human evaluators can relate. It is the first time such an approach is designed with consideration for human judges. Our main goal is to increase the objectivity of the results and reduce judges' mistakes due to out-of-context dialogs they have to evaluate.

- We present MEED2, the second generation of the MEED model, which is more controllable and interpretable, and is capable of generating responses that have finer-grained emotions and intents. We are the first to consider modeling a fine-grained set of empathetic response intents in an empathetic dialog model, which ensures a more precise learning of the emotional interactions revealed in the dialog data. To facilitate the training of our empathetic dialog model, we curated a large-scale dialog dataset from movie subtitles. To effectively evaluate our empathetic dialog model, we carefully designed a crowdsourcing experiment that enabled the workers to work on the tasks more easily. A total number of 6,000 dialogs were evaluated, which, to our knowledge, has never been attempted before for the evaluation of empathetic dialog systems.

- To further improve the accuracy of the emotion and intent labeling of the dialog dataset we obtained while developing MEED2, we adopted a semi-supervised learning technique to curate EDOS, a large-scale dialog dataset containing 1M emotional dialogs labeled with 32 fine-grained emotions, eight empathetic response intents (the plus categories), and *neutral*. We grew a seed dataset that was manually labeled by crowdsourcing workers, by assigning high-confidence labels predicted by an iteratively trained classifier to the semantically similar dialogs. Compared to existing dialog datasets tagged with emotions, EDOS is significantly larger ($\approx$ 40 times larger than EmpatheticDialogues),

and contains more fine-grained emotions and empathetic response strategies. We outline the complex pipeline used to derive this dataset. We analyze the quality of the dataset by comparing with a state-of-the-art gold standard dataset using visual validation methods.

- We present AFEC, a knowledge graph capturing social intelligence from people's daily conversations. We designed a completely automatic curation pipeline to create AFEC from casual conversations crawled from online forums, which can be used as an external resource to improve the performance of dialog systems. As a by-product of our knowledge graph, we obtained a large-scale casual conversation dataset that has a good trade-off between size and quality. We designed a simple retrieval-based chatbot using the knowledge graph, without the efforts of training a neural network, and the comparison with existing empathetic dialog models showed that our retrieval model is capable of generating much more diverse responses, yet still producing quality responses.

1.4 Thesis Structure

This thesis is organized as follows:

- We continue the thesis by discussing some background information in Chapter 2, including the concept of emotion and various emotion models, the concept of empathy, dialog systems in general, and empathetic dialog systems in particular.
- Chapter 3 presents a multi-turn emotionally engaging dialog model called MEED, by describing the model architecture, the training data, and the evaluation through automatic metrics and human judgement.
- Chapter 4 presents MEED2, the second generation of MEED, by describing its key components, the process of creating emotion labeled dialog data from movie subtitles, and how the model was evaluated through crowdsourcing platform.
- Chapter 5 describes a curation pipeline of the EDOS dataset, which adopts a semi-supervised learning framework to iteratively train an dialog emotion/intent classifier. It also presents some analysis against the state-of-the-art gold standard.
- Chapter 6 presents AFEC, a knowledge graph capturing social intelligence from casual conversations. It describes the detailed curation pipeline, the development of a retrieval-based chatbot, and its evaluation against several baselines.
- Finally, Chapter 7 concludes the thesis by discussing some of the findings and lessons learned while studying the aforementioned research problems. It also presents some future directions worth investigating.

2 Background

2.1 Emotion

Emotion is a complex concept, and the psychological literature does not have a consensus on its definition. The word "emotion" originates from the French word *émouvoir*, which means "to stir up." According to the Dictionary of Psychology[1] of the American Psychological Association (APA), emotion is "a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event." Scherer (1987, 2000, 2001) defined emotion to be an episode of coordinated changes in several components, including at least neurophysiological activation, motor expression, and subjective experience, but also maybe action tendencies, and cognitive processes, in reaction to significant internal or external events for the organism. Plutchik (2001) defined emotion to be a complex chain of loosely connected events that starts with a stimulus and involves feelings, psychological changes, impulses to action and specific, goal-oriented behaviors.

Though psychological researchers have different definitions for emotion, they generally agree upon two aspects: (1) Emotion is an reaction to a stimulus event of significant concern for the individual. For example, we feel *happy* when we receive a job offer, and we feel *angry* when someone threatens or attacks us. (2) Emotion involves a series of interrelated and synchronized changes in the individual's organismic subsystems. This includes physiological symptoms (for example the change of heart rate and blood pressure), motor expressions (both facial and vocal expressions), and subjective feelings or experience of emotional states.

How to separate or compare one emotion from another, is a contentious topic in affective science and emotion research. There are basically two ways that researchers have attempted classifying emotions: (1) viewing emotions as discrete categories; (2) defining emotions in multiple dimensions.

[1] https://dictionary.apa.org

2.1.1 Categorical Emotion Models

According to the discrete emotion theory, all people possess a universally recognizable collection of core emotions. No matter an individual's racial or cultural background, these particular core emotions are biologically based emotional responses whose expression and recognition are fundamentally the same for everyone. There is disagreement on the exact number of core emotions, despite the fact that many psychologists have embraced the principle of basic emotions. Plutchik (1984) suggested a psychoevolutionary technique to categorizing general emotional reactions, and proposed eight fundamental emotions: *anger*, *fear*, *sadness*, *disgust*, *surprise*, *anticipation*, *trust*, and *joy*. Ekman (1992) proposed six basic emotions: *fear*, *anger*, *joy*, *sadness*, *disgust*, and *surprise*. Izard et al. (1993) identified 12 distinct emotions and studied the role of these emotions in the development of personality.

When curating the EmpatheticDialogues dataset, Rashkin et al. (2019) considered a set of 32 emotion labels, which cover a wide range of positive and negative emotions, and were chosen by aggregating the labels from multiple other emotion prediction datasets. Welivita and Pu (2020) extended the 32 emotions with a set of eight empathetic response intents, plus the *neutral* category, which was derived from manually labeling a subset of the listener utterances. Compared with existing taxonomies of dialog acts/intents, which are either too general or too specific, these eight extra response intents are more associated with empathy; i.e., they help make the listener utterances more empathetic, thus promoting the emotional experience shared by the interlocutors. The whole EmpatheticDialogues dataset was then labeled with the new taxonomy using an automatic technique.

2.1.2 Dimensional Emotion Models

Another viewpoint of classifying emotions is to define them in continuous multi-dimensional space. This idea dates back to more than one century ago, when Wundt and Judd (1902) proposed that emotions can be described by three dimensions: *pleasurable* versus *unpleasurable*, *arousing* versus *subduing*, and *strain* versus *relaxation*. According to dimensional models of emotion, all emotional states are caused by a single, interrelated neurophysiological system, whereas the theory of basic emotions contends that many emotions derive from various brain systems (Posner et al., 2005). Russell (1980) proposed the circumplex model of emotion, where the emotions are distributed in a circular two-dimensional space, with the horizontal axis being *valence* (the pleasantness of a stimulus) and the vertical axis being *arousal* (the intensity of emotion). The circumplex model of emotion was then described to represent core affect, the most elementary consciously accessible affective feelings (Russell and Barrett, 1999). Osgood et al. (1957) added an extra dimension to the two-dimensional valence-arousal emotion model, namely *dominance*, to represent the degree of control exerted by a stimulus. Mehrabian (1980) proposed the PAD emotion model that uses three dimensions, *pleasure*, *arousal*, and *dominance*. The PAD model has also been used to analyze nonverbal communication like body language (Mehrabian, 2017).

2.2 Empathy

Empathy is believed to be an innate ability of human beings (Roth-Hanania et al., 2011) and plays an important role in people's social communication (Valente, 2016). The definition of empathy has been debatable since the introduction of the word into the English language (Lanzoni, 2018; Hall and Schwartz, 2019). It covers a wide range of behaviors, such as having compassion for others and wanting to be of assistance to them, feeling feelings similar to those of another person, and identifying another person's thoughts or feelings. Broadly speaking, empathy has been studied from two perspectives:

- **Cognitive Empathy.** Cognitive empathy is the ability to understand another's perspective (Davis, 1983). One of the key parts of cognitive empathy is perspective-taking, which places oneself in the position of someone else and get a better understanding of the experience. Cognitive empathy could also include the purposeful use of perspective-taking to accomplish specific goals.
- **Emotional Empathy.** Emotional empathy (Mehrabian and Epstein, 1972), or affective empathy, is the ability to respond with an appropriate emotion to another's mental states (Davis, 1983). This includes having sympathy and compassion for other people, and feeling the individual's own distress in response to other people's suffering. For example, we start to feel sad when we are sitting close to a friend as he/she begins to cry.

Compared with cognitive empathy, which is a more rational and logical process, emotional empathy is more spontaneous and based on emotional contagion (Shamay-Tsoory et al., 2009). In fact, cognitive empathy could be demonstrated without using any aspect of emotion.

2.3 Dialog Systems

A dialog system, or conversational agent, is a piece of software that can conduct conversations with humans. Depending on the style of the conversations, dialog systems could be command-based, menu-based, etc. However, in the context of current NLP research, dialog systems often refer to software that is capable of having natural language conversations with the users. Dialog systems typically fall into two categories: task-oriented dialog systems and open-domain dialog systems. Task-oriented dialog systems are used to assist users in completing certain tasks, such as making reservations at restaurants, booking flights, offering customer support, etc. These systems are usually domain-specific. Examples of task-oriented dialog systems are voice assistants on existing technological gadgets like Siri, Google Assistant, Alexa, and Cortana. Though users can also conduct chitchat with these voice assistants, their main job is to help users accomplish various tasks, related to work or life, on the devices. Open-domain dialog systems, on the other hand, are intended to have extended conversations with users, frequently imitating the open-domain or "chitchat" characteristics of human-human conversations. Examples are Microsoft's XiaoIce (Zhou et al., 2020) and Meta's BlenderBot (Roller et al., 2021).

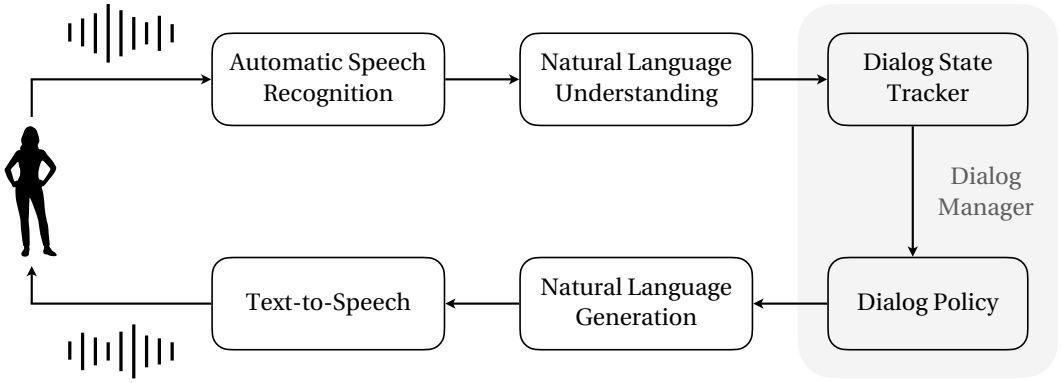


Figure 2.1: Architecture of a typical task-oriented dialog system, adapted from Figure 1 in the review by Williams et al. (2016). The automatic speech recognition module and the text-to-speech module are meant for spoken language processing and are not necessary for text-only task-oriented dialog systems.

Next we are going to give a brief introduction to these two classes of dialog systems, with more focus on open-domain dialog systems.

2.3.1 Task-Oriented Dialog Systems

Task-oriented dialog systems are mostly based on *frames*. A frame, which consists of a collection of *slots*, each of which can take a set of possible *values*, is a type of knowledge structure that represents the kinds of intentions the system can extract from user utterances. Earlier task-oriented dialog systems (Bobrow et al., 1977; Ward and Issar, 1994) adopted hand-designed rules to accomplish the slot-filling task. They will continuously ask questions until the frame is fully filled, and then query the relevant databases to return the answer. These systems also have other components like automatic speech recognition (ASR), to transcribe the audio input into a strings of words, and the natural language generation module, which is usually based on templates. Modern task-oriented dialog systems, whose architecture is shown in Figure 2.1, are also based on frames, but have more complicated modules. The *dialog state tracker* keeps track of the dialog's current state, including the user's most recent dialog act and all of the slot-filler constraints they have so far expressed. For the task of slot-filling, a common method is to train a sequence model based on BERT (Devlin et al., 2019) to map from the input representations to slot fillers, domain, and intent. The *dialog policy* makes the final call on what the system will say or do next. The dialog policy could be estimated by a neural classifier, or more complicatedly, through reinforcement learning (Fazel-Zarandi et al., 2017). Finally, the natural language generation module on modern task-oriented dialog systems is also more sophisticated, and can be neural-based and condition on specific context to generate more natural responses.

Since this thesis primarily focuses on open-domain dialog systems, we do not go into more details here. For more thorough reviews of task-oriented dialog systems, please refer to

Williams et al. (2016); Chen et al. (2017); Zhang et al. (2020c).

2.3.2 Open-Domain Dialog Systems

Compared with task-oriented dialog systems, open-domain dialog systems are more challenging to develop because they are supposed to conduct extended conversations with users, mimicking the "chitchat" characteristic of human-human conversations, and the goal to achieve in open dialogs is often open-ended and not clearly defined. Generally speaking, open-domain dialog systems can be classified into two broad categories: rule-based and corpus-based. Rule-based models use keyword matching and templates to generate responses. Corpus-based models make use of a corpus to generate responses, usually by information retrieval methods or neural generative approaches.

Rule-Based Models

Rule-based dialog models take advantage of a set of manually defined rules that map out conversations like a flowchart. The advantage of rule-based dialog models is that they are less expensive to train (though experts might be needed to define the rules), but these models could not deal with scenarios that fall outside what their rules have defined. Earlier attempts on chatbots often adopt this type of response generation scheme. For example, ELIZA (Weizenbaum, 1966) is a chatbot originally designed to simulate the role of a psychologist. It relies on a set of patterns that work like regular expressions to recognize certain parts of the user's input and then transform it into a response. These rules are ranked based on how common their associated keywords are, with more specific keywords ranking higher. A later chatbot also focusing on psychology, PARRY (Colby et al., 1971), included certain affect variables to model its own levels of anger and fear, in addition to the regular expressions similar to those of ELIZA. A.L.I.C.E. (Wallace, 2009), a more recent rule-based chatbot inspired by and adapted from ELIZA, uses AIML (Artificial Intelligence Markup Language) files to define its heuristic conversation rules. It won the bronze Loebner Prize (a competition that honors computer programs that deemed to be the most human-like by the judges) three times (in 2000, 2001, and 2004), but failed to pass the Turing test.

Retrieval-Based Models

One type of corpus-based dialog models is retrieval-based. These dialog models usually rely on a corpus of human-human conversations and make use of information retrieval techniques to select out the most relevant candidate as the response. In its simplest form, for an input context x and a corpus $\mathscr{D}$, we compute the tf-idf vectors for both x and $\mathscr{D}$ using conventional information retrieval approaches, and then select the turn $\hat{y} \in \mathscr{D}$ that has the highest cosine similarity value with x:

$$\hat{y} = \arg\max_{y \in \mathscr{D}} \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}. \tag{2.1}$$

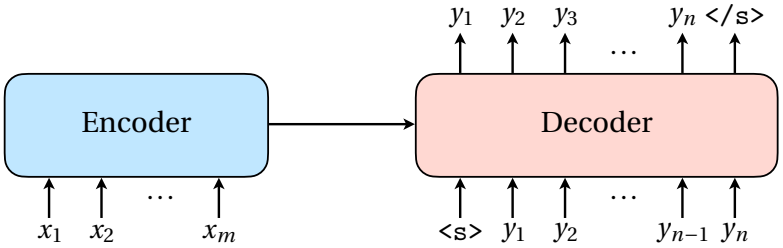


Figure 2.2: The encoder-decoder architecture for generative dialog models. The encoder encodes the dialog context x into vector representations, and the decoder conditions on these encoded representations to generate the response y, often in an autoregressive way.

We then return $\hat{y}$ as the response. Alternatively, we can also select the turn following $\hat{y}$ as the response. More recently, researchers started to use neural networks to learn a matching score instead of using the cosine similarity, and also represent the context and the candidate response using more sophisticated neural models (Hu et al., 2014; Lowe et al., 2015; Humeau et al., 2020).

The approaches described above are often referred to as representation-based methods, where the interaction between the input context x and the candidate response y happens in a later stage, after they have been encoded into, for example, vector representations. Another group of models is based on the deep interaction between x and y, where the interaction happens in an earlier stage, and it allows the model to learn a fused representation at the end for the matching task (Wu et al., 2017; Zhou et al., 2018c; Tao et al., 2019; Yuan et al., 2019). In this case, the context x usually consists of multiple utterances. The model first fuses the representations of y and each utterance of x, respectively, and then aggregates these fused representations to obtain an aggregated feature, based on which the matching score can be calculated.

Finally, with pre-trained language models being popular recently, researchers started to apply them for the task of response selection (Henderson et al., 2019; Gu et al., 2020; Xu et al., 2021). This is usually done by concatenating the context x and the candidate response y into one sequence and feed it into a pre-trained language model such as BERT, so that the representation, interaction, and aggregation actions can be carried out at the same time by a unified model.

Generative Models

Another way to utilize a corpus to generate responses is to adopt a machine learning approach and take advantage of neural networks. Ritter et al. (2011) first considered using statistical machine translation approach to generate responses to Twitter posts. This idea was then generalized to an encoder-decoder architecture, as shown in Figure 2.2, where the encoder encodes the input context x into vector representations, and the decoder conditions on these representations to generate the response y, often in an autoregressive way. In particular, Vinyals and Le (2015) took inspiration from Sutskever et al. (2014) and made use of the

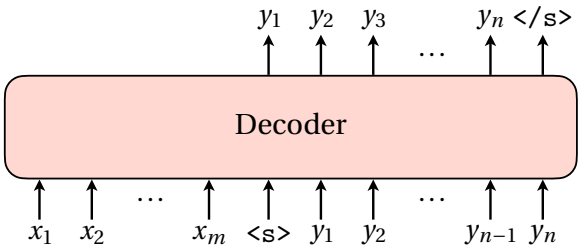


Figure 2.3: The decoder-only architecture for generative dialog models. The dialog context x is directly fed into the decoder as a sequence, and the decoder continues the sequence to generate the response y.

sequence-to-sequence (seq2seq) network to train a neural conversation model on a closed-domain IT helpdesk troubleshooting dataset and an open-domain movie transcript dataset. Shang et al. (2015) applied attention mechanism (Bahdanau et al., 2015) to the same structure and trained the model on Weibo data, a popular Twitter-like microblogging service in China. To model the hierarchical utterance-word structure of the context, Serban et al. (2016) adopted the hierarchical recurrent encoder-decoder (HRED) model (Sordoni et al., 2015a) to build an end-to-end dialog system. To give attention to different parts of the context while generating responses, Xing et al. (2018) proposed the hierarchical recurrent attention network (HRAN), using a hierarchical attention mechanism.

To take advantage of the large-scale pre-trained language models, there is also work on building open-domain dialog systems using a decoder-only architecture resembling a language model, as shown in Figure 2.3. This is usually done by concatenating all the utterances of the context x into one sequence and letting the decoder continue to generate the response y. Wolf et al. (2019) proposed the TransferTransfo model based on the pre-trained GPT model (Radford et al., 2018), which was fine-tuned with a multi-task objective. The empathetic chatbot CAiRE (Lin et al., 2020), also adapted from GPT, is an end-to-end dialog model capable of recognizing user emotions and responding in an empathetic manner. Zhang et al. (2020b) presented DIALOGPT on the basis of GPT-2 (Radford et al., 2019) and trained it on 147M conversation-like exchanges extracted from Reddit comments. Recently, Thoppilan et al. (2022) presented LaMDA (Language Models for Dialog Applications), a Transformer (Vaswani et al., 2017) decoder-only language model dedicated to dialog application, which have up to 137B parameters and was pre-trained on 1.56T words of public web text and dialog data.

Though the advent of neural networks brought huge advancements to open-domain dialog generation, we still face several key challenges. For example, generative neural-based dialog models are known to suffer from the diversity issue, where the generated responses are often generic and uninformative like "I don't know." There already exists some work that focuses on addressing this diversity issue (Li et al., 2016a,c; Zhao et al., 2017b; Du et al., 2018; Vijayakumar et al., 2018). Some other work attempted at keeping the persona of the chatbot consistent (Li et al., 2016b; Zhang et al., 2018; Qian et al., 2018; Zhang et al., 2019a,b), and others tried to ground the response generation process with external knowledge sources (Ghazvininejad

et al., 2018; Zhou et al., 2018b; Dinan et al., 2019; Zhao et al., 2020; Li et al., 2022b). For a more thorough review of these challenges, please refer to Huang et al. (2020).

2.4 Emotional and Empathetic Dialog Systems

One of the trends in the research community of dialog generation is to incorporate emotion information into the design of dialog systems. The goal is to generate emotionally richer responses, and/or make the chatbot more sound more empathetic, i.e., recognizing and understanding the user's emotions and then responding in a way that is emotionally appropriate. In Table 2.1, we summarize some of the existing work on emotional and empathetic dialog models, with respect to their emotion models, neural architectures, and evaluation methods.

One line of research focuses on making the generated response emotionally richer. Asghar et al. (2018) appended the original word embeddings in the seq2seq model with a VAD affect vector (Warriner et al., 2013). VAD is a vector model, as opposed to a categorical model such as LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001), that represents a given emotion in each of the valence, arousal, and dominance dimensions. The proposed neural affect dialog model aims at generating explicit responses given a particular utterance. To do so, they designed three affect-related loss functions, namely minimizing affective dissonance, maximizing affective dissonance, and maximizing affective content. The paper also proposed the affectively diverse beam search during decoding, so that the generated candidate responses are as affectively diverse as possible. Later on, Zhong et al. (2019) proposed an affect-rich dialog model using biased attention mechanism on emotional words in the input message, by taking advantage of the VAD embeddings. The model was trained with a weighted cross-entropy loss function, which encourages the generation of emotional words.

Some other work adopts the setting in which a pre-defined emotion label is given to the model so that it generates responses accordingly. The Emotional Chatting Machine (ECM) (Zhou et al., 2018a) takes a post and generates a response in a predefined emotion category. The main idea is to use an internal memory module to capture the emotion dynamics during decoding, and an external memory module to model emotional expressions explicitly by assigning different probability values to emotional words as opposed to regular words. Zhou and Wang (2018) extended the standard seq2seq model to a conditional variational autoencoder combined with policy gradient techniques. The model takes a post and an emoji as input, and generates the response with target emotion specified by the emoji. Hu et al. (2018) built a tone-aware chatbot for customer care on social media, by deploying extra meta information of the conversations in the seq2seq model. Specifically, a tone indicator is added to each step of the decoder during the training phase. Colombo et al. (2019) proposed EMOTICONS, an affect-driven dialog system that generates emotional responses in a controlled way using a continuous representation of emotions. Song et al. (2019) proposed EmoDS (Emotional Dialogue System), which is able to generate responses with coherent structures according to a desired emotion, expressed either explicitly or implicitly.

Table 2.1: A summary of some existing work on emotional and empathetic dialog models.

Model	**Emotion Model**	**Emotion Interaction**	**Base**	**Human Evaluation**
Affective Seq2Seq (Asghar et al., 2018)	*Continuous* Valence, arousal, dominance	Heuristic rules	RNN (LSTM)	5 raters on 100 dialogs
Emotional Chatting Machine (Zhou et al., 2018a)	*Categorical* 6 emotions	Emotion label as input	RNN (GRU)	3 raters on 200 dialogs
MojiTalk (Zhou and Wang, 2018)	*Categorical* 64 emojis	Emotion label as input	RNN (GRU)	100 dialogs
Affect-Rich Seq2Seq (Zhong et al., 2019)	*Continuous* Valence, arousal, dominance	Encourage affect-rich words	RNN (LSTM)	5 raters on 100 dialogs
EMOTICONS (Colombo et al., 2019)	*Categorical* 6 emotions	Emotion label as input	RNN (GRU)	3/18/22 raters on 40/45/120 dialogs
Emotion-Aware Chat Machine (Wei et al., 2019)	*Categorical* 6 emotions	Emotion selection by model	RNN (GRU)	3 raters on 200 dialogs
EmoDS (Song et al., 2019)	*Categorical* 6 emotions	Emotion label as input	RNN (LSTM)	3 raters on 200 dialogs
MoEL (Lin et al., 2019)	*Categorical* 32 emotions	Multiple emotion listeners	Transformer	100 dialogs
EmpDG (Li et al., 2020)	*Categorical* 7 emotions	Emotional discriminator	RNN (LSTM)	100 dialogs
MIME (Majumder et al., 2020)	*Categorical* 32 emotions	Mimicking the user emotion	Transformer	3 raters on 128 dialogs
CoMAE (Zheng et al., 2021)	*Categorical* 10 emotions; 9 dialog acts	Communication mechanism predicted by model	Transformer	3 raters on 200 dialogs
CEM (Sabour et al., 2022)	*Categorical* 32 emotions	Commonsense inference	Transformer	3 raters on 100 dialogs
KEMP (Li et al., 2022a)	*Categorical* 32 emotions	Emotional context graph	Transformer	3 raters on 100 dialogs

In addition to the aforementioned work, some also considered modeling more sophisticated emotion interactions. Wei et al. (2019) proposed the EACM (Emotion-Aware Chat Machine), which is based on the seq2seq network and consists of an emotion selector that selects the desired emotion out of six basic labels, and a response generator that generates the corresponding response. MoEL (Lin et al., 2019) is an end-to-end empathetic dialog model that uses multiple decoders (listeners) to react to each context emotion accordingly. According to the emotion classification distribution, a meta-listener then combines the output states of each listener, to generate the final empathetic response. Majumder et al. (2020) proposed the MIME model, which generates empathetic responses by exploiting the assumption that an empathetic conversational agent would often mimic the user's emotion to a certain degree, depending on whether it is positive or negative. The model also introduces some stochasticity into the emotion mixture to generate more varied responses. Sabour et al. (2022) introduced the CEM model by taking commonsense knowledge into consideration. When generating the empathetic response, CEM first queries COMET (Bosselut et al., 2019), a GPT-2 based model fine-tuned on the ATOMIC knowledge graph (Sap et al., 2019), to get the commonsense inferences of the input context, and then uses a knowledge selector to fuse the obtained information.

3 Multi-Turn Emotionally Engaging Dialog Generation

This chapter is based on the work of Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu (Xie et al., 2020). The author of this thesis (Yubo Xie) was mainly responsible for designing, implementing and training the dialog model, as well as its automatic evaluation.

3.1 Introduction

Many application areas show significant benefits of integrating affect information in natural language dialogs. In earlier work on human-computer interaction, Klein et al. (2001) found user's frustration caused by a computer system can be alleviated by computer-initiated emotional support, by providing feedback on emotional content along with sympathy and empathy. More recently, Hu et al. (2018) developed a customer support neural chatbot, capable of generating dialogs similar to the humans in terms of empathic and passionate tones, potentially serving as proxy customer support agents on social media platforms. In a qualitative study (Zamora, 2017), participants expressed an interest in chatbots capable of serving as an attentive listener and providing motivational support, thus fulfilling users' emotional needs. Several participants even noted a chatbot is ideal for sensitive content that is too embarrassing to ask another human. Finally, Bickmore and Picard (2005) showed a relational agent with deliberate social-emotional skills was respected more, liked more, and trusted more, even after four weeks of interaction, compared to an equivalent task-oriented agent.

Recent development in neural language modeling has generated significant excitement in the open-domain dialog generation community. The success of sequence-to-sequence (seq2seq) learning (Sutskever et al., 2014; Cho et al., 2014) in the field of neural machine translation has inspired researchers to apply the recurrent neural network (RNN) encoder-decoder structure to response generation (Vinyals and Le, 2015). Following the standard seq2seq structure, various improvements have been made on the neural conversation model. For example, Shang et al. (2015) applied attention mechanism (Bahdanau et al., 2015) to the same structure on Twitter-style microblogging data. Li et al. (2016a) found the original version tend to favor short and dull responses. They fixed this problem by increasing the diversity of the response.

Li et al. (2016b) modeled the personalities of the speakers, and Xing et al. (2017) developed a topic aware dialog system. We call work in this area globally neural dialog generation.

More recently, researchers started incorporating affect information into neural dialog models. While a central theme seems to be making the responses emotionally richer, existing approaches mainly follow two directions. In one, an emotion label is explicitly required as input so that the machine can generate sentences of that particular emotion label or type (Zhou et al., 2018a; Huang et al., 2018; Zhou and Wang, 2018; Colombo et al., 2019; Song et al., 2019; Shen and Feng, 2020). In another group of work, the main idea is to develop handcrafted rules to direct the machines to generated responses of the desired emotions (Asghar et al., 2018; Zhong et al., 2019). Both approaches require an emotion label as input (either given or handcrafted), which might be unpractical in real dialog scenarios.

Furthermore, to the best of our knowledge, the psychology and social science literature does not provide clear rules for emotional interaction. It seems such social and emotional intelligence is captured in our conversations. This is why we decided to take the automatic and data-driven approach. In this chapter, we describe an end-to-end Multi-turn Emotionally Engaging Dialog model (MEED), capable of recognizing emotions and generating emotionally appropriate and human-like responses with the ultimate goal of reproducing social behaviors that are habitual in human-human conversations. We chose the multi-turn setting because a model suitable for single-turn dialogs cannot effectively track earlier context in multi-turn dialogs, both semantically and emotionally. Since being able to track several turns is really important, we made this design decision from the beginning, in contrast to most related work where models are only trained and tested on single-turn dialogs. While using a hierarchical mechanism to track the conversation history in multi-turn dialogs is not new (e.g., HRAN by Xing et al. (2018)), to combine it with an additional emotion RNN to process the emotional information in each history utterance has never been attempted before.

Our contributions are threefold. (1) We describe in detail a novel emotion-tracking dialog generation model that learns the emotional interactions directly from the data. This approach is free of human-defined heuristic rules, and hence, is more robust and fundamental than those described in existing work. (2) We compare our model, MEED, with the generic seq2seq model and the hierarchical model of multi-turn dialogs (HRAN). Offline experiments show that our model outperforms both seq2seq and HRAN by a significant amount. Further experiments with human evaluation show our model produces emotionally more appropriate responses than both baselines, while also improving the language fluency. (3) We illustrate a human-evaluation procedure for judging machine produced emotional dialogs. We consider factors such as the balance of positive and negative emotions in test dialogs, a well-chosen range of topics, and dialogs that our human evaluators can relate. It is the first time such an approach is designed with consideration for human judges. Our main goal is to increase the objectivity of the results and reduce judges' mistakes due to out-of-context dialogs they have to evaluate.

3.2 Related Work

3.2.1 Neural Dialog Generation

Vinyals and Le (2015) were one of the first to model dialog generation using neural networks. Their seq2seq framework was trained on an IT Helpdesk Troubleshooting dataset and the OpenSubtitles dataset (Lison and Tiedemann, 2016). Shang et al. (2015) further trained the seq2seq model with attention mechanism on a self-crawled Weibo (a popular Twitter-like social media website in China) dataset. Meanwhile, Xu et al. (2017) built a customer service chatbot by training the seq2seq model on a dataset collected with conversations between customers and customer service accounts from 62 brands on Twitter.

The standard seq2seq framework is applied to single-turn response generation. In multi-turn settings, where a context with multiple history utterances is given, the same structure often ignores the hierarchical characteristic of the context. Some recent work addresses this problem by adopting a hierarchical recurrent encoder-decoder (HRED) structure (Sordoni et al., 2015a; Serban et al., 2016, 2017). To give attention to different parts of the context while generating responses, Xing et al. (2018) proposed the hierarchical recurrent attention network (HRAN), using a hierarchical attention mechanism. However, these multi-turn dialog models do not take into account the turn-taking emotional changes of the dialog.

3.2.2 Neural Dialog Models with Affect Information

Recent work on incorporating affect information into natural language processing tasks has inspired our current work. They can be mainly described as affect language models and emotional dialog systems.

Ghosh et al. (2017) made the first attempt to augment the original LSTM language model with affect treatment in what they called Affect-LM. At training time, Affect-LM can be considered as an energy based model where the added energy term captures the degree of correlation between the next word and the affect information of the preceeding text. At text generation time, affect information is also used to increase the appropriate selection of the next word. A key component in Affect-LM is the use of a well established text analysis program, LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001). For every sentence (e.g., "I unfortunately did not pass my exam"), the model generates five emotion features denoting (*sad*: 1, *angry*: 1, *anxiety*: 1, *negative emotion*: 1, *positive emotion*: 0). This makes Affect-LM both capable of distinguishing affect information conveyed by each word in the language modeling part and aware of the preceeding text's emotion in each generation step. In a similar vein, Asghar et al. (2018) appended the original word embeddings with a VAD affect model (Warriner et al., 2013). VAD is a vector model, as opposed to a categorical model (LIWC), representing a given emotion in each of the valence, arousal, and dominance axes. In contrast to Affect-LM, Asghar's neural affect dialog model aims at generating explicit responses given a particular utterance. To do so, the authors designed three affect-related loss functions, namely

minimizing affective dissonance, maximizing affective dissonance, and maximizing affective content. The paper also proposed the affectively diverse beam search during decoding, so that the generated candidate responses are as affectively diverse as possible. However, literature in affective science does not necessarily validate such rules. In fact, the best strategy to speak to an angry customer is the de-escalation strategy (using neutral words to validate anger) rather than employing equally emotional words (minimizing affect dissonance) or words that convey happiness (maximizing affect dissonance).

The Emotional Chatting Machine (ECM) (Zhou et al., 2018a) takes a post and generates a response in a predefined emotion category. The main idea is to use an internal memory module to capture the emotion dynamics during decoding, and an external memory module to model emotional expressions explicitly by assigning different probability values to emotional words as opposed to regular words. Zhou and Wang (2018) extended the standard seq2seq model to a conditional variational autoencoder combined with policy gradient techniques. The model takes a post and an emoji as input, and generates the response with target emotion specified by the emoji. Hu et al. (2018) built a tone-aware chatbot for customer care on social media, by deploying extra meta information of the conversations in the seq2seq model. Specifically, a tone indicator is added to each step of the decoder during the training phase.

In parallel to these developments, Zhong et al. (2019) proposed an affect-rich dialog model using biased attention mechanism on emotional words in the input message, by taking advantage of the VAD embeddings. The model was trained with a weighted cross-entropy loss function, which encourages the generation of emotional words.

3.2.3 Summary

As much as these work in the above section inspired our work, our approach in generating affect dialogs is significantly different. Most of related work focused on integrating affect information into the transduction vector space using either VAD or LIWC, we aim at modeling and generating the affect exchanges in human dialogs using a dedicated embedding layer. The approach is also completely data-driven, thus absent of hand-crafted rules. To avoid learning obscene and callous exchanges often found in social media data like tweets and Reddit threads (Rashkin et al., 2019), we opted to train our model on movie subtitles, whose dialogs were carefully created by professional writers. We believe the quality of this dataset can be better than those curated by crowdsource platforms. For modeling the affect information, we chose to use LIWC because it is a well-established emotion lexical resource, covering the whole English dictionary whereas VAD only contains 13K lemmatized terms.

3.3 Model

In order to track previous turns in the dialog history (both semantically and emotionally), we adopt a multi-turn setting. Some previous dialog models use a vanilla seq2seq structure as

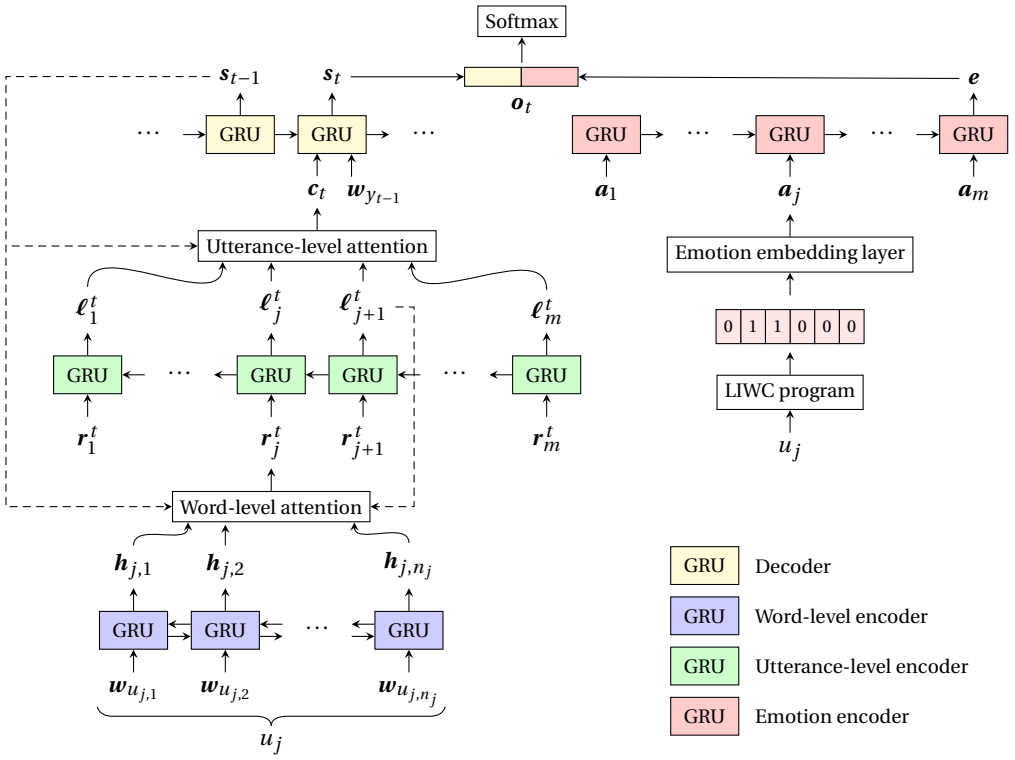


Figure 3.1: The overall architecture of the MEED model.

the backbone, as they often assume a single-turn setting, and it suffices to just use a normal RNN to encode the input. Although we could also use a vanilla seq2seq structure to encode the input dialog history that contains multiple utterances (by concatenating them into one single sequence), the hierarchical structure of the dialog context is ignored and thus the model will not be able to differentiate between specific utterances in the context. This leads to degradation of the quality of the generated responses. Therefore, we adopt a hierarchical architecture to encode the input, which better fits our multi-turn setting. In this section, we describe our model one element at a time, from the basic structure, to the hierarchical component, and finally the emotion embedding layer.

We first consider the problem of generating response y given a context x consisting of multiple previous utterances by estimating the probability distribution $p(y|x)$ from a data set $\mathscr{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ containing N context-response pairs. Here

$$x^{(i)} = \left(u_1^{(i)}, u_2^{(i)}, \ldots, u_{m_i}^{(i)}\right) \tag{3.1}$$

is a sequence of m_i utterances, and

$$u_j^{(i)} = \left(u_{j,1}^{(i)}, u_{j,2}^{(i)}, \ldots, u_{j,n_{ij}}^{(i)}\right) \tag{3.2}$$

is a sequence of n_{ij} words. Similarly,

$$y^{(i)} = \left(y_1^{(i)}, y_2^{(i)}, \ldots, y_{T_i}^{(i)}\right) \tag{3.3}$$

is the response with T_i words.

Usually the probability distribution $p(y|x)$ can be modeled by an RNN language model conditioned on X. When generating the word y_t at time step t, the context x is encoded into a fixed-sized dialog context vector $\boldsymbol{c}_t$ by following the hierarchical attention structure in HRAN (Xing et al., 2018). Additionally, we extract the emotion information from the utterances in x by leveraging an external text analysis program, and use an RNN to encode it into an emotion context vector $\boldsymbol{e}$, which is combined with $\boldsymbol{c}_t$ to produce the distribution. The overall architecture of the model is depicted in Figure 3.1. We are going to elaborate on how to obtain $\boldsymbol{c}_t$ and $\boldsymbol{e}$, and how they are combined in the decoding part.

3.3.1 Hierarchical Attention

The hierarchical attention structure involves two encoders to produce the dialog context vector $\boldsymbol{c}_t$, namely the word-level encoder and the utterance-level encoder. The word-level encoder is essentially a bidirectional RNN with gated recurrent units (GRU) (Cho et al., 2014). For utterance u_j in x ($j = 1, 2, \ldots, m$), the bidirectional encoder produces two hidden states at each word position k, the forward hidden state $\boldsymbol{h}_{jk}^{\mathrm{f}}$ and the backward hidden state $\boldsymbol{h}_{jk}^{\mathrm{b}}$. The final hidden state $\boldsymbol{h}_{jk}$ is then obtained by concatenating the two,

$$\boldsymbol{h}_{jk} = \mathrm{concat}\left(\boldsymbol{h}_{jk}^{\mathrm{f}}, \boldsymbol{h}_{jk}^{\mathrm{b}}\right). \tag{3.4}$$

The utterance-level encoder is a unidirectional RNN with GRU that goes from the last utterance in the context to the first, with its input at each step as the summary of the corresponding utterance, which is obtained by applying a Bahdanau-style attention mechanism (Bahdanau et al., 2015) on the word-level encoder output. Using this backward RNN, we make sure that the information contained in the last utterance populates at each time step of the RNN, and with the following attention mechanism, the model would focus more on the last utterance. More specifically, at decoding step t, the summary of utterance u_j is a linear combination of $\boldsymbol{h}_{jk}$, for $k = 1, 2, \ldots, n_j$,

$$\boldsymbol{r}_j^t = \sum_{k=1}^{n_j} \alpha_{jk}^t \boldsymbol{h}_{jk}. \tag{3.5}$$

Here α_{jk}^t is the word-level attention score placed on $\boldsymbol{h}_{jk}$, and can be calculated as

$$a_{jk}^t = \boldsymbol{v}_a^T \tanh\left(\boldsymbol{U}_a \boldsymbol{s}_{t-1} + \boldsymbol{V}_a \boldsymbol{\ell}_{j+1}^t + \boldsymbol{W}_a \boldsymbol{h}_{jk}\right), \tag{3.6}$$

$$\alpha_{jk}^t = \frac{\exp\left(a_{jk}^t\right)}{\sum_{k'=1}^{n_j} \exp\left(a_{jk'}^t\right)}, \tag{3.7}$$

where $\boldsymbol{s}_{t-1}$ is the previous hidden state of the decoder, $\boldsymbol{\ell}^t_{j+1}$ is the previous hidden state of the utterance-level encoder, and $\boldsymbol{v}_a$, $\boldsymbol{U}_a$, $\boldsymbol{V}_a$ and $\boldsymbol{W}_a$ are word-level attention parameters. The final dialog context vector $\boldsymbol{c}_t$ is then obtained as another linear combination of the outputs of the utterance-level encoder $\boldsymbol{\ell}^t_j$, for $j = 1, 2, \ldots, m$,

$$\boldsymbol{c}_t = \sum_{j=1}^{m} \beta^t_j \boldsymbol{\ell}^t_j. \tag{3.8}$$

Here β^t_j is the utterance-level attention score placed on $\boldsymbol{\ell}^t_j$, and can be calculated as

$$b^t_j = \boldsymbol{v}_b^T \tanh\left(\boldsymbol{U}_b \boldsymbol{s}_{t-1} + \boldsymbol{W}_b \boldsymbol{\ell}^t_j\right), \tag{3.9}$$

$$\beta^t_j = \frac{\exp\left(b^t_j\right)}{\sum_{j'=1}^{m} \exp\left(b^t_{j'}\right)}, \tag{3.10}$$

where $\boldsymbol{s}_{t-1}$ is the previous hidden state of the decoder, and $\boldsymbol{v}_b$, $\boldsymbol{U}_b$ and $\boldsymbol{W}_b$ are utterance-level attention parameters.

3.3.2 Emotion Encoder

The main objective of the emotion embedding layer is to recognize the affect information in the given utterances so that the model can respond with emotionally appropriate replies. To achieve this, we need an encoder to distinguish the affect information in the context, in addition to its semantic meaning. Equally we need a decoder capable of selecting the best and most human-like answers.

We are able to achieve this goal, i.e., capturing the emotion information carried in the context x, in the encoder, thanks to LIWC. We make use of the five emotion-related categories, namely *positive emotion*, *negative emotion*, *anxious*, *angry*, and *sad*. This set can be expanded to include more categories if we desire a richer distinction. See the discussion section for more details on how to do this. Using the newest version of the program LIWC2015,[1] we are able to map each utterance u_j in the context to a six-dimensional indicator vector $\mathbf{1}(u_j)$, with the first five entries corresponding to the five emotion categories, and the last one corresponding to *neutral*. If any word in u_j belongs to one of the five categories, then the corresponding entry in $\mathbf{1}(u_j)$ is set to 1; otherwise, u_j is treated as neutral, with the last entry of $\mathbf{1}(u_j)$ set to 1. For example, assuming u_j = "he is worried about me", then

$$\mathbf{1}(u_j) = [0, 1, 1, 0, 0, 0], \tag{3.11}$$

since the word "worried" is assigned to both *negative emotion* and *anxious*. We apply a dense layer with sigmoid activation function on top of $\mathbf{1}(u_j)$ to embed the emotion indicator vector

[1] https://liwc.wpengine.com/

into a continuous space,

$$\boldsymbol{a}_j = \sigma(\boldsymbol{W}_e \mathbf{1}(u_j) + \boldsymbol{b}_e), \tag{3.12}$$

where $\boldsymbol{W}_e$ and $\boldsymbol{b}_e$ are trainable parameters. The emotion flow of the context x is then modeled by an unidirectional RNN with GRU going from the first utterance in the context to the last, with its input being $\boldsymbol{a}_j$ at each step. The final emotion context vector $\boldsymbol{e}$ is obtained as the last hidden state of this emotion encoding RNN. Here, unlike the utterance-level and word-level RNNs, we use a forward RNN that goes from the first utterance to the last one in the dialog context. This is because at each time step, the input is just a six-dimensional zero-one vector, and such amount of information can be sufficiently processed and encoded by a forward RNN without attention mechanism.

3.3.3 Decoding

The probability distribution $p(y \mid x)$ can be written as

$$\begin{aligned} p(y \mid x) &= p(y_1, y_2, \ldots, y_T \mid x) \\ &= p(y_1 \mid \boldsymbol{c}_1, \boldsymbol{e}) \prod_{t=2}^{T} p(y_t \mid y_1, \ldots, y_{t-1}, \boldsymbol{c}_t, \boldsymbol{e}). \end{aligned} \tag{3.13}$$

We model the probability distribution using an RNN language model along with the emotion context vector $\boldsymbol{e}$. Specifically, at time step t, the hidden state of the decoder $\boldsymbol{s}_t$ is obtained by applying the GRU function,

$$\boldsymbol{s}_t = \mathrm{GRU}(\boldsymbol{s}_{t-1}, \mathrm{concat}(\boldsymbol{c}_t, \boldsymbol{w}_{y_{t-1}})), \tag{3.14}$$

where $\boldsymbol{w}_{y_{t-1}}$ is the word embedding of y_{t-1}. Similar to Affect-LM Ghosh et al. (2017), we then define a new feature vector $\boldsymbol{o}_t$ by concatenating $\boldsymbol{s}_t$ (which we refer to as the language context vector) with the emotion context vector $\boldsymbol{e}$,

$$\boldsymbol{o}_t = \mathrm{concat}(\boldsymbol{s}_t, \boldsymbol{e}), \tag{3.15}$$

on which we apply a softmax layer to obtain a probability distribution over the vocabulary,

$$\boldsymbol{p}_t = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{o}_t + \boldsymbol{b}), \tag{3.16}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters. Each term in Equation (3.13) is then given by

$$p(y_t \mid y_1, \ldots, y_{t-1}, \boldsymbol{c}_t, \boldsymbol{e}) = \boldsymbol{p}_{t, y_t}. \tag{3.17}$$

We use the cross-entropy loss as our objective function

$$\mathscr{L} = -\frac{1}{\sum_{i=1}^{N} T_i} \sum_{i=1}^{N} \log p\big(y^{(i)} \,\big|\, x^{(i)}\big). \tag{3.18}$$

Table 3.1: Statistics of the Cornell Movie-Dialogs Corpus and the DailyDialog dataset.

	Cornell	**DailyDialog**
Number of dialogs	83,097	13,118
Number of utterances	304,713	102,977
Average number of turns	3.7	7.9
Average number of words per utterance	12.5	14.6
Training set size	142,450	46,797
Validation set size	10,240	10,240

3.4 Evaluation

We trained our model using two different datasets and compared its performance with HRAN as well as the basic seq2seq model by performing both offline and online testings.

3.4.1 Datasets

We used two different dialog corpora to train our model—the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011) and the DailyDialog dataset (Li et al., 2017).

- **Cornell Movie-Dialogs Corpus**. The dataset contains 83,097 dialogs (220,579 conversational exchanges) extracted from raw movie scripts. In total there are 304,713 utterances.
- **DailyDialog**. The dataset is developed by crawling raw data from websites used for language learners to learn English dialogs in daily life. It contains 13,118 dialogs in total.

We summarize some of the basic information regarding the two datasets in Table 3.1.

In our experiments, the models were first trained on the Cornell Movie-Dialogs Corpus, and then fine-tuned on the DailyDialog dataset. We adopted this training pattern because the Cornell dataset is bigger but noisier, while DailyDialog is smaller but more daily-based. To create a training set and a validation set for each of the two datasets, we took segments of each dialog with number of turns no more than six,[2] to serve as the training/validation examples. Specifically, for each dialog $D = (u_1, u_2, \ldots, u_M)$, we created $M-1$ context-response pairs, namely $U_i = (u_{s_i}, \ldots, u_i)$ and $y_i = u_{i+1}$, for $i = 1, 2, \ldots, M-1$, where $s_i = \max(1, i-4)$. We filtered out those pairs that have at least one utterance with length greater than 30. We also reduced the frequency of those pairs whose responses appear too many times (the threshold is set to 10 for Cornell, and 5 for DailyDialog), to prevent them from dominating the learning procedure. See Table 3.1 for the sizes of the training and validation sets. The test set consists of 100 dialogs with four turns. We give more detailed description of how we created the test set in the section of human evaluation.

[2]We chose the maximum number of turns to be six because we would like to have a longer context for each dialog while at the same time keeping the training procedure computationally efficient.

3.4.2 Baselines and Implementation

Our comparision is based on three multi-turn dialog generation models: the standard seq2seq model (denoted as S2S), HRAN, and our proposed model, MEED. Our choice of including S2S is rather obvious. Including HRAN instead of other neural dialog models with affect information was not an easy decision. As mentioned in the related work, Asghar's affective dialog model, the affect-rich conversation model, and the Emotional Chatting Machine do not learn the emotional exchanges in the dialogs. This leaves us wondering whether using a multi-turn neural model can be as effective in learning emotional exchanges as MEED. In addition, comparing S2S and HRAN also gives us an idea of how much the hierarchical mechansim is improving upon the basic model. In order to adapt S2S to the multi-turn setting, we concatenate all the history utterances in the context into one.

For all the models, the vocabulary consists of 20,000 most frequent words in the Cornell and DailyDialog datasets, plus three extra tokens: `<unk>` for words that do not exist in the vocabulary, `<go>` indicating the begin of an utterance, and `<eos>` indicating the end of an utterance. Here we summarize the configurations and parameters of our experiments:

- We set the word embedding size to 256. We initialized the word embeddings in the models with word2vec (Mikolov et al., 2013) vectors first trained on Cornell and then fine-tuned on DailyDialog, consistent with the training procedure of the models.
- We set the number of hidden units of each RNN to 256, the word-level attention depth to 256, and utterance-level 128. The output size of the emotion embedding layer is 256.
- We optimized the objective function using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.001.
- For prediction, we used beam search Tillmann and Ney (2003) with a beam width of 256.

We have made the source code publicly available.[3]

3.4.3 Evaluation Metrics

The evaluation of chatbots remains an open problem in the field. Recent work (Liu et al., 2016) has shown that the automatic evaluation metrics borrowed from machine translation such as BLEU score (Papineni et al., 2002) tend to align poorly with human judgement. Therefore, in this chapter, we mainly adopt human evaluation, along with perplexity and BLEU score, following the existing work.

[3] https://github.com/yuboxie/meed

Automatic Evaluation

Perplexity is a measurement of how a probability model predicts a sample. It is a popular method used in language modeling. In neural dialog generation community, many researchers have adopted this method, especially in the beginning of this field (Vinyals and Le, 2015; Serban et al., 2016; Xing et al., 2018; Zhou and Wang, 2018; Zhou et al., 2018a; Zhong et al., 2019). It measures how well a dialog model predicts the target response. Given a target response $y = (y_1, y_2, \ldots, y_T)$, the perplexity is calculated as

$$\begin{aligned} \mathrm{PPL}(y) &= p(y_1, y_2, \ldots, y_T)^{-1/T} \\ &= \exp\left[-\frac{1}{T} \sum_{t=1}^{T} \log p(y_t \mid y_1, \ldots, y_{t-1}) \right]. \end{aligned} \tag{3.19}$$

Thus a lower perplexity score indicates that the model has better capability of predicting the target sentence, i.e., the humans' response. Some researchers (Shang et al., 2015; Li et al., 2016d; Zhong et al., 2019) argue that perplexity score is not the ideal measurement because for a given context history, one should allow many responses. This is especially true if we want our conversational agents to speak more diversely. However, for our purpose, which is to speak emotionally appropriately and as human-like as possible, we believe this is a good measure. We do recognize that it is not the only way to measure chatbots' performance. This is why we also conducted human evaluation experiment.

BLEU score is often used to measure the quality of machine-translated text. Some earlier work of dialog response generation (Li et al., 2016a,b) adopted this metric to measure the performance of chatbots. However, recent study (Liu et al., 2016) suggests that it does not align well with human evaluation. Nevertheless, we still include BLEU scores in this chapter, to get a sense of comparison with perplexity and human evaluation results.

Human Evaluation

Human evaluation has been widely used to evaluate open-domain dialog generation tasks. This approach can include any criterion as we judge appropriate. Most commonly, researchers have included the model's ability to generate grammatically correct, contextually coherent, and emotionally appropriate responses, of which the latter two properties cannot be reliably evaluated using automatic metrics. Recent work (Asghar et al., 2018; Zhong et al., 2019; Zhou et al., 2018a) on affect-rich conversational chatbots turned to human opinion to evaluate both fluency and emotionality of their models. But such human experiments are sensitive to risk factors if the experiment is not carefully designed. They include whether the intructions are clear, whether they have been tested with users before hand, and whether there is a good balance of the human judgement tasks. Further, if a test set for human evaluation is prepared by randomly sampling the dialogs from the dataset, it may include out-of-context dialogs, causing confusion and ambiguity for human evaluators. Unbalanced emotional distribution of the test dialogs may also lead to biased conclusions since the chatbot's abilities are evaluated

on the unrepresentative sample.

To take into account the above issues, we took several iterations to prepare the instructions and the test set before conducting the human evaluation experiment. Part of our test set comes from the DailyDialog dataset, which consists of meaningful complete dialogs. To compensate for the inbalance, we further curated more negative emotion dialogs so that the final set has equal emotion distributions. We provide the details about the test data preparation process and the evaluation experiment below.

Preparation of Natural Dialog Test Set We first selected the emotionally colored dialogs with exactly four turns from the DailyDialog dataset. In the dataset each dialog turn is annotated with a corresponding emotional category, including the neutral one. For our purposes we filtered out only those dialogs where more than a half of utterances have non-neutral emotional labels, resulting in 78 emotionally positive dialogs and 14 emotionally negative dialogs. We recruited two human workers to augment the data to produce more emotionally negative dialogs. Both of them were PhD students from our university (males, aged 24 and 25), fluent in English, and not related to the authors' lab. We found them via email and messaging platforms, and offered 80 CHF (or roughly US $80) gift coupons as incentive for each participant. The workers fulfilled the tasks in Google form[4] following the instructions and created five negative dialogs with four turns, as if they were interacting with another human, in each of the following topics: *relationships*, *entertainment*, *service*, *work and study*, and *everyday situations*. The Google form was released on 31 January 2019, and the workers finished their tasks by 4 February 2019. Subsequently, to form the final test set, we randomly selected 50 emotionally positive and 50 emotionally negative dialogs from the two pools of dialogs described above.

Human Evaluation Experiment Design In the final human evaluation of the model, we recruited four more PhD students from our university (1 female and 3 males, aged 22–25). Three of them are fluent English speakers and one is a native speaker. The recruitment proceeded in the same manner as described above; the raters were offered 80 CHF (or roughly US $80) per participant gift coupons for fulfilling the task, and extra 20 CHF (or roughly US $20) coupon was promised as a bonus to the rater judged to be the most serious. For the evaluation survey, we also leveraged Google form. Specifically, we randomly shuffled the 100 dialogs in the test set, then we used the first three utterances of each dialog as the input to the three models being compared (S2S, HRAN, and MEED), and obtain the respective responses. Dialog contexts and three models' responses were included into Google form. According to the context given, the raters were instructed to evaluate the quality of the responses based on three criteria:

[4]We provide the link to the form used for creating the dialogs: https://forms.gle/rPagMZYuYJ3M3Sq8A, hoping to help other researchers reproduce the same procedure. However, due to privacy concerns, we do not plan to release this dataset.

Table 3.2: Perplexity and average BLEU scores achieved by MEED, compared with S2S and HRAN. Avg. BLEU: average of BLEU-1, -2, -3, and -4. Validation set 1 comes from the Cornell dataset, and validation set 2 comes from the DailyDialog dataset.

	Perplexity			**Avg. BLEU**		
Model	**Valid Set 1**	**Valid Set 2**	**Test Set**	**Valid Set 1**	**Valid Set 2**	**Test Set**
S2S	43.136	25.418	19.913	1.639	2.427	3.720
HRAN	46.225	26.338	20.355	1.701	2.368	2.390
MEED	**41.862**	**24.341**	**19.795**	**1.829**	**2.635**	**4.281**

1. *Grammatical correctness*—whether or not the response is fluent and free of grammatical mistakes;
2. *Contextual coherence*—whether or not the response is context sensitive to the previous dialog history;
3. *Emotional appropriateness*—whether or not the response conveys the right emotion and feels as if it had been produced by a human.

For each criterion, the raters gave scores of either 0, 1 or 2, where 0 means bad, 2 means good, and 1 indicates neutral. For this survey, the Google form was launched on 12 February 2019, and all the submissions from our raters were collected by 14 February 2019.

3.4.4 Results and Analysis

In this subsection, we present the experimental results of the automatic evaluation metric as well as human judgement, followed by some analysis.

Automatic Evaluation Results

Table 3.2 gives the perplexity and BLEU scores obtained by the three models on the two validation sets and the test set. As shown in the table, MEED achieves the lowest perplexity and the highest BLEU score on all three sets. We conducted t-test on the perplexity and BLEU scores obtained, and results show significant improvements of MEED over S2S and HRAN (with p-value < 0.05).

Human Evaluation Results

Table 3.3, 3.4 and 3.5 summarize the human evaluation results on the responses' grammatical correctness, contextual coherence, and emotional appropriateness, respectively. In the tables, we give the percentage of votes each model received for the three scores, the average score obtained, and the agreement score among the raters. Note that we report Fleiss' κ

Table 3.3: Human evaluation results on grammatical correctness of MEED, compared with S2S and HRAN.

Model	+2	+1	0	Avg. Score	r
S2S	98.0	0.8	1.2	1.968	0.915
HRAN	98.5	1.3	0.2	1.982	0.967
MEED	99.5	0.3	0.2	**1.992**	0.981

Table 3.4: Human evaluation results on contextual coherence of MEED, compared with S2S and HRAN.

Model	+2	+1	0	Avg. Score	κ
S2S	25.8	19.7	54.5	0.713	0.389
HRAN	37.3	21.2	41.5	0.958	0.327
MEED	38.5	22.0	39.5	**0.990**	0.356

score (Fleiss and Cohen, 1973) for contextual coherence and emotional appropriateness, and Finn's r score (Finn, 1970) for grammatical correctness. We did not use Fleiss' κ score for grammatical correctness. As agreement is extremely high, this can make Fleiss' κ very sensitive to prevalence (Hripcsak and Heitjan, 2002). On the contrary, we did not use Finn's r score for contextual coherence and emotional appropriateness because it is only reasonable when the observed variance is significantly less than the chance variance (Tinsley and Weiss, 1975), which did not apply to these two criteria. As shown in the tables, we got high agreement among the raters for grammatical correctness, and fair agreement among the raters for contextual coherence and emotional appropriateness.[5] For grammatical correctness, all three models achieved high scores, which means all models are capable of generating fluent utterances that make sense. For contextual coherence and emotional appropriateness, MEED achieved higher average scores than S2S and HRAN, which means MEED keeps better track of the context and can generate responses that are emotionally more appropriate and natural. We first conducted Friedman test (Howell, 2016) and then t-test on the human evaluation results (contextual coherence and emotional appropriateness), showing the improvements of MEED over S2S are significant (with p-value < 0.01).

The comparison between perplexity scores and human evaluation results further confirms the fact that in the context of dialog response generation, perplexity does not align with human judgement. In Table 3.2, for all the three sets, HRAN performs worse than S2S in terms of perplexity. However, for all of the three criteria in human evaluation, HRAN actually outperforms S2S. Based on this, we conclude that perplexity alone is not enough for evaluating a dialog system.

[5] https://en.wikipedia.org/wiki/Fleiss%27_kappa#Interpretation

Table 3.5: Human evaluation results on emotional appropriateness of MEED, compared with S2S and HRAN.

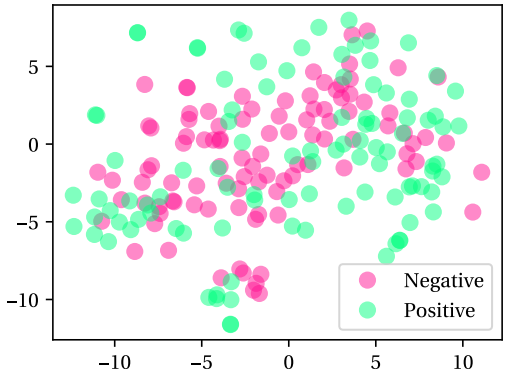
Model	+2	+1	0	Avg. Score	κ
S2S	21.8	25.2	53.0	0.688	0.361
HRAN	30.5	28.5	41.0	0.895	0.387
MEED	32.0	27.8	40.2	**0.917**	0.337

Visualization of Output Layer Weights

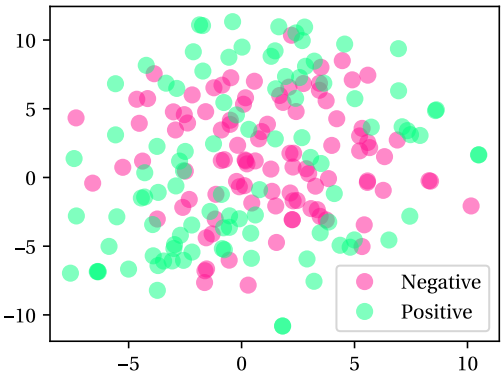
We may wonder how HRAN and MEED differ in terms of the distributional representations of their respective vocabularies (words in the language model, and affect words). We decided to visualize the output layer weights as word embedding representations using dimensionality reduction technique for the various models.

In the decoding phase, Equation (3.16) takes $\boldsymbol{o}_t$, the concatenation of the language context vector $\boldsymbol{s}_t$ and the emotion context vector $\boldsymbol{e}$, and generates a probability distribution over the vocabulary words by applying a softmax layer. The weight matrix of this softmax layer is denoted as $\boldsymbol{W}$, whose shape is $|V| \times 2d$, where $|V|$ is the vocabulary size and $d = 256$ is the hidden state size of the RNNs. Thus the ith row of the weight matrix $\boldsymbol{W}_i$ can be regarded as a vector representation of the ith word in the vocabulary. Since we concatenate the language context vector and the emotion context vector as the input to the softmax layer, the first half of the weight vector $\boldsymbol{W}_i$ corresponds to the language context vector, and the second half corresponds to the emotion context vector. We refer to them as language model weights and emotion weights, respectively. If the emotion embedding layer is learning and distinguishing affect states correctly, we will see clear differences in the visualization.

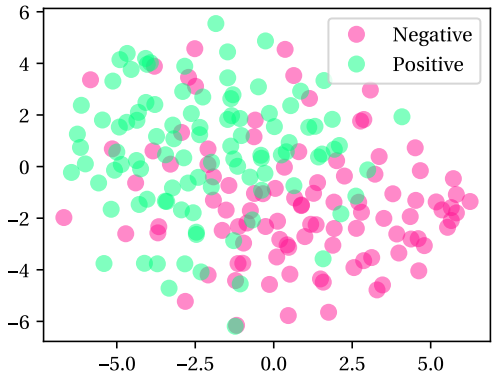
Using t-SNE (Maaten and Hinton, 2008) (with a perplexity value of 30.0), we are able to reduce the dimensionality of the weights to two, and visualize them in a straightforward way. For better illustration, we selected 100 most frequent (emotionally) positive words and 100 most frequent negative words from the vocabulary, and used t-SNE to project the corresponding language model weights and emotion weights to two dimensions. Figure 3.2 gives the results in three subplots. Since HRAN does not have the emotion context vector, we just visualized the whole output layer weight vector, which does a similar job as the language model weights in MEED. We can observe from the first two plots that positive words (green dots) and negative words (red dots) are scattered around and mixed with each other in the language model weights for HRAN and MEED respectively, which means no emotion information is captured in these weights. On the contrary, the emotion weights in MEED, in the last plot, have a clearer clustering effect, i.e., positive words are mainly grouped on the top-left, while negative words are mainly grouped at the bottom-right. This gives the hint that the emotion encoder in MEED is capable of tracking the emotion states in the conversation history.



(a) Language model weights of HRAN.



(b) Language model weights of MEED.



(c) Emotion weights of MEED.

Figure 3.2: t-SNE visualization of the output layer weights in HRAN and MEED. 100 most frequent positive words and 100 most frequent negative words are shown. The weight vectors in MEED are separated into two parts and visualized individually.

Case Study

We present four sample dialogs in Table 3.6, along with the responses generated by the three models. Dialog 1 and 2 are emotionally positive and dialog 3 and 4 are negative. For the first two examples, we can see that MEED is able to generate more emotional content (like "fun" and "congratulations") that is appropriate according to the context. For dialog 4, MEED responds in sympathy to the other speaker, which is consistent with the second utterance in the context. On the contrary, HRAN poses a question in reply, contradicting the dialog history.

3.5 Discussion

In this section, we briefly discuss how our framework can incorporate other components, as well as several directions to extend it.

Table 3.6: Sample responses generated by MEED, compared with S2S and HRAN. For each dialog, the ground truth (last turn) is included in a pair of parentheses.

#	Context	Model Responses
1	A: I'm happy to see you again. B: Mee too. A: We should do this more often. (B: Okay, I'll give you a ring next week.)	S2S: I hope so. HRAN: Thanks a lot. MEED: That sounds like fun.
2	A: Thank god! I am finished writing that service guide! It took me forever! B: When did you finish? A: This morning! No more overtime, and no more headaches! (B: Well, I'm glad to hear it. Have a cup of coffee!)	S2S: When? HRAN: Why? MEED: Congratulations!
3	A: I think that's settled. B: I'm tired of your cut-and-dried opinions. Who do you think you are! A: How dare you speak to me like this. (B: Why not?)	S2S: What is it? HRAN: What is it? MEED: Are you serious?
4	A: This concert was awful. B: Agreed, the musicians were not in harmony. A: It was too painful. Never again. (B: That's for sure!)	S2S: Congratulations! HRAN: Why not? MEED: That's true.

3.5.1 Emotion Recognition

To extract the affect information contained in the utterances, we used the LIWC text analysis program. We believe this emotion recognition step is vital for a dialog model to produce emotionally appropriate responses. However, the choice of emotion classifier is not strictly limited to LIWC. It could be replaced by other well-established affect recognizer or one that is more appropriate to the target domain. For example, we can consider using more fine-grained emotion categories from GALC (Scherer, 2005), or using DeepMoji (Felbo et al., 2017), which was trained on millions of tweets with emoji labels and is more suitable for tweet-like conversations. However, for DeepMoji, the 64 categories of emojis do not have a clear and exact correspondence with standardized emotion categories, nor to the VAD vectors.

3.5.2 Training Data

We pre-trained our model on the Cornell movie subtitles and then fine-tuned it with the DailyDialog dataset. We adopted this particular training order because we would like our chatbot to talk more like human chit-chats, and the DailyDialog dataset, compared with the bigger Cornell dataset, is more daily-based. Since our model learns how to respond properly in a data-driven way, we believe having a training dataset with good quality while being large enough plays an important role in developing an engaging and user-friendly chatbot. Thus, we could train our model on the multi-turn conversations extracted from the much bigger

OpenSubtitles corpus and then fine-tune on the EmpatheticDialogues dataset.[6]

3.5.3 Evaluation

Evaluation of dialog models remains an open problem in the response generation field. Early work (Ritter et al., 2011; Sordoni et al., 2015b; Li et al., 2016b) on response generation used automatic evaluation metrics borrowed from the machine translation field, such as the BLEU score, to evaluate dialog systems. Later on, Liu et al. (2016) showed that these metrics correlate poorly with human judgement. Recently, a number of researchers begain developing automatic and data-driven evaluation methods (Lowe et al., 2017; Tao et al., 2018), with the ultimate goal of replacing human evaluation. However they are still in an early stage. In this chapter, we used both perplexity measures and human judgement in our experiments to finalize our model. In other words, using the perplexity measures, we were able to determine when to stop training our model. But this condition does not gurantee the optimal results until human judgement test can validate them. We thus highly recommend this combination, which is also a common practice in the research community (Xing et al., 2018; Zhou and Wang, 2018; Zhou et al., 2018a; Zhong et al., 2019).

3.5.4 Model Extensions

Our model uses RNNs to encode the input sequences, and GRU cells to capture long-term dependency among different positions in the sequences. Recent advances in natural language understanding have proposed new network architectures to process text input. Specifically, the Transformer (Vaswani et al., 2017) uses pure attention mechanisms without any recurrence structures. Compared with RNNs, the Transformer can capture better long-term dependency due to the self-attention mechanism, which is free of locality biases, and is more efficient to train because of better parallelization capability. Following the Transformer architecture, researchers found that pre-training language models on huge amounts of data could largely boost the performance of downstream tasks, and published many pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Therefore, we could adopt the Transformer architecture to replace the RNNs in our model, and initialize our encoder with pre-trained language models. We hope to increase the performance of response generation.

3.6 Chapter Summary

We believe reproducing conversational and emotional intelligence will make social chatbots more believable and engaging. In this chapter, we proposed a multi-turn dialog system capable of recognizing and generating emotionally appropriate responses, which is the first step toward such a goal. We have demonstrated how to do so by (1) modeling utterances with extra affect vectors, (2) creating an emotional encoding mechanism that learns emotion exchanges in the

[6] https://github.com/facebookresearch/EmpatheticDialogues

dataset, (3) curating a multi-turn and balanced dialog dataset, and (4) evaluating the model with offline and online experiments. For future directions, we would like to investigate the diversity issue of the responses generated, possibly by extending the mutual information objective function (Li et al., 2016a) to multi-turn settings. As an extension of the MEED model, we can adopt the Transformer architecture (possibly with pre-trained language model weights), and train our model on a much larger dataset, by extracting multi-turn dialogs from the OpenSubtitles corpus.

4 Empathetic Dialog Generation with Fine-Grained Intents

This chapter is based on the work of Yubo Xie and Pearl Pu (Xie and Pu, 2021). The author of this thesis (Yubo Xie) was mainly responsible for curating the data, designing, implementing, and training the dialog model, and its evaluation.

4.1 Introduction

Empathy is considered to be an innate ability of human beings (Roth-Hanania et al., 2011) and plays an important role in people's social communication (Valente, 2016). It has been shown that integrating empathy into dialog systems could improve user experience for human-computer interaction (Liu and Picard, 2005). One of the empathetic components is the capacity to respond with an appropriate emotion to another person's mental states (Shamay-Tsoory et al., 2009). In this regard, many existing neural dialog systems (Zhou et al., 2018a; Huang et al., 2018; Zhou and Wang, 2018; Colombo et al., 2019; Song et al., 2019; Shen and Feng, 2020) generate emotional responses conditioned on a pre-specified emotion label. However, this might be impractical when deploying the chatbots in reality, since an extra label is required as input. Other neural dialog systems (Asghar et al., 2018; Li and Sun, 2018; Zhong et al., 2019; Lin et al., 2019; Li et al., 2022a) adopt manually defined rules, either explicitly or implicitly, to decide the emotion state for the response to be generated, e.g., following/reversing the speaker's emotion, or just maximizing the emotion content in the response. However, such deterministic rules are not confirmed by psychology literature, and they ignore the subtle interactions captured in human conversations, where the listener often exhibits empathetic intents that are more neutral. Figure 4.1 gives an example of a situation where responding with the same or opposite emotion fails to drive the conversation towards an empathetic direction. In fact, as revealed by Welivita and Pu (2020), listeners are much more likely to respond with *questioning* to sad or angry emotions of another person, than expressing similar or opposite emotions.

Therefore, it is necessary to incorporate these additional empathetic response intents explicitly into the design of dialog systems. Existing neural dialog systems adopt an empathetic dialog

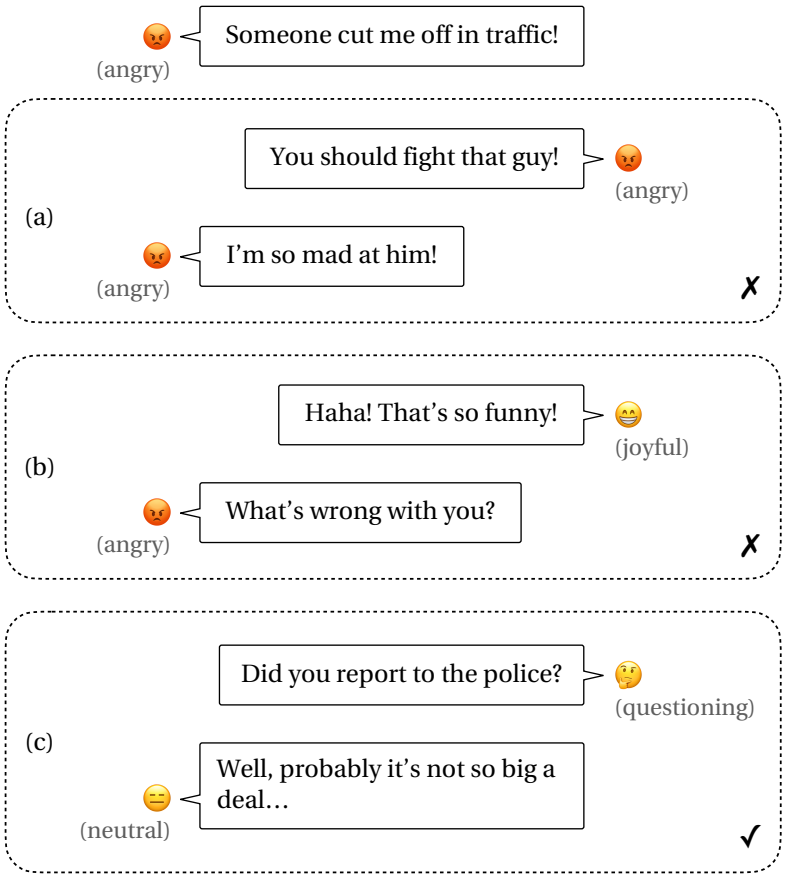


Figure 4.1: Three ways of responding to a speaker's utterance. Note that simply following the speaker's emotion state (a) or reversing it (b) still leaves the speaker in angry state (or even escalates the situation). Responding with questioning (c) successfully calms down the speaker and drives the conversation to a more manageable direction.

dataset that either has no neutral category (Rashkin et al., 2019), or the neutral category is a conglomerate of intents that cannot be clearly defined. This is why this category is often called *other*, which shows it is not sufficiently treated (Chatterjee et al., 2019; Li et al., 2017). While Xu et al. (2018) proposed to model open-domain dialog generation as the selection of dialog acts that control the generation of responses, they did not specifically focus on the generation of empathetic dialogs.

Based on the taxonomy proposed by Welivita and Pu (2020), we incorporated an extra set of eight empathetic response intents (*questioning*, *agreeing*, *acknowledging*, *sympathizing*, *encouraging*, *consoling*, *suggesting*, and *wishing*) plus *neutral* into the design of an empathetic dialog model, in addition to the 32 emotion categories proposed by Rashkin et al. (2019). Emotional experience is primarily a reaction to an external event, for example, a loud sound, a surprising result on an exam, etc. In the case of dialogs, this emotional experience is shared by the interlocutors. When a listener wants to acknowledge or console the speaker, for example, he or she is expressing an emotional intent. This is the reason we treat all of these additional

categories as well as the 32 emotion categories as dialog intents.

Overall speaking, our contributions are as follows: (1) We are the first to consider modeling a fine-grained set of empathetic response intents in an empathetic dialog model, which ensures a more precise learning of the emotional interactions revealed in the dialog data; (2) To facilitate the training of our empathetic dialog model, we curated a large-scale dialog dataset from movie subtitles; (3) To effectively evaluate our empathetic dialog model, we carefully designed a crowdsourcing experiment that enabled the workers to work on the tasks more easily. A total number of 6,000 dialogs were evaluated, which, to our knowledge, has never been attempted before for the evaluation of empathetic dialog systems.

4.2 Related Work

4.2.1 Empathetic Dialog Generation

Lubis et al. (2018) designed a hierarchical encoder-decoder model that captures the user's emotion state and takes it into account when generating the response. Shin et al. (2020) adopted a reinforcement learning framework that provides a higher reward to the generative model if it promotes the user's future emotion state. Li et al. (2020) adopted an adversarial learning framework and proposed two discriminators to evaluate if the generated response is empathetic and elicits more positive emotions by considering the emotion words in the gold response and the next reply. However, these models do not have a clear mechanism for controlling the emotion state of the generated response. Wei et al. (2019) is the closest to our work, but they only considered a limited number of emotion categories, and thus the model is not able to convey certain subtle emotions and different empathetic response intents of the listener can not be effectively learned.

In Chapter 3, we proposed MEED, a multi-turn emotionally engaging dialog model by modeling the emotion states in the dialog history. Compared with MEED, MEED2 introduces more fine-grained emotion categories and an additional set of empathetic response intents that are more neutral. Moreover, we learn the emotion interactions more explicitly, which allows the model to have more controllability and interpretability.

4.2.2 Emotional Dialog Datasets

Most of the existing emotional dialog datasets are small in size and have limited number of emotion categories. Li et al. (2017) created the DailyDialog dataset from English learning websites, consisting of 13K multi-turn dialogs manually labeled with 7 emotions. The EmotionLines dataset (Hsu et al., 2018) contains 2,000 dialogs collected from Friends TV scripts and EmotionPush chat logs, labeled with 7 emotions. Poria et al. (2019) extended the EmotionLines dataset to a multimodal setting, containing 1,433 dialogs from Friends TV scripts. Chatterjee et al. (2019) proposed the EmoContext dataset collected from users' interaction

Table 4.1: Statistics of the OpenSubtitles dialogs after cleaning.

Total number of dialogs	4,010,009
Total number of turns	18,849,440
Total number of tokens	312,574,468
Average number of turns per dialog	4.70
Average number of tokens per turn	16.58
Average number of tokens per dialog	77.95

with a conversational agent, which contains 38K dialogs labeled with 4 emotions. Rashkin et al. (2019) curated the EmpatheticDialogues dataset containing 25K dialogs collected from a crowdsourcing platform by letting workers communicate with each other based on 32 emotion categories.

4.3 Data Curation

Existing empathetic dialog corpora are usually limited in size and training solely on these datasets could not give us a chatbot with desirable performance. Therefore, we would like to take advantage of transfer learning and pre-train the dialog model on a huge amount of dialog data (not necessarily empathetic), and then fine-tune it on a possibly much smaller empathetic dialog dataset.

4.3.1 Extracting Dialogs from Movie Subtitles

To obtain a large-scale dialog dataset, we relied on the OpenSubtitles2018 corpus (Lison et al., 2018), which contains text collected from movie subtitles spread over 60 languages, and is a good source of human conversations written by professional screenwriters. We only used the English part, which has 447K subtitle files, 441M sentences and 3.2B tokens. Due to the lack of speaker information in the OpenSubtitles corpus, before extracting the dialogs, we followed the same procedure proposed by Lison and Meena (2016) and built an SVM classifier to determine whether two consecutive lines in one subtitle file are actually spoken by the same character and should be in the same dialog turn. As a result, we obtained a turn segmentation accuracy of 76.69%.

We then separated these turns into dialogs by adopting a heuristic rule based on timestamps: for each subtitle file, we calculate the gap between the starting time of each turn and the ending time of its previous turn. If this time gap is greater than 5 seconds, we cut off at this position and regard these two turns as belonging to different dialogs. An exception is when the timestamp information is missing for one of the two turns. In this case, we just regard them as belonging to one dialog. In this way, we obtained 9M dialogs from the whole English OpenSubtitles corpus. To further clean the dataset, we applied a sequence of steps to remove undesirable utterances. As a result, we obtained 4M cleaned OpenSubtitles dialogs.

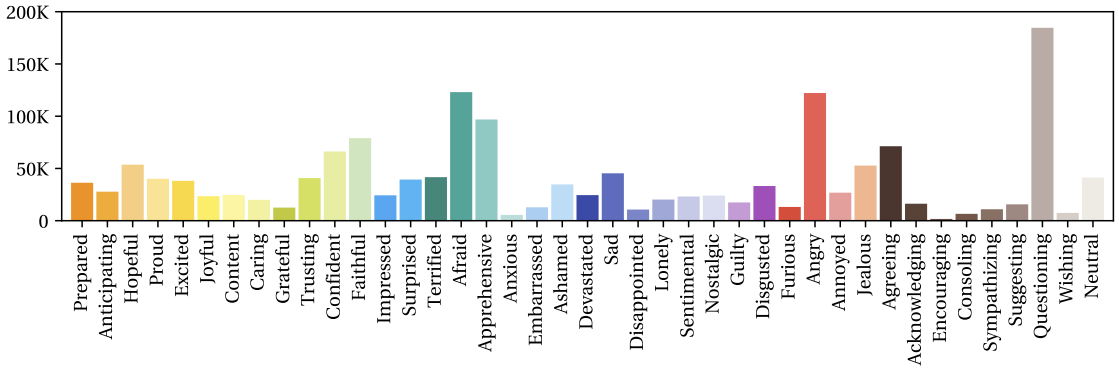


Figure 4.2: Distribution of emotions/intents in the emotional dialogs in OpenSubtitles.

See Appendix A.1 for the detailed cleaning procedure. Table 4.1 lists some statistics of the OpenSubtitles dialogs.

4.3.2 Emotional Dialogs in OpenSubtitles

Many existing emotional dialog datasets are small in size due to the expensive procedure of data collection, usually done manually by human. In this chapter, we created a large-scale empathetic dialog dataset by first training a sentence-level fine-grained emotion classifier and then selecting out emotional dialogs from the cleaned OpenSubtitles dataset aforementioned.

To build the emotion classifier, we followed Welivita and Pu (2020) and fine-tuned RoBERTa (Liu et al., 2019) on the situation sentences from the EmpatheticDialogues (Rashkin et al., 2019) training set (labeled with 32 fine-grained emotions), and 7K listener utterances labeled with 8 empathetic intents (*questioning*, *agreeing*, *acknowledging*, *sympathizing*, *encouraging*, *consoling*, *suggesting*, and *wishing*) plus one *neutral* category (all other not mentioned intents). The 7K intent-labeled utterances were obtained by first manually labeling 521 sentences and then expanding through searching most frequent n-grams for each intent. The classifier achieved an accuracy of 65.88% on the EmpatheticDialogues test set. We applied the obtained classifier on all cleaned OpenSubtitles dialogs, and calculated a probability distribution over the 41 categories for each utterance. We then define the emotionality of each utterance as the sum of the probability values of the 32 emotion categories, and the emotionality of each dialog as the averaged emotionality values of its utterances. We selected the top 1M dialogs with highest emotionality values to form the dataset of emotional dialogs in OpenSubtitles. Figure 4.2 gives the distribution of emotions/intents of the last utterance. Some samples of the OpenSubtitles dialogs can be found in Appendix A.4. The datasets along with the code of our model are publicly available.[1]

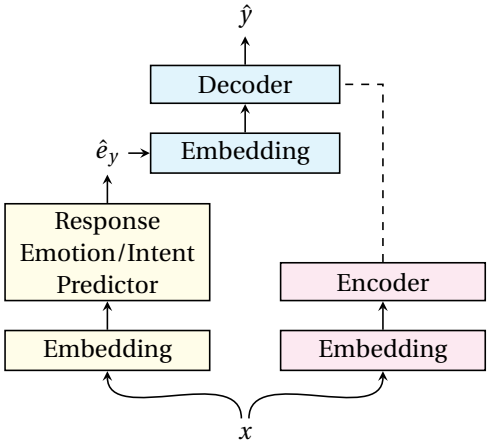


Figure 4.3: Overall architecture of MEED2 showing how the model works in inference mode. Dashed line denotes multi-head attention.

4.4 An Empathetic Dialog Model

We propose an empathetic dialog model that incorporates the fine-grained set of empathetic response intents, by training a classifier that predicts the response emotion/intent, and based on that, generates the response accordingly. Compared with our previous multi-turn emotionally engaging model (MEED), this model uses Transformer (Vaswani et al., 2017) as its backbone, which allows for more parallelization in the training procedure, and is able to capture longer dependency in the input than the RNN architecture. We also designed a response emotion/intent predictor that explicitly predicts the emotion/intent of the response to be generated. This allows for more interpretability of the responses produced by the model—in addition to the generated response, the model also gives an emotion/intent label that it deems the most appropriate to respond to the user's input, which ideally should be compatible with the generated response. Furthermore, one could also replace this predicted emotion/intent label with a custom one and feed it into the decoder to generate the desired response. In this sense, the model becomes more controllable, due to the separately trained response emotion/intent predictor.

The problem could be defined as follows: given a dialog context x consisting of one or more utterances $u_1, u_2, \ldots, u_m$, spoken between two people, try to generate a response $\hat{y}$ that not only follows the dialog context but also is emotionally appropriate. Our model consists of three modules: (1) an encoder responsible for encoding the input x into vector representations; (2) a response emotion/intent predictor which takes x as input and decides in which emotion/intent the model should respond; (3) a decoder responsible for generating the actual response. We use Transformer encoder structure for our encoder and emotion/intent predictor, and Transformer decoder structure for our decoder. Figure 4.3 gives an overall depiction of the whole model architecture. All the three modules have the same input representation, which

[1] https://github.com/yuboxie/meed2

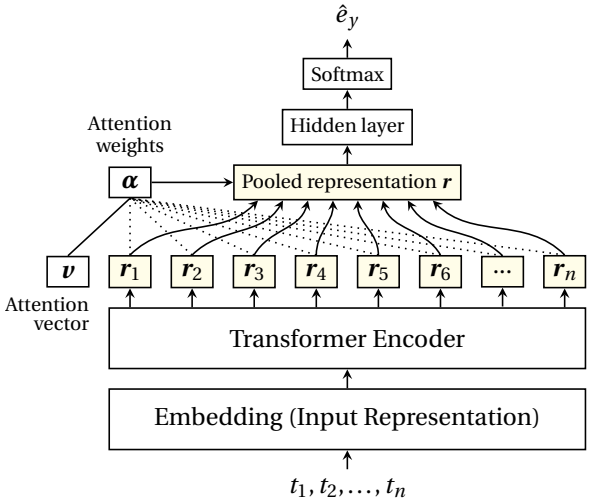


Figure 4.4: A detailed illustration of the response emotion/intent predictor in MEED2. Dotted lines denote attention mechanism.

we describe in detail next.

4.4.1 Input Representation

The input representation is illustrated in Figure 4.5. We use the RoBERTa tokenizer to tokenize the utterances $u_1, u_2, \ldots, u_m$ in the input dialog context x, and concatenate them by two special tokens: `<s>` and `</s>`, as shown in the figure. For our model to have a better understanding of the input dialog context, in addition to the word embeddings and position embeddings in the original Transformer architecture, we also have emotion embeddings. Specifically, for each utterance u_i, we use the same emotion classifier described in Section 4.3.2 to obtain an emotion representation in the form of a probability distribution on 41 emotions/intents. The label with maximum probability value is denoted as e_{u_i}, representing the emotion/intent expressed by utterance u_i. Similar to word embeddings, we embed this emotion/intent e_{u_i} into a vector space with the same dimensionality as other embeddings, so that they could add up. The same emotion embedding is used for all the tokens in the same utterance. To further differentiate between the speakers, we augment the input representation with segment embeddings. Utterances spoken by the same person would have the same segment embedding. The encoder and decoder share the same embedding tables.

4.4.2 Response Emotion/Intent Predictor

We relied on a data-driven approach to decide the emotion/intent of the response to be generated, by designing an emotion/intent classifier to predict the emotion/intent of the ground-truth response y, based on the context x. As shown in Figure 4.4, we use a Transformer encoder to get a context-dependent vector representation $\boldsymbol{r}_i$ for each of the input token

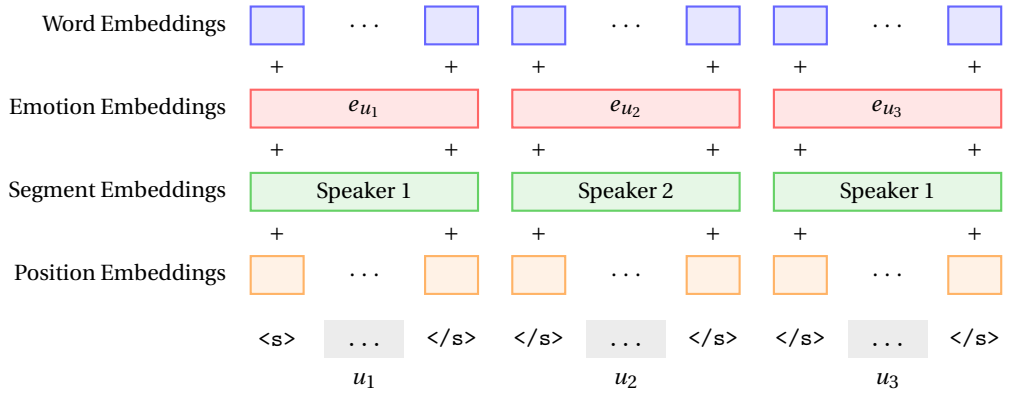


Figure 4.5: Input representation of the MEED2 model, which is the sum of four types of embeddings: word embeddings, emotion embeddings, segment embeddings, and position embeddings.

t_i. To pool these high-level representations into a single vector, we use a simple attention mechanism and incorporate a trainable vector $\boldsymbol{v}$ to obtain an attention weight α_i for $\boldsymbol{r}_i$,

$$\alpha_i = \frac{\exp(\boldsymbol{v}^T \boldsymbol{r}_i)}{\sum_{j=1}^{n} \exp(\boldsymbol{v}^T \boldsymbol{r}_j)}. \tag{4.1}$$

The aggregate representation $\boldsymbol{r}$ is then

$$\boldsymbol{r} = \sum_{i=1}^{n} \alpha_i \boldsymbol{r}_i. \tag{4.2}$$

$\boldsymbol{r}$ is fed into a hidden layer followed by a softmax layer to produce $\hat{e}_y$, denoting the predicted emotion/intent of the response to be generated.

4.4.3 Training

The response emotion/intent predictor is trained separately from the encoder/decoder, which means the training phase is a bit different from what is illustrated in Figure 4.3. In particular, the response emotion/intent predictor is independently trained to minimize the cross entropy loss of $\hat{e}_y$ with respect to e_y (true emotion/intent of y). While training the encoder and decoder simultaneously, we just feed e_y into the embedding layers of the decoder, and try to minimize the cross entropy loss of $\hat{y}$ with respect to y.

We also experimented with jointly training the response emotion/intent predictor and the encoder/decoder, by combining two loss functions like in a multi-task setting. However, we found the generated responses quite generic compared with training the two components separately, plus joint training also introduces more hyperparameters to be tuned. Moreover, having them trained separately endows the decoder with more controllability—the decoder is able to generate responses according to a specified emotion/intent label.

4.5 Evaluation

We trained our empathetic dialog model and the baselines on three datasets and evaluated them in *held-out* setting (meaning the test data comes from the same domain as the training data) and *zero-shot* setting (meaning the test data comes from a different domain than the training data), using both automatic metrics and human judgement via crowdsourcing.

4.5.1 Datasets

Three datasets were involved in the evaluation:

- **OpenSubtitles dialogs**. As described in Section 4.3.1, these dialogs were obtained by segmenting the movie subtitles. Note that for the purpose of pre-training, we excluded the emotional dialogs in OpenSubtitles (containing 1M dialogs), resulting in around 3M dialogs. We denote this dataset as OS.
- **Emotional dialogs in OpenSubtitles**. The curation process is described in Section 4.3.2. The total number of dialogs is 1M. We denote this dataset as EDOS.
- **EmpatheticDialogues dataset**. This dataset is created by Rashkin et al. (2019) and contains 24,850 dialogs collected from crowdsourcing. We denote this dataset as ED.

We split each of the three datasets into training set (80%), validation set (10%), and test set (10%). Among the dialogs of each test set, we further randomly selected out 2,000 to form a combined test set of 6,000 dialogs, for the purpose of evaluating the models on automatic metrics and human judgement via crowdsourcing.

4.5.2 Baselines

Similar to the work of Rashkin et al. (2019), we adopted the full Transformer model as our baseline, and based on the training strategies, we have the following variants:

- **Pre-trained**. To take advantage of transfer learning, we pre-trained the full Transformer model on the curated OS dataset, which contains around 3M dialogs. The large scale of this training set is expected to provide a good starting point for fine-tuning.
- **Fine-tuned**. We took the pre-trained full Transformer, and then fine-tuned it on two smaller dialog datasets: our curated EDOS dataset, and the ED dataset, respectively.
- **Raw**. To test the effectiveness of pre-training, we directly trained the full Transformer on the ED dataset, and then compared it with the fine-tuned models.

Note that we did not include the EmoPrepend-1 model by Rashkin et al. (2019) as our baseline, because in their paper, its human evaluation performance is actually reported to be worse

than the fine-tuned Transformer. All the models have a hidden size of 300, and were trained until the minimum validation loss was reached. For inference we used beam search with beam size 32 and 4-gram repeats blocking. Further details regarding the implementation parameters can be found in Appendix A.2.

4.5.3 Automatic Evaluation

Most of the existing automatic metrics directly compare the generated response with the ground-truth provided by human, often in a simple way. Due to the inherent diversity of human conversations, this is not suitable for dialog models, since for the same prompt, there could exist many responses that are equally good. In fact, Liu et al. (2016) has shown that word-overlap-based metrics (specifically BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004)) and word embedding metrics all exhibit weak or no correlation with human judgements. To this end, we did not adopt these metrics in our experiment, but instead considered the following:

- **Perplexity**. Perplexity is a model-dependent metric that measures how well a probability model predicts a given sample. In our case, a lower perplexity score indicates better capability of generating the ground-truth response.
- **Distinct-1 and -2**. The Distinct-1 and -2 metrics (Li et al., 2016a) measure the diversity of the generated responses by calculating the ratio of unique unigrams or bigrams over the total number of unigrams or bigrams in the generated responses.
- **Sentence Embedding Similarity**. For this metric, we use Sentence-BERT (Reimers and Gurevych, 2019) to obtain an embedding for the generated response as well as the ground-truth, and then calculate the cosine similarity between the two embeddings.

The results of automatic evaluation are shown in Table 4.2. Our model (MEED2) achieves lower perplexity scores than the corresponding full Transformer on all the three datasets. Here we have an extra model configuration, Raw (ED), to compare with Fine-tuned (ED), in order to see the effects brought by pre-training. As we can see, without pre-training on OS, the model gets much worse performance on the perplexity scores. This indicates that pre-training and then fine-tuning is preferred to directly training on a target dataset. On Distinct-1 and -2, our model always has a higher score than the corresponding full Transformer model, suggesting that by injecting additional emotion information, the dialog system could be guided to generate more diverse responses. We also observe that on the ED dataset, our model fine-tuned on EDOS actually has the highest Distinct scores, even though it has never seen the ED data. We conjecture that this is because the EDOS dataset is much bigger than the ED dataset, and contains text that is more diverse. Table 4.3 lists the weighted precision, recall, and F-1 scores of the response emotion/intent predictor for different model configurations.

If we consider a zero-shot setting, meaning the model is evaluated on data from a different

Table 4.2: Automatic evaluation results of MEED2 and its baselines. Here PPL denotes perplexity, D1 and D2 denote Distinct-1 and -2, and SES denotes the sentence embedding similarity. X→Y means pre-training on X and then fine-tuning on Y.

	OS				EDOS				ED			
Model	**PPL**	**D1**	**D2**	**SES**	**PPL**	**D1**	**D2**	**SES**	**PPL**	**D1**	**D2**	**SES**
Pre-trained (OS)	24.8	0.046	0.159	0.172	37.8	0.046	0.154	0.126	564.6	0.044	0.167	0.178
Fine-tuned (EDOS)	26.9	0.044	0.139	0.162	32.3	0.056	0.165	0.137	452.6	0.031	0.107	0.176
Fine-tuned (ED)	88.9	0.030	0.109	**0.174**	140.8	0.028	0.096	0.130	19.3	0.026	0.091	**0.316**
Raw (ED)	793.9	0.009	0.032	0.144	1615.0	0.008	0.027	0.098	35.8	0.008	0.029	0.278
MEED2 (OS)	**22.0**	**0.064**	**0.210**	0.168	31.9	0.061	0.197	0.130	487.3	0.046	0.171	0.174
MEED2 (OS→EDOS)	22.8	0.057	0.196	0.168	**28.5**	**0.070**	**0.225**	**0.171**	391.7	**0.051**	**0.199**	0.207
MEED2 (OS→ED)	84.3	0.038	0.153	0.165	125.7	0.036	0.138	0.116	**17.2**	0.036	0.140	0.299

Table 4.3: Weighted precision, recall and F-1 scores of the response emotion/intent predictor in MEED2 on the three datasets. X→Y means pre-training on X and then fine-tuning on Y.

	OS			EDOS			ED		
Model	**P**	**R**	**F-1**	**P**	**R**	**F-1**	**P**	**R**	**F-1**
Random	.1484	.0240	.0285	.0382	.0250	.0266	.0989	.0165	.0215
MEED2 (OS)	**.2210**	**.3960**	.2312	.0109	.1040	.0198	.0942	.3070	.1442
MEED2 (OS→EDOS)	.2012	.1480	.1537	**.1029**	**.1495**	**.0917**	.1288	.2630	.1674
MEED2 (OS→ED)	.2166	.3265	**.2502**	.0253	.0870	.0239	**.2660**	**.3530**	**.2864**

Table 4.4: Human evaluation results of MEED2 and its baselines on each of the three test sets. Numbers have been normalized across the three quality categories on each test set. X→Y means pre-training on X and then fine-tuning on Y.

	OS			EDOS			ED		
Model	**Good**	**Okay**	**Bad**	**Good**	**Okay**	**Bad**	**Good**	**Okay**	**Bad**
Pre-trained (OS)	.3097	.2878	.4025	.2975	.2933	.4091	.1799	.3037	.5164
MEED2 (OS)	.3166	.3158	.3676	.3073	.3288	.3639	.1863	.3088	.5049
MEED2 (OS→EDOS)	.3175	.3036	.3789	.2926	.3034	.4040	.2097	.2891	.5012
MEED2 (OS→ED)	**.3513**	.3125	.3362	**.3535**	.3093	.3372	**.4890**	.3033	.2077

domain than its training data, we see from Table 4.2 that all models achieves higher perplexity scores on zero-shot test data. In particular, models trained on the OS (EDOS) dataset achieves lower perplexity on the EDOS (OS) dataset, compared with the results on the ED dataset. This is because OS and EDOS dialogs are actually curated from the same source, while the source of ED data is quite different. Moreover, models trained on EDOS has better perplexity scores on OS dataset, due to the performance boost brought by fine-tuning.

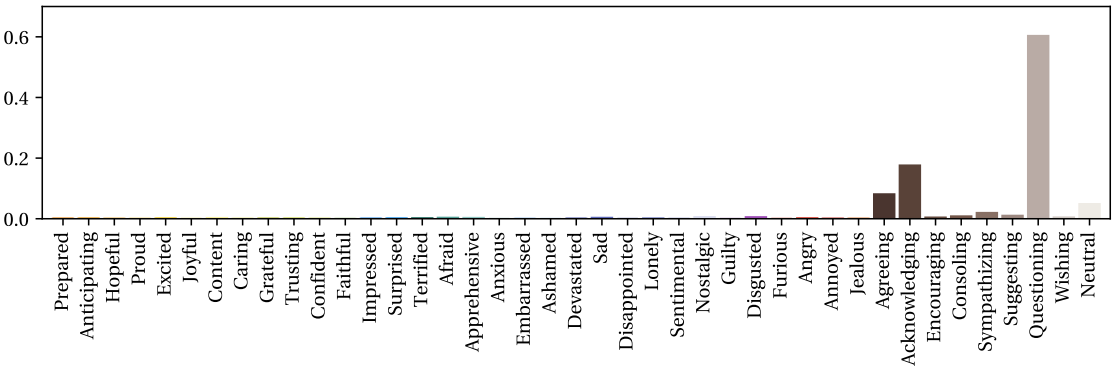


Figure 4.6: Distribution of emotions/intents in the responses generated by MEED2 (OS→ED) rated as *good*.

Table 4.5: Some samples of the responses generated by MEED2 and its baselines. X → Y means pre-training on X and then fine-tuning on Y.

Context (from OS): – Well, now, you know, most of your average people... won 't know a venomous reptile from a large garden worm. **Pre-trained (OS)**: I don 't know what you're talking about. **MEED2 (OS) (questioning)**: What do you mean, you won 't know? **MEED2 (OS → EDOS) (terrified)**: I'm not afraid of snakes. **MEED2 (OS → ED) (terrified)**: Oh my gosh, I hate snakes. What kind of snake was it? **Ground-truth (afraid)**: Oh, no. It has to be an absolute, authentic, poisonous snake... or he won 't work with it.
Context (from EDOS): – Yes, it's been a long wait. The second half of King of Comedy will now begin! – Hyo-jin. Yes. I'm excited to see what stars the second half will bring. – Yes. Welcome our first contestants. Who are they? **Pre-trained (OS)**: The second half will bring the second half will be the second half. **MEED2 (OS) (questioning)**: Who are they? What do they look like? **MEED2 (OS → EDOS) (excited)**: Let's begin! **MEED2 (OS → ED) (excited)**: I can't wait to see who they are. **Ground-truth (wishing)**: Warriors. Welcome!
Context (from ED): – Even though I could've gotten more money at another job, I've stayed with my current employer. It is a good place to work. – well that's good to hear I guess. what do you do? – I work in public relations. I create campaigns for companies. **Pre-trained (OS)**: what do you do? **MEED2 (OS) (questioning)**: what do they do in public relations? **MEED2 (OS → EDOS) (questioning)**: What are you doing in public affairs? **MEED2 (OS → ED) (acknowledging)**: that sounds like a great job. **Ground-truth (questioning)**: what's your most successful campaign so far?

4.5.4 Human Evaluation via Crowdsourcing

Human evaluation for dialog models has been widely adopted due to the limitations of automatic metrics. However, the experiment should be carefully designed so that the raters clearly understand the instructions and are constantly engaged in the evaluation tasks. Moreover, most of the existing work only recruited a limited number of raters to evaluate a test set of small size, therefore leading to possibly biased results. In this chapter, we carefully designed a human evaluation experiment that enables the raters to work on the evaluation tasks more easily and at the same time keeps them engaged by incorporating bonus checkpoints.

A New Evaluation Strategy

We conducted our human evaluation experiment on Amazon Mechanical Turk (MTurk). The 6,000 test dialogs were randomly shuffled and then split into 600 Human Intellligence Tasks (HITs), with each HIT containing 10 dialogs to be evaluated. For each test dialog, we included

the generated responses from four candidate models, i.e., Pre-trained (OS), MEED2 (OS), MEED2 (OS → EDOS), and MEED2 (OS → ED). Existing human experiments in dialog evaluation adopt either Likert Scale or side-by-side comparison (A/B testing). Likert Scale allows accurate evaluation of single items, but lacks reference and comparison; while A/B testing allows comparison, it doesn't scale. Our method is the first one that combines these two strategies and leverages on the merits of both. We allow the workers to drag and drop multiple candidate responses to one of the three pre-defined areas: *good*, *okay*, and *bad*, according to whether the response is emotionally appropriate following the given dialog context. In this way, it is easier for the workers to finish the tasks, and we also benefit from the accurate scoring results. In order to make the workers more engaged in the evaluation, and also encourage those providing high-quality answers, for each HIT we attached a bonus task to three ED dialogs, by adding the ground-truth response as a candidate. If the worker successfully put the ground-truth into the *good* or *okay* category, he or she will receive a bonus point. We gave a bonus of $0.1 to those workers who obtained all the three bonus points. More details of the human evaluation setup, including screenshots of the interface, can be found in Appendix A.3.

Human Evaluation Results

In total we received 24,000 answers from the MTurk experiment (4 answers for each of the 6,000 dialogs). We discarded answers from low-quality workers, i.e., those who provided the same answer for almost all dialogs, and those who completed the tasks in less than five minutes and failed to obtain at least two bonus points. Then, to calculate the human evaluation scores, we further selected out those assignments with at least two bonus points, and obtained a total number of 21,630 answers. The human evaluation results on the three individual test sets are shown in Table 4.4. From the table we see that our model outperforms the full Transformer on all three datasets (Pre-trained (OS) v.s. MEED2 (OS)), and of all the four model configurations, our model trained on ED achieves the highest percentage of good response on all three datasets, meaning training on ED enables the model to gain both good held-out performance and good zero-shot performance. Compared with our model only pre-trained on OS, it achieves better performance on OS and ED if fine-tuned on EDOS, but not on EDOS itself, meaning this model has a good zero-shot performance but the held-out performance is somehow lower. This could be explained by the unbalanced emotion/intent distribution in the OS dataset. As discussed in Section 4.5.3, for our model trained on OS, the response emotion/intent predictor would usually predict the dominating "questioning" category. For EDOS dialogs, since the response emotion/intent is more difficult to predict, responding in questions is probably safer.

We also investigated the distribution of emotions/intents in the generated responses, to see which emotions/intents are more preferred by the workers. For responses generated by MEED2 (OS → ED) that are rated as *good*, we gathered the predicted emotions/intents and calculated a probability distribution over the 41 categories, which is shown in Figure 4.6. We can see that *questioning*, *acknowledging* and *agreeing* are the major categories. This shows

that our model tends to generate responses with the empathetic intents, and they are indeed more preferred by the human evaluators.

4.5.5 Case Study

In this section, we give some sample responses generated by the models in Table 4.5. We took one dialog from each test set (OS, EDOS and ED). We can observe that most of the generated responses are syntactically correct (exceptional cases are from Pre-trained (OS)). The models could understand the dialog context and generate appropriate responses. For example, in the first dialog, our models fine-tuned on EDOS and ED recognize and understand the word "reptile" in the context, and then as response, generate the word "snakes." We can also observe from the table that the response emotions predicted by our models (fine-tuned on EDOS and ED) are reasonable and follow the emotions embedded in the dialog context. Moreover, the generated responses are indeed consistent with the predicted emotions. Note that our model trained on OS has a big chance of predicting the "questioning" category, which is due to the unbalanced distribution in the training set. More samples of the generated responses can be found in Appendix A.4.

4.6 Chapter Summary

In this chapter, we emphasize the importance of incorporating more fine-grained empathetic response intents into the design of empathetic dialog models. To this end, we proposed an empathetic dialog model capable of learning the emotion/intent interactions from the dialog data at a more precise level, and producing empathetic responses accordingly. To facilitate the training process, we also curated a large-scale dialog dataset from the OpenSubtitles corpus. Pre-training dialog models on this dataset could largely boost the performance of down-stream empathetic response generation. Our model was evaluated through a carefully designed human evaluation experiment on the crowdsourcing platform, on a large test set never attempted before. As future work, we would like to improve the accuracy of the response emotion/intent predictor in the model, which we found plays a vital role in generating empathetic responses.

5 A Large-Scale Dataset for Empathetic Response Generation

This chapter is based on the work of Anuradha Welivita, Yubo Xie, and Pearl Pu (Welivita et al., 2021). The author of this thesis (Yubo Xie) was mainly responsible for curating the data and expanding the human-labeled data using the semi-supervised learning framework.

5.1 Introduction

Using domain-specific datasets, researchers are more and more inclined to fine-tune pre-trained language models to accomplish specified tasks (Devlin et al., 2019; Liu et al., 2019; Rashkin et al., 2019). The development of conversational agents with empathy, or those that can recognize and respond to human emotions, is one such area. The objective of the empathetic response generation task is to produce responses to previous dialog turns that are syntactically correct, contextually relevant, and—most importantly—emotionally appropriate. Such tasks necessitate the development and availability of sizable dialog datasets, where each utterance is tagged with the appropriate emotions and intents. Despite the fact that many similar datasets have been created in the past (Busso et al., 2008; Poria et al., 2019; Li et al., 2017; Rashkin et al., 2019), their size is constrained due to the expense of manual labor, making them insufficient to train robust conversational agents. Given the high cost of gathering and manually annotating such gold standard data, replacing them with automatically annotated silver standard data is becoming increasingly popular (Filannino and Di Bari, 2015). We demonstrate how such a large-scale, high-quality silver standard dataset may be curated and utilized to fine-tune language models for the generation of empathetic responses.

Social chitchat can disclose a variety of nuanced emotions. Due to the subtle variations present in human emotion, there are numerous categories of emotions that can be distinguished. For instance, despite the fact that *sadness* and *disappointment* are both negative emotions, they are pursued and handled differently in human conversations. Additionally, the listener's response to emotion is not necessarily a direct reflection of the speaker's emotion. Instead, it can be more neutral and convey a specific intent, as shown by the dialog example in Table 5.1.

Table 5.1: An example showing the listener's reactions to emotions do not always mirror the speaker's emotions.

Speaker:	I've been hearing some strange noises around the house at night. (*afraid*)
Listener:	Oh no! That's scary! What do you think it is? (*neutral: acknowledging; questioning*)
Speaker:	I don't know, that's what's making me anxious. (*anxious*)
Listener:	I'm sorry to hear that. (*neutral: sympathizing*)

Welivita and Pu (2020) analyzed the listener responses in EmpatheticDialogues (Rashkin et al., 2019) and discovered eight listener specific empathetic response intents contained in emotional dialogs: *questioning*; *agreeing*; *acknowledging*; *sympathizing*; *encouraging*; *consoling*: *suggesting*; and *wishing*. They have annotated the EmpatheticDialogues dataset with 32 fine-grained emotions, eight empathetic response intents, and the *neutral* category, and discovered frequent emotion-intent exchange patterns in empathetic conversations. They observe that this type of dataset tagged with fine-grained emotions and intents can be used to train neural chatbots to generate emotionally appropriate responses. But for this purpose, a large-scale emotion and intent labeled dataset is even more desirable. Curating such a dataset is technically challenging since (1) annotating such a large-scale dataset require costly human labor, and (2) given the fine-granularity of the emotion and intent labels, the human labeling task is more difficult and error-prone compared with the more coarse-grained *angry-happy-sad* emotion categories. As a result, existing manually labeled emotional dialog datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and DailyDialog (Li et al., 2017) are smaller in scale and contain only a limited set of emotions (emotions derived from basic emotion models such as the Ekman's). Most importantly, existing datasets fail to distinguish between *neutral* and *questioning*, or any of the other eight empathetic response intents. They combine everything into a big label *neutral* or *other* when the utterance is not emotional. But *questioning*, *agreeing*, *acknowledging*, *sympathizing*, *encouraging*, *consoling*, *suggesting*, and *wishing* are important details in constructing empathetic dialogs. These eight response intents, which we call the plus categories, are novel in our work and contribute to the model's learning of important response patterns in the data.

To fill the above gap, we curate a novel large-scale silver dialog dataset, **EDOS** (**E**motional **D**ialogs in **O**pen**S**ubtitles), containing 1M emotional dialogs from movie subtitles, in which each dialog turn is automatically annotated with 32 fine-grained emotions, eight plus categories as well as the *neutral* category. Movie subtitles are extensively used for emotion analysis in text in earlier and recent research (Kayhani et al., 2020; Merdivan et al., 2020; Giannakopoulos et al., 2009). According to the Nature article "How movies mirror our mimicry" (Ball, 2011), screenwriters mine everyday discourse to make dialogs appear authentic, and audiences use language devices in movies to shape their own discourse. Hence, it can be one of the major

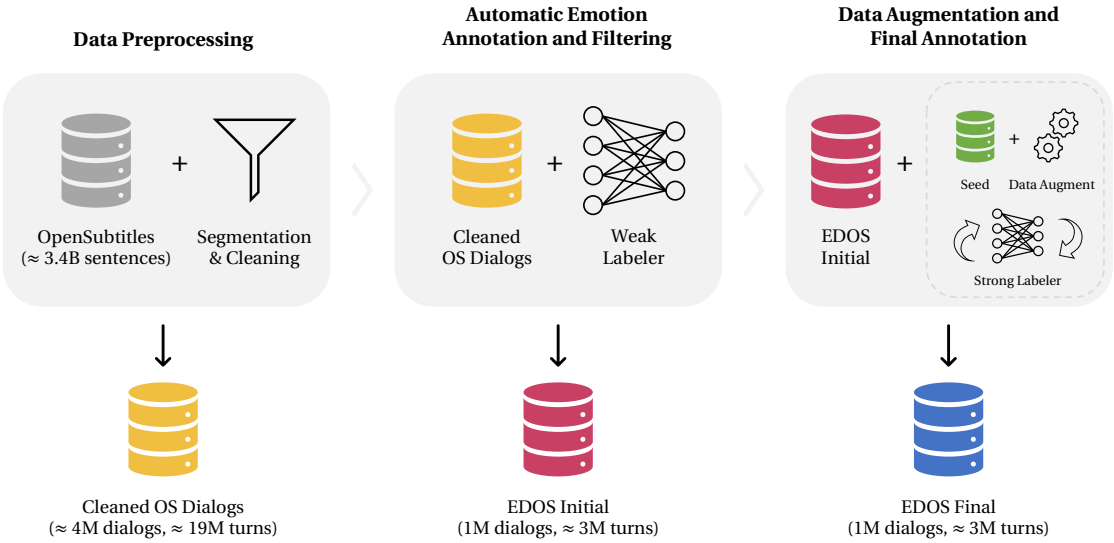


Figure 5.1: Steps for curating the EDOS dataset.

sources to train chatbots and learn emotional variations and corresponding response strategies in dialogs. To reduce the cost of human labeling and the complexity of labeling dialogs with fine-grained emotions and intents, we devised a semi-automated human computation task to collect fine-grained emotion and intent labels for a small set of movie dialogs (9K). We then followed automatic data augmentation techniques to expand the labeled data and trained a dialog emotion classifier to automatically annotate 1M emotional dialogs.

The process of curating the dataset involved several stages. First, we applied automatic turn and dialog segmentation methods, data cleaning and removal of duplicates on movie subtitles in the OpenSubtitles (OS) corpus (Lison et al., 2018) and obtained close to 4M dialogs. Then, we applied a weak labeler (a BERT-based sentence-level classifier) trained on the EmpatheticDialogues dataset (Rashkin et al., 2019), to label utterances in OS dialogs and filtered 1M emotional dialogs (EDOS initial). Thereafter, we applied data augmentation techniques on a small set of human-annotated data and used the manually annotated and extended labels to train a strong labeler that is used to annotate dialogs in EDOS initial and obtained the final 1M EDOS dataset. We evaluated the quality of the resultant dataset by comparing it against the EmpatheticDialogues dataset by means of visual validation methods. Figure 5.1 summarizes the process of creating EDOS. The data curation pipeline we followed substantially reduced the cost of human labor while ensuring quality annotations.

Our contributions in this chapter are three-fold: (1) We curate a large-scale dialog dataset, EDOS, containing 1M emotional dialogs labeled with 32 fine-grained emotions, eight empathetic response intents (the plus categories), and *neutral*. Compared to existing dialog datasets tagged with emotions, EDOS is significantly larger (≈ 40 times larger than EmpatheticDialogues), and contains more fine-grained emotions and empathetic response strategies. (2) We outline the complex pipeline used to derive this dataset. (3) We analyze the quality of the

dataset by comparing with a state-of-the-art gold standard dataset using visual validation methods.

5.2 Related Work

IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), DailyDialog (Li et al., 2017), EmotionLines (Hsu et al., 2018), and EmoContext (Chatterjee et al., 2019) are some existing state-of-the-art dialog datasets with emotion labels. However, these datasets are limited in size and are labeled with only a small set of emotions without any response strategies. Table 5.2 shows a summary of the size and the labels in these datasets. All the datasets compared here are in the English language.

Herzig et al. (2016) detected customer emotions and agent emotional techniques (e.g., *apology* and *empathy*) in customer support dialogs. They curated a dialog dataset from two customer support Twitter accounts and manually annotated the customer turns with one of 9 emotions and the agent turns with one of 4 emotional techniques. But emotions expressed by customers in social media service dialogs are mainly negative (e.g., *anger* and *frustration*), and the customer service agents also respond in a restricted manner, which limits the utility of this dataset, in addition to its small size.

The EmpatheticDialogues dataset (Rashkin et al., 2019) contains 25K open-domain dialogs grounded on 32 emotions. The 32 emotions range from basic emotions derived from biological responses (Ekman, 1992; Plutchik, 1984) to larger sets of subtle emotions derived from contextual situations (Skerry and Saxe, 2015). Welivita and Pu (2020) manually analyzed a subset of the listener turns in EmpatheticDialogues and identified eight listener-specific response intents. They developed a sentence-level weak labeler using which they annotated the entire dataset with 32 emotions, eight empathetic response intents, and the *neutral* category. However, due to the limited size of EmpatheticDialogues, it is difficult to be used for data-intensive applications. To address the above limitations, we curate EDOS containing 1M movie dialogs. We label each dialog turn with 32 emotions, eight empathetic response intents, and *neutral* using our own dialog emotion and intent classifier. Table 5.2 compares EDOS to state-of-the-art emotion annotated dialog datasets.

5.3 Methodology

This section describes the dialog selection process, the design of the human annotation task, the data augmentation techniques used to expand human-labeled dialogs, and the development of a strong labeler to annotate the dataset.

Table 5.2: Comparison of emotion annotated dialog datasets available in the literature against EDOS.

Dataset	**Labels**	**# Dialogs**	**# Utterances**	**Public?**
IEMOCAP (Busso et al., 2008)	*Joy*, *sadness*, *anger*, *frustrated*, *excited*, and *neutral*	151	7,433	Yes
Twitter customer support (Herzig et al., 2016)	Customer emotions: *confusion*, *frustration*, *anger*, *sadness*, *happiness*, *hopefulness*, *disappointment*, *gratitude*, and *politeness*; Agent emotions: *empathy*, *gratitude*, *apology*, and *cheerfulness*	2,413	14,078	No
DailyDialog (Li et al., 2017)	*Joy*, *surprise*, *sadness*, *anger*, *disgust*, *fear*, and *neutral*	13,118	102,977	Yes
EmotionLines (Hsu et al., 2018)	*Joy*, *surprise*, *sadness*, *anger*, *disgust*, *fear*, and *neutral*	2,000	29,245	Yes
MELD (Poria et al., 2019)	*Joy*, *surprise*, *sadness*, *anger*, *disgust*, *fear*, and *neutral*	1,433	13,708	Yes
EmoContext (Chatterjee et al., 2019)	*Joy*, *sadness*, *anger*, and *other*	38,424	115,272	Yes
EmpatheticDialogues (Rashkin et al., 2019; Welivita and Pu, 2020)	32 fine-grained emotions (positive and negative), *neutral*, and 8 empathetic response intents: *questioning*, *agreeing*, *acknowledging*, *sympathizing*, *encouraging*, *consoling*, *suggesting*, and *wishing*	24,850	107,220	Yes
EDOS	32 fine-grained emotions (positive and negative), 8 empathetic response intents, and *neutral*	1,000,000	3,488,300	Yes

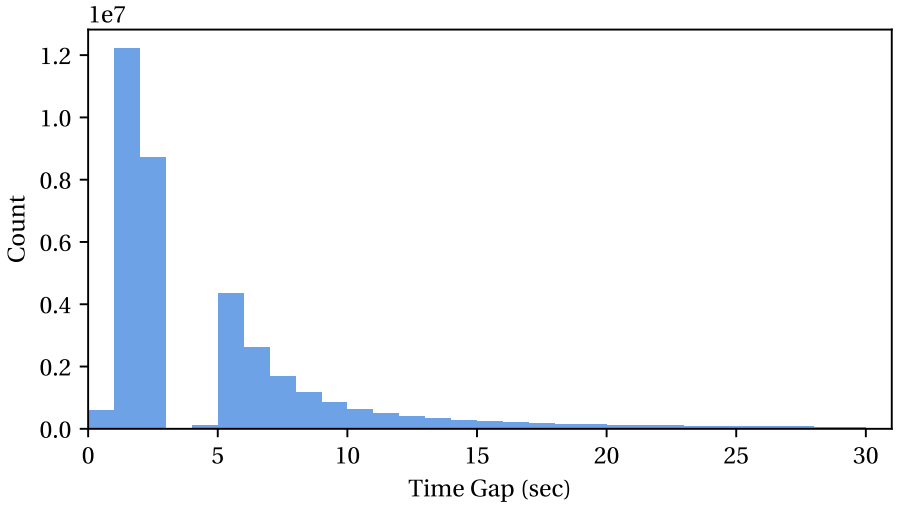


Figure 5.2: Histogram of time intervals (in seconds) between adjacent subtitle blocks in the OpenSubtitles corpus.

5.3.1 Dialog Curation from Movie Subtitles

The OpenSubtitles2018 corpus consists of 3.7M movie and TV subtitles. It comprises 3.4B sentences and 22.2B tokens, and we only use the English part, which has 447K subtitle files, 441M sentences and 3.2B tokens. It is an excellent source to learn emotional variations in dialog and corresponding response mechanisms. But due to the absence of speaker markers, movie subtitles do not contain an explicit dialog turn structure (who speaks what) and specific indicators where one dialog ends and the next dialog begins. To overcome the first issue, we reproduced the work by Lison and Meena (2016) to build an SVM-based classifier that determines if two consecutive sentences are part of the same dialog turn. Our classifier achieved a segmentation accuracy of 76.69%, which is close to the accuracy of 78% that the authors claim. The set of features that gave the best turn segmentation accuracy are:

- Unigram and bigram features of adjacent sentences after lemmatization;
- First and final tokens of adjacent sentences;
- First and final bi-grams of adjacent sentences;
- Whether the two sentences belong to the same subtitle block or not (boolean);
- Genre of the movie (*Drama*, *Crime*, *Musical*, etc.);
- Sentence density of the subtitles file (no. of sentences/subtitle duration);
- Quadratic combinations of the above features with itself and the rest.

After performing turn segmentation on the OpenSubtitles corpus, we divided the turns into separate dialogs based on a simple heuristic. If the difference between the end time of the

previous turn and the start time of the current turn is more than 5 seconds, we take these two turns as belonging to 2 different dialogs. An exception occurs if this timestamp information is missing in at least one of the turns. In this case, we assume that these two turns appear in the same subtitle block and consider them as belonging to the same dialog. This way, we formed 9M dialogs from the OpenSubtitles corpus altogether. The choice of 5 seconds to separate dialogs is based on a histogram of time intervals between adjacent subtitle blocks in the OpenSubtitles corpus, which is depicted in Figure 5.2. As can be observed in the histogram, most of the time gaps fall below 3 seconds, and there is a clear drop between 3–5 seconds.

To further clean the dialogs, we removed character names, the repetitive dialog turns, turns that start with "previously on..." (narration at the beginning of TV episodes), turns with character length less than 2 or greater than 100, turns with an alphabetic proportion less than 60%, and turns with a lot of repetitive tokens. When a dialog turn was removed, all the turns following that turn were also removed from the dialog to maintain consistency. After that, all the dialogs left with only one turn were removed from the corpus. We removed dialogs from movies of the genre "Documentary" since they do not correspond to actual dialogs. We also ran a profanity check on all the utterances using the Python package `profanity-check`.[1] We set a threshold of 0.7, and removed 24,898 dialogs and 84,657 utterances. As a result, we obtained a cleaned OS dialog dataset consisting of 4M dialogs.

To select out dialogs containing emotional statements and empathetic responses from the cleaned OS dialogs dataset, we employed a weak labeler (a BERT-based sentence level classifier) trained on 25K situation descriptions from EmpatheticDialogues (Rashkin et al., 2019) tagged with 32 emotion classes, and 7K listener utterances tagged with eight empathetic response intents and the *neutral* category (Welivita and Pu, 2020). The classifier had a high top-1 classification accuracy of 65.88%. We call it a weak labeler since it predicts emotion or intent only at the sentence level and is trained on a different dataset other than OS. We filtered the top 1M dialogs having the highest label confidence as predicted by this classifier to form the 1M EDOS (initial) dataset.

5.3.2 Human Computation

To train a dialog emotion classifier that can identify both fine-grained emotions and empathetic response intents, we devised an Amazon Mechanical Turk (AMT) experiment to collect an initial set of ground truth labels for OS dialogs. But annotating dialog turns with one of 41 labels is a daunting task. To make the task less exhaustive, we devised a semi-automated approach using our weak labeler. By applying the weak labeler on each turn of the cleaned OS dialog dataset, we selected out the turns having prediction confidence ≥ 0.9, along with their dialog history. Next, we ranked these dialogs according to their readability and selected the highest readable dialogs from each class to be labeled. This is to reduce the time spent by the workers in having to read long and complicated dialogs. The steps followed to compute the

[1] https://pypi.org/project/profanity-check/

Table 5.3: The results of the AMT task for curating EDOS.

Description	Result
Total number of dialogs	10,250
Number of dialogs labeled with majority vote	8,913 (86.96%)
Inter-annotator agreement (Fleiss' κ)	0.46 (moderate)
Percentage of times workers got 3/5 quiz questions correct	77.75%
Number of dialogs in which the workers manually specified the label	425

dialogs' readability are included in Appendix A.6. Workers had to select a label from the top-3 predictions made by the weak labeler. If none of the top-3 predictions matched, they could manually specify the correct class. The main purpose of incorporating a weak labeler here was to make the task less daunting for the crowdsourcing worker. Otherwise, having to choose a label out of 41 labels may lead to even worse results due to the complicated nature of the task. The risk of reduced data reliability is avoided by taking only the labels with the majority vote. The AMT task's user interface design is included in Appendix A.7.

After ranking the dialogs according to readability, we selected the top 250 dialogs in each category for the AMT task. We bundled 15 dialogs in a HIT with 5 quiz questions that served as checkpoints to evaluate the crowdsourcing workers' quality. Situation descriptions from the EmpatheticDialogues dataset for which we already knew the emotion labels were used to formulate the quiz questions. Finally, we obtained dialogs where we had 2 out of 3 worker agreements, which resulted in 8,913 dialogs altogether. Table 5.3 shows the results of the AMT task.

5.3.3 Data Augmentation and Annotation

To scale up the training data obtained from the AMT task, we utilized a distant learning technique using dialog embeddings and self-labeling (Triguero et al., 2015), a semi-supervised learning technique. The first approach we used is Sentence-BERT (SBERT) proposed by Reimers and Gurevych (2019), which uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. Using this approach, we obtained semantically similar dialogs to those annotated by crowdsourcing workers and tagged them with the same class label. Among several models the authors have proposed, we used the *roberta-base-nli-stsb-mean-tokens* model, fine-tuned on the NLI (Bowman et al., 2015) and STS benchmark (STSb) (Cer et al., 2017) datasets, since it has reported a high Spearman's rank correlation of 84.79 ± 0.38 between the cosine-similarity of the sentence embeddings and the gold labels in the STS benchmark test set outperforming the existing state-of-the-art. It is also more efficient than the *roberta-large* model. Before proceeding, we split the crowd-annotated dialogs into 60% training, 20% validation, and 20% testing (balanced across all class labels). Then, we followed the following steps to extend the dialogs using Sentence-BERT:

1. Using the Sentence-BERT model, first, we computed dialog turn embeddings (each with a vector representation with dimension 768) for all the turns (≈19M) in the cleaned OS dataset.

2. Then, we calculated dialog embeddings for human-annotated and unlabeled dialogs from the cleaned OS dialogs dataset. For this, we applied a decaying weight starting from the last turn and took the weighted average of the turn embeddings of each dialog. We used half decaying; i.e., if we have a dialog with turn embeddings $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$, the final dialog embedding would be $(4/7)\,\boldsymbol{v}_3 + (2/7)\,\boldsymbol{v}_2 + (1/7)\,\boldsymbol{v}_1$.

3. Next, we calculated the cosine similarity between annotated and unlabeled dialog embeddings and ranked the results.

4. Finally, we applied a similarity threshold and obtained all the unlabeled dialogs with a cosine similarity that exceeds this threshold and tagged them with the same crowd-annotated class label. Here, we used a threshold of 0.92 after manually inspecting a random subset of the results obtained for a range of thresholds. Examples from this stage are shown in Appendix A.8.

We extended the original crowd-annotated dialog dataset by 3,196 more dialogs with distantly annotated class labels using the above method.

Using the crowd-annotated and extended labels, we trained an initial classifier that we used to annotate the rest of the dialogs and add more labels to our dataset that had annotation confidence over 0.9. This method is termed self-labeling (Triguero et al., 2015), a semi-supervised learning technique that can be used to grow labeled data. With this, we were able to extend the labeled data by 4,100 more dialogs. Next, we again applied Sentence-BERT over the self-labeled data and extended them by 2,118 more dialogs. Finally, we were able to have around 14K labeled dialogs altogether. We used this data to train a final dialog emotion classifier to annotate the rest of the unlabeled data. This resulted in a classifier with precision 64.11%, recall 64.59%, macro F1-score 63.86%, and accuracy 65.00%, which is comparable with the state-of-the-art dialog emotion classifiers (as denoted in Table 5.4). The performance of the classifier over the iterations is shown in Table 5.5. We now elaborate its architecture design in next section.

Dialog Emotion Classifier

Our dialog emotion classifier, as shown in Figure 5.3, consists of a representation network that adopts the BERT architecture, an attention layer that aggregates all hidden states at each time step, a hidden layer, and a softmax layer. We used the BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters as the representation network. It was initialized with weights from RoBERTa (Liu et al., 2019). We fed in a dialog turn along with the preceding context in the reverse order as input to the representation network. To give more

Table 5.4: Comparison of the performance of the dialog emotion classifier used for annotation with performance of the state-of-the-art dialog emotion classifiers. Here we use macro averaging for the F1 score.

Classifier	**Dataset**	**# Labels**	**F1**	**Accuracy**
EmotionX-AR (Khosla, 2018)	EmotionLines (Hsu et al., 2018)	4 emotions	-	62.50 (Friends) 62.48 (EmotionPush)
CMN (Hazarika et al., 2018b)	IEMOCAP (Busso et al., 2008)	6 emotions	56.13	56.56
ICON (Hazarika et al., 2018a)	IEMOCAP (Busso et al., 2008)	6 emotions	57.90	58.30
IAAN (Yeh et al., 2019)	IEMOCAP (Busso et al., 2008)	6 emotions	-	64.70
DialogueRNN (Majumder et al., 2019)	IEMOCAP (Busso et al., 2008)	6 emotions	62.75	63.40
DialogueGCN (Ghosal et al., 2019)	IEMOCAP (Busso et al., 2008); MELD (Poria et al., 2019)	IEMOCAP: 6 emotions; MELD: 7 emotions	64.18 (IEMOCAP) 58.10 (MELD)	65.25 (IEMOCAP)
Ours	OS dialogs	32 emotions, 8 empathetic response intents, and *neutral*	63.86	65.00

Table 5.5: Precision, recall, F1, and accuracy scores of the dialog emotion classifier over the semi-supervised learning iterations. All scores are reported on the human-annotated test set. Here we use macro averaging.

Iteration	Training Data	P	R	F1	Acc
Base	Dialogs from AMT (5K)	0.6333	0.6394	0.6328	**0.6517**
First	5K + Similar dialogs from SBERT (3K)	0.6355	0.6354	0.6292	0.6455
Second	5K + 3K + Self-labeled dialogs (4K)	0.6396	0.6427	0.6361	0.6488
Second (extended)	5K + 3K + 4K + Similar self-labeled dialogs from SBERT (2K)	**0.6411**	**0.6459**	**0.6386**	0.6500

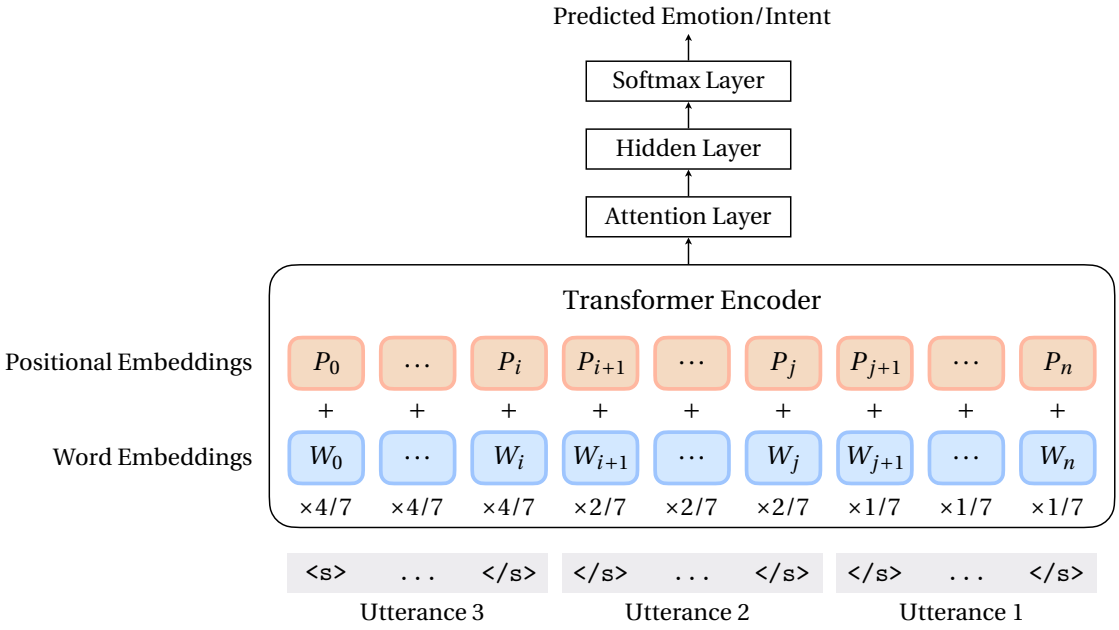


Figure 5.3: Architecture of the emotion/intent classifier used to label the EDOS dataset.

importance to the dialog turn for which prediction has to be made and the turns that immediately precede it, we multiplied the word embeddings belonging to each turn by a decreasing weight factor. The input representation is constructed by summing the corresponding word embeddings multiplied by the weighting factor and its position embeddings. More details including the hyper-parameters used are included in Appendix A.8.

5.4 Quality Analysis

The statistics of the EDOS dataset are given in Table 5.6. More detailed statistics including the number of dialogs per emotion are included in Appendix A.5. Table 5.7 shows some example dialogs taken from the EDOS dataset along with annotations and confidence scores. By observing the examples, it can be noticed that even for less confident predictions, the label

Table 5.6: Statistics of the EDOS dataset.

Total number of dialogs	1,000,000
Total number of turns	2,829,426
Total number of tokens	39,469,825
Average number of turns per dialog	2.83
Average number of tokens per turn	13.95
Average number of tokens per dialog	39.47

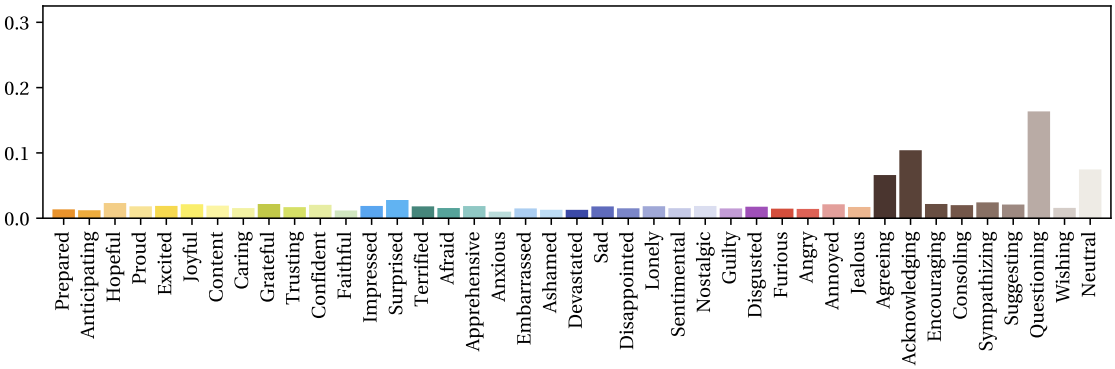
Table 5.7: Example dialogs from the EDOS dataset along with annotations and confidence scores.

Turn 1	**(Excited, 0.98)** The concert will start soon.
Turn 2	**(Questioning, 0.01)** Are you excited?
Turn 3	**(Proud, 0.99)** I am. Because one of my friends made his efforts to make the concert happen. He wanted to fulfill a promise he made to his first love.
Turn 4	**(Sentimental, 0.99)** I like their story very much. I want to dedicate this concert to everyone who has truly loved someone.
Turn 1	**(Apprehensive, 0.89)** Staying here might not be safe.
Turn 2	**(Questioning, 0.41)** Take the earliest flight tomorrow?
Turn 3	**(Caring, 0.94)** Take Josie to mother. My home is where you are.
Turn 4	**(Faithful, 0.86)** We're not leaving.

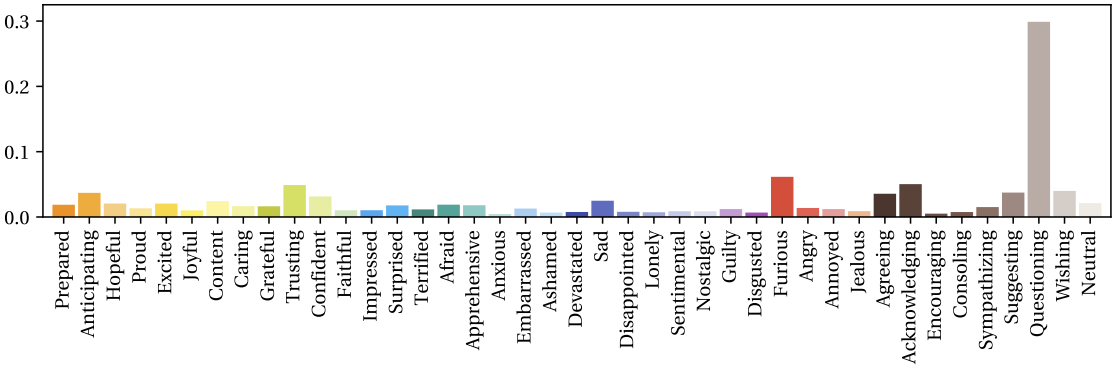
quite accurately describes the emotion or intent of the corresponding dialog turn.

We also conducted a qualitative comparison of the annotations in the EDOS dataset with EmpatheticDialogues (Rashkin et al., 2019; Welivita and Pu, 2020), a state-of-the-art gold standard dataset for empathetic conversations. Figure 5.4 compares the distributions of emotions and intents in the two datasets. It is observed that in both datasets, intent categories take prominence over individual emotion classes. This is in par with observations of Welivita and Pu (2020), where they notice that one or more intents from the taxonomy of empathetic intents are mostly utilized when responding to emotions in dialog, rather than similar or opposite emotions. Especially, the intent *questioning* takes the highest percentage among the annotations in EmpatheticDialogues and EDOS. We also computed the KL-divergence (≥ 0) of the emotion and intent distribution of EDOS with respect to that of EmpatheticDialogues, which measures how one probability distribution is different from a second, reference probability distribution (Kullback and Leibler, 1951). It resulted in a KL-divergence value of 0.2447, which indicates a considerable similarity between the two distributions (the lower the KL divergence, the more similar the distributions are).

Figure 5.5 and 5.6 show the emotion-intent flow patterns in EmpatheticDialogues and EDOS. In the visualization corresponding to EmpatheticDialogues, the first and third dialog turns correspond to the speaker and the second and fourth dialog turns correspond to the listener. However, in EDOS, we cannot distinguish the dialog turns as speaker and listener turns



(a) EmpatheticDialogues dataset.



(b) EDOS dataset.

Figure 5.4: Comparison of distribution of emotions and intents in the EmpatheticDialogues and EDOS datasets.

due to the absence of speaker annotations. Though this is the case, we could still observe some conversational dynamics present in EmpatheticDialogues are preserved in EDOS. For example, in both datasets, the speaker mostly starts the conversation with some emotional statement and in the subsequent turn, the response tends to be of the intent *questioning*. In both datasets, intents *agreeing* and *acknowledging* follow emotions seen in the first turn irrespective of whether they are positive or negative. As the dialogs proceed, it could be seen in both datasets the emotions deescalate as more empathetic response intents emerge.

5.5 Chapter Summary

In this chapter, we curated a large-scale dialog dataset, EDOS, comprising of 1M emotional dialogs from movie subtitles. This dataset is significantly larger in size and contains more fine-grained emotion categories and empathetic response intents than the existing emotional dialog datasets. To facilitate annotation, we utilized data augmentation techniques to extend a small set of manually annotated data and trained a dialog emotion classifier having comparable accuracy to the state-of-the-art. The data augmentation and automatic annotation procedure we employed significantly reduced the manual annotation cost and time.

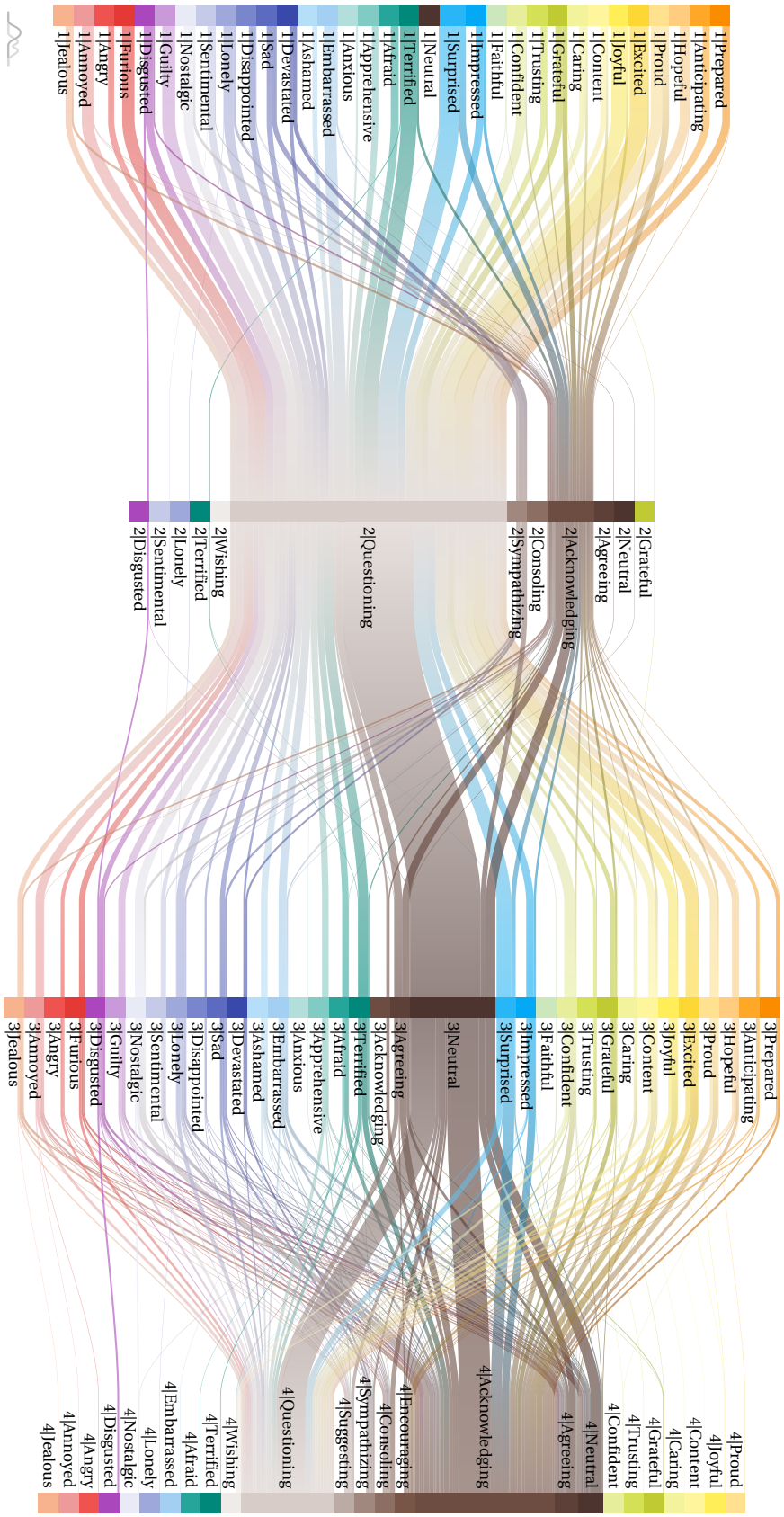


Figure 5.5: The emotion-intent flow pattern in the EmpatheticDialogues dataset. For simplicity, only the first four dialog turns are visualized.

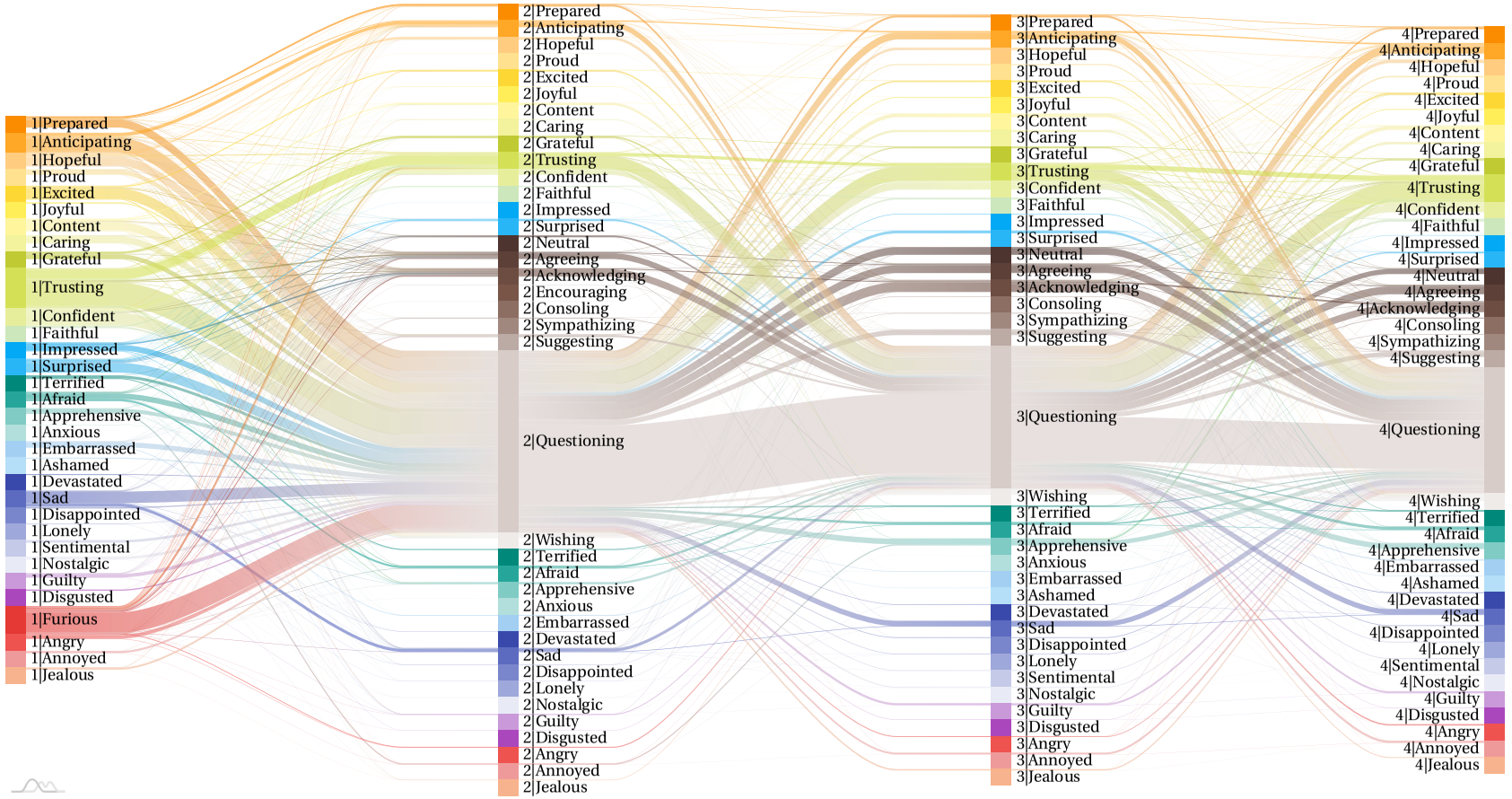


Figure 5.6: The emotion-intent flow pattern in the EDOS dataset. For simplicity, only the first four dialog turns are visualized.

Obtaining a large dataset is important only if the quality can be assured. The qualitative comparison conducted between EDOS and the state-of-the-art EmpatheticDialogues dataset by means of visual validation was one way to confirm that. The results of the comparison confirmed that most of the conversational dynamics present in EmpatheticDialogues were observed in EDOS. We also proposed some experimental baselines by training a Transformer model for empathetic response generation on the OS, EDOS, and EmpatheticDialogues datasets and tested them in held-out and zero-shot settings (see Chapter 4). The results showed that the model fine-tuned on EDOS scored the best in terms of diversity metrics. This dataset can be readily utilized to develop empathetic conversational agents and for fine-grained emotion analysis in dialogs. The pipeline we present can be used to create similar large-scale datasets in similar or even different domains.

As future work, we plan to utilize this dataset to further conduct experiments on empathetic response generation. Since it is annotated with emotions and intents, we will use it for experiments involving controllable and interpretable response generation. Particularly, the plus categories present in the dataset can be utilized to condition the chatbot's response generation process, making it possible to control and interpret the generated responses. The dataset can also be used to train state-of-the-art dialog emotion classifiers.

6 Capturing Social Intelligence in Casual Conversations

6.1 Introduction

Social intelligence is our ability to optimally understand our social environment and react wisely (Ganaie and Mudasir, 2015). As one type of commonsense knowledge, social intelligence encompasses various abilities such as emotion reading, empathy, and knowledge of social rules. According to Daniel (2006), social intelligence consists of two ingredients: *social awareness*, the ability of understanding others' feelings and grasping the complexity of social situations, and *social facility*, the ability of making smooth and effective interactions based on social awareness. For example, if someone says "my job interview didn't go well," one could immediately sense that the person is upset and might be seeking consolation. Based on such social awareness, a good way to respond could be "don't give up and I'm sure you'll do better next time." Such examples of applying social intelligence exist everywhere in our daily life and can easily be found in human-to-human conversations.

Having desirable social intelligence does not always seem easy and straightforward even for human beings, not to mention AI systems. Previous work has attempted at creating knowledge graphs to explicitly represent different commonsense knowledge so that AI systems could utilize them as external resources to perform various types of commonsense reasoning. This ranges from knowledge about common concepts (Speer et al., 2017; Tandon et al., 2017) to inferences over common events (Sap et al., 2019; Zhang et al., 2020a). However, to the best of our knowledge, there is still no work on building a knowledge graph which captures the social intelligence that people display when conducting social conversations with each other in day-to-day social environments. By inspecting the knowledge graph, one could learn how to appropriately respond to another person with desired emotions and intents. The knowledge graph can also serve as an external resource for the development of both open-domain and task-oriented dialog systems.

In this chapter, we present **AFEC** (**AF**fable and **E**ffective **C**onverser), an automatically curated knowledge graph that captures social intelligence in casual conversations between real people on various topics commonly found in daily life. Figure 6.1 gives a snippet of our knowledge

Figure 6.1: A snippet of AFEC, our knowledge graph of social intelligence in casual conversations. Red nodes represent speaker utterances, which start a conversation, and blue nodes represent the corresponding listener utterances, labeled with the desired emotions necessary for continuing the conversation.

graph, showing different ways people respond to some experience shared by another person. Due to the limitations of existing open-domain dialog datasets (large-scale datasets are usually noisy, and manually labeled datasets are small in size), we decided to manually crawl conversational data from the `r/CasualConversation` subreddit,[1] which covers a wide range of topics in people's day-to-day social communication. After preprocessing, we grouped and merged the utterances into nodes of the knowledge graph. Each node is a short utterance, with the red ones representing the speaker (usually sharing some past experience) and the blue ones representing the listener (responding to the speaker). Following the taxonomy of empathetic response intents proposed by Welivita and Pu (2020), we labeled each node with one of the 41 categories of emotions/intents, by training a Transformer (Vaswani et al., 2017) classifier on manually labeled dialog data. Unlike other commonsense knowledge graphs that deal with concepts or events, ours focuses on verbal skills to demonstrate empathy, which are ubiquitous in people's daily social environments, and was curated automatically from data, without the efforts of manual selecting and labeling.

[1] https://www.reddit.com/r/CasualConversation/

Our contributions are as follows: (1) We designed a completely automatic curation pipeline to create a knowledge graph of social intelligence covering people's daily communication, which can be used as an external resource to improve the performance of dialog systems; (2) As a by-product of our knowledge graph, we obtained a large-scale casual conversation dataset that has a good trade-off between size and quality; (3) We designed a simple retrieval-based chatbot using the knowledge graph, without the efforts of training a neural network, and the comparison with existing empathetic dialog models showed that our retrieval model is capable of generating much more diverse responses, yet still producing quality responses.

6.2 Related Work

6.2.1 Commonsense Knowledge

It is believed that Bar-Hillel (1960) was the first to mention the importance of incorporating commonsense knowledge into natural language processing systems, in the context of machine translation. Broadly speaking, there are two types of commonsense knowledge. One type of commonsense knowledge that humans generally acquire is "naive physics," which involves inference of how physical objects interact with each other. For example, if one is told that a glass of water falls onto the floor, he/she will most likely infer that the glass shatters and the floor becomes wet. Another type of commonsense knowledge that humans have is "intuitive psychology." This type of knowledge enables us to infer people's behaviors, intents, or emotions. For example, one could easily tell that a person who has lost his/her job probably feels upset. There is abundant recent work that explicitly represents commonsense knowledge into a structured knowledge graph. Cyc (Lenat and Guha, 1989) is a project aiming at integrating ontologies and commonsense knowledge from all different domains into one knowledge base, and based on that, achieving the ability of knowledge inference like human beings. Concepts in Cyc are called "constants" and categorized into *individuals*, *collections*, *truth functions*, and *functions*. OpenCyc 4.0 is the most recent public version and contains 239,000 concepts and 2,039,000 facts. ConceptNet (Speer et al., 2017) is a directed graph whose nodes are concepts, and the edges represent assertions of commonsense about the concepts, e.g., *IsA*, *IsUsedFor*, *MotivatedByGoal*, etc. The nodes are natural language phrases, e.g., noun phrases, verb phrases, or clauses. The latest version is ConceptNet 5.5, which contains over 8 million nodes and over 21 million links. SenticNet (Cambria et al., 2020) incorporates a set of semantics, sentics, and polarity associated with 200,000 natural language concepts. Specifically, semantics define the denotative information associated with natural language phrases, sentics define the emotion categorization values (expressed in terms of four affective dimensions) associated with these concepts, and polarity is floating number between -1 and $+1$. WebChild (Tandon et al., 2017) is a large-scale commonsense knowledge graph that was automatically extracted and disambiguated from Web contents, using semi-supervised label propagation over graphs of noisy candidate assertions. It contains triples that connect nouns with adjectives via fine-grained relations such as *hasShape*, *hasTaste*, *evokesEmotion*, etc. The newest version WebChild 2.0 was released in 2017 and contains over 2 million concepts

Figure 6.2: The overall workflow for curating AFEC.

with 18 million assertions. ATOMIC (Sap et al., 2019) is a commonsense knowledge graph consisting of 877K textual descriptions of inferential knowledge obtained from crowdsourcing. It focuses on *if-then* relations between events and possible inferences over the events. The base events were extracted from a variety of corpora including stories and books. The ATOMIC knowledge graph contains a total number of 309,515 nodes and 877,108 triples. ASER (Zhang et al., 2020a) is a large-scale eventuality knowledge graph automatically extracted from more than 11-billion-token unstructured textual data. It contains 15 relation types belonging to five categories, 194 million unique eventualities, and 64 million edges between them. The eventualities were extracted from a wide range of corpora from different sources, according to a selected set eventuality patterns. The eventuality relations were also automatically extracted using a selected set of seed connectives.

6.2.2 Emotional Dialog Data

There are not many emotional dialog datasets that are publicly available, and most of them are limited in size. Li et al. (2017) created the DailyDialog dataset from English learning websites, consisting of 13K multi-turn dialogs manually labeled with 7 emotions. The EmotionLines dataset (Hsu et al., 2018) contains 2,000 dialogs collected from Friends TV scripts and EmotionPush chat logs, labeled with 7 emotions. Chatterjee et al. (2019) proposed the EmoContext dataset collected from users' interaction with a conversational agent, which contains 38K dialogs labeled with 4 emotions. Rashkin et al. (2019) curated the EmpatheticDialogues dataset containing 25K dialogs collected from a crowdsourcing platform by letting workers communicate with each other based on 32 emotion categories. In Chapter 5, we adopted a semi-supervised approach to label the OpenSubtitles dialog data, and obtained a large-scale dialog dataset EDOS that contains 1M empathetic dialogs. However, the dialogs curated from movie subtitles are often noisy and don't fully overlap with daily social topics. Therefore, it is desirable to find conversations from sources that fit more with daily social environments.

6.3 Data Curation

Since our goal is to build a knowledge graph of social intelligence from people's everyday social interactions, it is necessary to have a large-scale daily dialog dataset. Existing dialog datasets (that are publicly available) mainly falls into two categories: task-oriented dialogs

and open-domain dialogs. Task-oriented dialogs are extracted from some specific domain, for example the Ubuntu Dialogue Corpus (Lowe et al., 2015), which includes technology related conversations mostly. These dialogs do not fit into the day-to-day social settings, thus not applicable to our case. Open-domain dialogs, on the other hand, covers a much broader range of topics in daily life, thus more suitable for our purpose. However, existing open-domain dialog datasets have either low quality or limited size. For example, dialog datasets created from movie subtitles, for example OpenSubtitles (Lison et al., 2018), often have a large scale and a diverse range of topics, but the separation between two conversation scenes is hard to detect precisely, and the text quality cannot always be guaranteed due to transcription errors. On the other hand, open-domain dialog datasets of high quality, for example EmpatheticDialogues (Rashkin et al., 2019), are often limited in size, because they are created manually or through crowdsourcing.

Due to the limitations of existing dialog datasets, we would like to curate a large-scale dialog dataset containing high quality casual conversations in daily life, by crawling from online resources. The overall workflow is illustrated in Figure 6.2. Next we are going to describe each step in detail.

Crawling from Reddit Reddit[2] is a discussion website where users can post what they want to share in subreddits with different themes. Among all subreddits, `r/CasualConversation` is a subreddit for users to talk about common topics, like "Today I finish to pay to my loan" and "I'm starting a new job today." We used the Pushshift Reddit API[3] to crawl submission and comment posts on `r/CasualConversation`. Limiting the publishing time from January 1, 2016 to December 31, 2021, we obtained a total number of 387,594 submissions and 4,152,652 comments in English. Among all the comments, there are 1,908,867 comments that directly reply to submission posts. To form the conversations, the submissions are considered as the speaker utterances and we only selected the comments that directly reply to submissions as the listener utterances.

Preprocessing To preprocess all the crawled text, we followed the following rules:

1. We replace the HTML escape characters with their normal ones (e.g., `>` becomes >);
2. We remove any content in brackets;
3. We remove redundant spaces (including breaklines) in the input text;
4. We discard the input text if it is empty;
5. We discard the input text if it is "[deleted]" or "[removed]";

[2] https://www.reddit.com/
[3] https://pushshift.io/

6. We discard the input text if it contains URL;
7. We discard the input text if it contains “r/<subreddit>”, “u/<username>” or “reddit”;
8. We discard the input text if the percentage of alphabetical letters is less than 70%;
9. We discard the input text if the number of tokens is less than 2.

Summarization After preprocessing, there still exist some long utterances (for example some submissions have one or more blocks of text) containing multiple sentences. To keep the crawled conversations succinct, we applied the SMMRY algorithm[4] to summarize these utterances into one sentence.

Dependency Parsing We observed that a submission usually consists of two parts: a title and possibly a block of description text further explaining the situation, and in most cases the title itself is enough to summarize what the speaker is going to express. However, we need to deal with the cases where the title is just a phrase and does not fully describe the whole submission. Therefore, we would like to check if a speaker utterance indicates some specific actions, propositions or statements. Specifically, we used the SpaCy[5] package to generate a dependency parsing tree, and if the root of the tree is a verb, we kept the sentence as the speaker utterance. Note that we first applied this procedure on the title (i.e., first try summarizing the title and then apply dependency parsing), and if the final result was discarded, we then applied the same procedure on the description text.

Clustering After looking into the the utterances filtered by previous steps, we observed that some speaker utterances are similar in meaning (the same case for listener utterances). Therefore, these similar utterances can be clustered into one node. We first encoded each utterance into a vector with dimension 768 using Sentence-BERT (Reimers and Gurevych, 2019), and then calculated the cosine similarity score between vector representations of two utterances. If the similarity value between two speaker utterances is more than 0.85 (and 0.80 for listener utterances), they are grouped together. We used a fast community detection algorithm[6] to generate all the speaker nodes and listener nodes. However, due to the number of listener utterances being too large, the algorithm does not fit into the memory. Thus, we evenly split the listener utterances into two parts and applied the algorithm to find the respective clusters, and then ran another round of community detection on the obtained clusters. To connect these nodes into one graph, an edge is added between a speaker node and a listener node, if at least one of the utterances in the speaker node matches one of the utterances in the listener node. In this step, we only considered utterances with length not greater than 40.

[4] https://smmry.com/about
[5] https://spacy.io/
[6] https://www.sbert.net/examples/applications/clustering/README.html#fast-clustering

Table 6.1: Some statistics of AFEC.

Total number of speaker nodes	131,038
Total number of listener nodes	637,628
Total number of edges	770,192
Average out-degree of speaker nodes	5.88
Average in-degree of listener nodes	1.21

Table 6.2: Taxonomy of emotions and intents used in AFEC.

Emotion	Prepared, anticipating, hopeful, proud, excited, joyful, content, caring, grateful, trusting, confident, faithful, impressed, surprised, terrified, afraid, apprehensive, anxious, embarrassed, ashamed, devastated, sad, disappointed, lonely, sentimental, nostalgic, guilty, disgusted, furious, angry, annoyed, jealous
Intent	Agreeing, acknowledging, encouraging, consoling, sympathizing, suggesting, questioning, wishing, neutral

Finally, we combined all the selected dialogs crawled from Reddit and the dialogs from EmpatheticDialogues dataset. To filter out potentially offensive contents, we ran a profanity check using the Python package `profanity-check`.[7] In the end, there are 149,332 speaker utterances and 803,320 listener utterances in total. After clustering, we have **131,038 speaker nodes** and **637,628 listener nodes**. Some statistics of the AFEC knowledge graph are listed in Table 6.1.

6.4 Node Labeling

When conducting social conversations, people have different patterns of responding to the other interlocutor, which are mostly driven by their emotional states and emotional intelligence. For example, empathy is an important aspect of social conversations, and depending on the other interlocutor's utterance, different people would have different ways to respond to certain emotional events (either joyful or sad), e.g., encouraging, consoling, or questioning. To this end, it is necessary to study the emotions and intents embedded in the utterances in our knowledge graph. To do this, we adopted the taxonomy of empathetic response intents proposed by Welivita and Pu (2020), which extends the 32 emotion categories of EmpatheticDialogues (Rashkin et al., 2019) with 8 fine-grained empathetic response intents, plus the *neutral* category. We then labeled each node in the graph with one of the 41 emotions/intents (listed in Table 6.2).

[7] https://pypi.org/project/profanity-check/

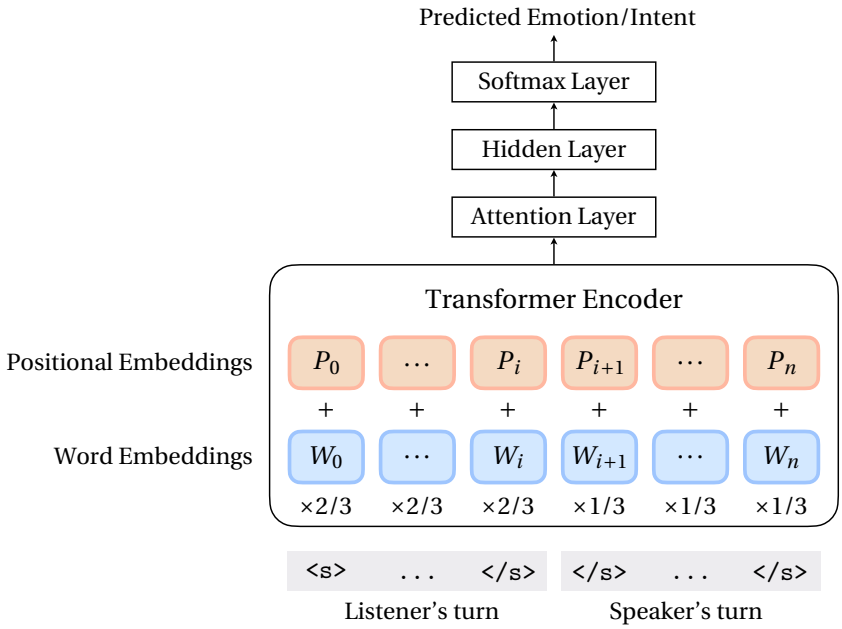


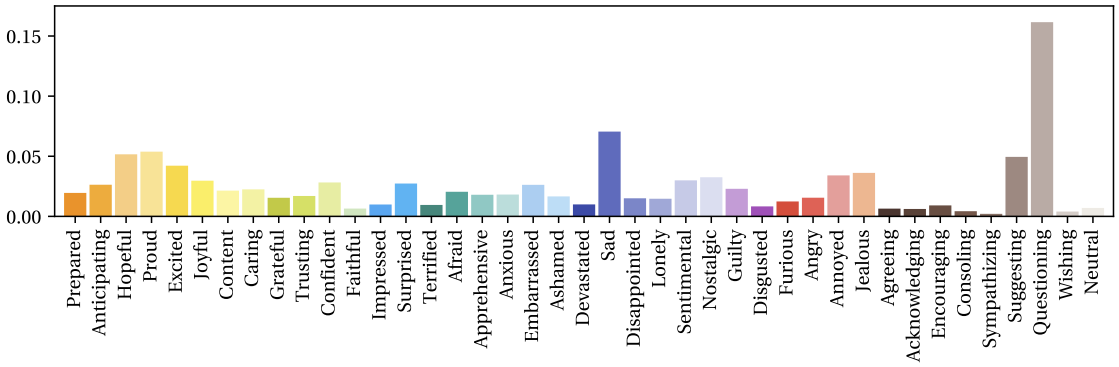
Figure 6.3: Architecture of the emotion/intent classifier used to label the nodes in AFEC.

Specifically, we followed the work of Chapter 5 and trained the same emotion/intent classifier on the dialog data labeled with crowdsourcing workers and then extended with distant learning (14K in total). Figure 6.3 depicts the architecture of the classifier. We adopted a Transformer encoder architecture with 12 layers, 768 hidden units, 12 multi-heads, and initialized the model with weights from the pre-trained language model RoBERTa (Liu et al., 2019). When labeling the speaker's turn (red nodes in the knowledge graph), we just feed the speaker's utterance as input into the model. When labeling the listener's turn (blue nodes in the knowledge graph), we append the corresponding speaker's utterance after the listener's utterance and apply a decaying weight factor so that the model pays more attention to the listener's utterance (as shown in Figure 6.3).

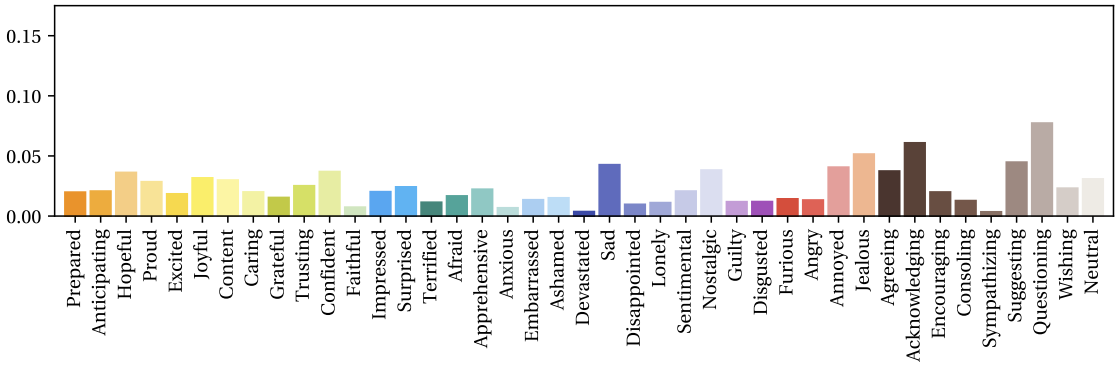
After obtaining the labels for each node, we visualized the distribution of the emotion/intent of the speaker nodes (Figure 6.4a) and that of the listener nodes (Figure 6.4b), respectively. We make the bars of the emotion categories more colorful than the intent categories. It can be observed that the listener nodes are generally less emotional than the speaker nodes, and for the listener nodes, the empathetic response intents, such as *agreeing*, *acknowledging*, *encouraging*, *consoling*, *sympathizing*, and *wishing*, are much more prominent than the speaker nodes. This is also consistent with the findings of Welivita and Pu (2020).

6.5 Experiments

Our curated casual conversational graph can serve as a good data source for the development of open-domain dialog models. In this section, we show how we can build a simple retrieval-



(a) Distribution of emotion/intent in the speaker nodes.



(b) Distribution of emotion/intent in the listener nodes.

Figure 6.4: Distributions of emotion/intent in AFEC.

based dialog model using the conversation graph, which is capable of generating very diverse and human-like responses, yet still achieves better performance than some of the existing generative dialog models, using both automatic and human evaluation.

6.5.1 Data Split

In order to evaluate the retrieval-based chatbot as well as the baselines, we split out roughly 10% of all the utterances in the knowledge graph, reserved as the test set for our following experiments. Specifically, we first selected out all the speaker utterances that come from the EmpatheticDialogues test set, and then, in the remaining speaker utterances, we randomly chose 10% and combined them with the previously selected EmpatheticDialogues utterances, to form the final test set. As a result, the test set consists of 15,212 speaker utterances and 82,284 listener utterances.

6.5.2 A Retrieval-Based Dialog Model

Based on the casual conversation knowledge graph, we are able to develop a simple retrieval-based dialog model, AFEC-Talk, by comparing the input utterance with the speaker nodes

Table 6.3: Groups of similar emotions and intents.

#	Emotion/Intent
1	prepared, confident, proud
2	content, hopeful, anticipating
3	joyful, excited
4	caring
5	faithful, trusting, grateful
6	jealous, annoyed, angry, furious
7	terrified, afraid, anxious, apprehensive
8	disgusted
9	ashamed, guilty, embarrassed
10	devastated, sad, disappointed, nostalgic, lonely
11	surprised
12	impressed
13	sentimental
14	neutral
15	agreeing, acknowledging
16	encouraging
17	consoling, sympathizing
18	suggesting
19	questioning
20	wishing

in the graph and then return the replies (listener nodes) that correspond to the most similar speaker node. In particular, for an input utterance x, we first encode it into a vector with dimension 768 using Sentence-BERT, which is denoted by $\boldsymbol{v}_x$. We then calculate the cosine similarity score between $\boldsymbol{v}_x$ and $\boldsymbol{v}_s$, where $\boldsymbol{v}_s$ is the vector representation (also encoded using Sentence-BERT) of a speaker node s in the graph. By sorting all the speaker nodes by their cosine similarity scores with x, we pick the most similar one, denoted by $\hat{s}$. Given $\hat{s}$, we can find all its corresponding listener nodes $L_{\hat{s}}$, among which we are going to select one utterance as the dialog model's response. We have the following strategies for selecting the reply:

1. **Select randomly**. We just randomly select a listener utterance from $L_{\hat{s}}$.
2. **Select the reply with the highest degree**. We select the listener utterance that has the most in-going edges, meaning that this utterance is more frequently used as a reply than other utterances in the graph.
3. **Select the reply that follows the emotion of the input utterance**. We simply select the reply that has the same or similar emotion/intent of the input utterance. If such reply can not be found, we just randomly pick one from $L_{\hat{s}}$. We consider emotions/intents to be similar if they belong to the same group, as defined in Table 6.3.
4. **Select the reply with empathetic response intents**. We select the reply that is labeled

with one of the 8 empathetic response intents (excluding *neutral*).

For all the strategies, if multiple candidate replies exist, we would randomly select one.

6.5.3 Baselines

Empathy is an important aspect of people's daily social communication, and is considered as an essential ability of open-domain chatbots. To this end, we compare our retrieval-based dialog model with some existing empathetic open-domain dialog models. Due to limited retrieval-based chatbots using knowledge graphs to generate empathetic responses, we decided to compare AFEC-Talk with existing end-to-end generative systems. Our baselines are:

- **MoEL** (Lin et al., 2019). MoEL is an end-to-end empathetic dialog model that uses multiple decoders (listeners) to react to each context emotion accordingly. According to the emotion classification distribution, a meta-listener then combines the output states of each listener, to generate the final empathetic response.

- **MIME** (Majumder et al., 2020). MIME generates empathetic responses by exploiting the assumption that an empathetic conversational agent would often mimic the user's emotion to a certain degree, depending on whether it is positive or negative. The model also introduces some stochasticity into the emotion mixture to generate more varied responses.

- **CEM** (Sabour et al., 2022). When generating the empathetic response, CEM first queries COMET (Bosselut et al., 2019), a GPT-2 (Radford et al., 2019) based model fine-tuned on ATOMIC, to get the commonsense inferences of the input context, and then uses a knowledge selector to fuse the obtained information.

- **BlenderBot** (Roller et al., 2021). BlenderBot was pre-trained on Reddit data and then fine-tuned on the BlendedSkillTalk dataset (Smith et al., 2020), which combines conversation skills from three individual dialog datasets that focus on engaging personality, empathy, and knowledge, respectively. We use the 90M model that is publicly available.

- **MEED2** (Chapter 4). MEED2 incorporates extra fine-grained empathetic intents into the emotion distribution, and uses a separate Transformer encoder to determine the emotion/intent of the response to be generated. The model was pre-trained on Open-Subtitles dialogs and fine-tuned on smaller dialog datasets. We use the version of MEED2 that was fine-tuned on the EmpatheticDialogues dataset.

Table 6.4: Automatic evaluation results. Dist-n denotes Distinct-n score. AFEC-Talk$_*$ denotes our retrieval-based dialog model with different reply selecting strategies: *rand* means selecting randomly; *hd* means selecting the reply with the highest degree; *follow* means following the input emotion/intent; *intent* means selecting the reply with one of the 8 empathetic response intents.

Model	BLEU-2	BLEU-4	ROUGE-2	METEOR	Dist-1	Dist-2	Dist-3
MoEL	0.0232	0.0016	0.0040	0.1098	0.0021	0.0092	0.0191
MIME	0.0215	0.0011	0.0034	0.1298	0.0011	0.0042	0.0074
CEM	0.0184	0.0013	0.0032	0.0862	0.0021	0.0080	0.0151
BlenderBot	0.0383	0.0022	0.0054	**0.1615**	0.0151	0.0873	0.1778
MEED2	0.0300	0.0025	0.0047	0.0972	0.0141	0.0681	0.1314
AFEC-Talk$_{rand}$	0.0475	**0.0223**	0.0110	0.1294	0.0702	**0.4527**	**0.7915**
AFEC-Talk$_{hd}$	**0.0528**	0.0181	**0.0124**	0.1274	0.0703	0.4286	0.7363
AFEC-Talk$_{follow}$	0.0464	0.0223	0.0110	0.1275	0.0696	0.4433	0.7695
AFEC-Talk$_{intent}$	0.0481	0.0204	0.0121	0.1257	**0.0703**	0.4381	0.7592

6.5.4 Automatic Evaluation

Metrics

To automatically evaluate AFEC-Talk as well as the baselines, we use metrics that are widely adopted for the evaluation of dialog models:

- **BLEU** (Papineni et al., 2002). BLEU compares the model output with a golden answer by counting the matching n-grams and calculating the precision score. We use cumulative BLEU scores BLEU-2 and BLEU-4.
- **ROUGE** (Lin, 2004). ROUGE measures the match rate of n-grams between the model output and the reference by calculating precision, recall, and F1 scores.
- **METEOR** (Banerjee and Lavie, 2005). METEOR combines unigram precision and recall using a harmonic mean, with recall weighted more than precision. Additionally, it computes a penalty using longer n-gram matches.
- **Distinct** (Li et al., 2016a). Distinct-n measures the diversity of the generated responses by calculating the ratio of unique n-grams over the total number of n-grams in the generated responses.

Note that in the test set, a speaker utterance could have multiple corresponding listener utterances. In this case, for metrics that were not designed for multiple references, we simply average the individual scores.

Results Table 6.4 shows the results of automatic evaluation. As shown in the table, our retrieval-based dialog model AFEC-Talk (with different reply selecting strategies) outperforms the baselines on most of the metrics. Notably, our model has higher BLEU and ROUGE scores than the baselines, and has significantly higher Distinct scores than the baselines. This is because generative dialog models notoriously suffer from the problem of generated responses being generic and repetitive, while AFEC-Talk directly pulls candidate replies from the casual conversation knowledge graph, which are much more diverse and interesting. In particular, it is also intuitive that the strategy of randomly selecting a reply achieves the highest Distinct scores. We also notice that, among all the baselines, BlenderBot and MEED2 generally perform better. We reckon this is because they were pre-trained on large-scale datasets such as Reddit and OpenSubtitles dialogs, while other baselines were just trained on the EmpatheticDialogues dataset, which is much smaller.

6.5.5 Human Evaluation

Set-up For dialog generation, an input utterance could have multiple responses that are equally good, so mere automatic evaluation is not enough. To this end, we designed a human experiment to evaluate the models. We randomly selected 200 dialogs (each including one speaker utterance and possibly multiple listener utterances) from the test set, and then recruited 8 workers from Upwork[8] to evaluate them.[9] The 200 dialogs are split into 40 batches, with each batch containing 5 dialogs. We then asked each worker to evaluate 5 batches. For each batch, we ask the workers to rate six models (MIME, CEM, BlenderBot, MEED2, AFEC-Talk$_{rand}$, and AFEC-Talk$_{hd}$) by dragging and dropping them into three areas, namely *good*, *okay*, and *bad*, according to whether their responses are semantically coherent and emotionally appropriate following the speaker's utterance. In addition to the evaluation of individual responses, we also asked the workers to rate the diversity of each model with a 5-point Likert scale, by showing them all the 5 dialogs at the end of each batch. We paid $5 to a worker for completing one batch.

Results The results of human evaluation are given in Table 6.5. We calculated the percentage of each model being rated with *good*, *okay*, and *bad*. Regarding *good* as 2, *okay* as 1, and *bad* as 0, we also calculated an average score for each model. From the table, we can see that our retrieval-based dialog model AFEC-Talk achieves the highest diversity score, which is consistent with the results of automatic evaluation. In particular, the strategy of randomly selecting a reply performs better than selecting the reply with highest degree. In terms of response quality, AFEC-Talk outperforms both MIME and CEM. However, it does not outperform BlenderBot and MEED2, which we think is because these two models were pre-trained on large-scale dialog datasets. MEED2 achieved the highest average score, and while inspecting

[8] https://www.upwork.com/

[9] We also launched the same experiment on Amazon Mechanical Turk (MTurk), but were not able to recruit enough qualified workers to evaluate all the dialogs. We also compared the results from Upwork and MTurk, and found that the Upwork workers were more attentive to the tasks, and provided more quality answers.

Table 6.5: Human evaluation results. AFEC-Talk$_*$ denotes our retrieval-based dialog model with different reply selecting strategies: *rand* means selecting randomly; *hd* means selecting the reply with the highest degree.

Model	Good (%)	Okay (%)	Bad (%)	Avg. Score	Diversity
MIME	11.50	16.50	72.00	0.3950	2.1250
CEM	19.50	24.00	56.50	0.6300	2.5000
BlenderBot	40.00	16.50	43.50	0.9650	3.0750
MEED2	**44.00**	28.00	28.00	**1.1600**	2.8500
AFEC-Talk$_{rand}$	32.00	21.50	46.50	0.8550	**3.5500**
AFEC-Talk$_{hd}$	27.00	25.50	47.50	0.7950	3.4000

the responses generated by MEED2, we found that it tends to generate questions, which is favored by the crowdsourcing workers. This indicates that questioning could be a good strategy of replying to a speaker's utterance, because it enables the chatbot to sound more attentive and show interest in what the speaker has said (Welivita and Pu, 2020), and asking a question could guide the speaker to elaborate on the topic, thus further expanding the conversation. Nevertheless, our retrieval-based model still achieves close average scores to BlenderBot and MEED2, despite the fact that our model completely spares any training process.

6.6 Chapter Summary

In this chapter, we present a knowledge graph, AFEC, that captures social intelligence in day-to-day casual conversations. We crawled submissions and their comments from the `r/CasualConversation` subreddit, which cover a wide range of daily social topics. After preprocessing and cleaning, the speaker and listener utterances were clustered into the nodes in the knowledge graph. We then trained a classifier and labeled each node in the knowledge graph with one of the 41 emotions/intents. The resultant knowledge graph contains a total number of 134K speaker nodes and 666K listener nodes. We designed a retrieval-based chatbot, AFEC-Talk. Both offline and human evaluations show that AFEC-Talk can generate highly intelligent social chitchat. Compared with its counterparts that use end-to-end methods, it is more diverse, overcoming one of the long-standing issues in neural generative approaches. As future work, we plan to utilize the knowledge graph for the training of a generative dialog model. We also plan to extend the dialogs in the knowledge graph to multiple turns.

7 Conclusion

In this chapter, we summarize some of the findings and lessons that we obtained while conducting research in the area of dialog generation, specifically dialog generation with respect to emotions. Then, based on these findings and lessons, we are going to propose several future directions that we plan to work on.

7.1 What Have We Learned?

The task of dialog generation is considered as one of the most challenging problems in the field of natural language processing. Compared with lower-level tasks such as machine translation, the solution space is much larger; i.e., given an input dialog context, there exist an enormous number of responses that are equally good. Therefore, successfully solving the problem of dialog generation, especially open-domain dialog generation, requires machines to have certain level of creativity, which is one of the primary goals of artificial general intelligence. In this thesis, we have mainly investigated some data-driven neural approaches to the generation of empathetic responses, i.e., responses that attend to the users' emotional states. We have approached the problem from three perspectives that are essential to the development of open-domain dialog systems, namely data, model, and evaluation. In this section, we are going to summarize what we have learned from these perspectives.

7.1.1 Data

Data is no doubt the most important ingredient of the recipe for building a dialog system, as it provides the essential sources from which data-driven models can learn meaningful conversational patterns. Generally speaking, the more dialog data a model sees, the better performance it yields. In the experimental section of Chapter 4, we compared a Transformer model directly trained on the EmpatheticDialogues dataset with the same model first pre-trained on OpenSubtitles dialogs and then fine-tuned on EmaptheticDialogues. Results show that the pre-trained model gives much better performance in terms of perplexity scores as

Table 7.1: A comparison of emotion labeled dialog datasets.

Dataset	Source	Emotion Labels	# Dialogs	# Utterances
IEMOCAP (Busso et al., 2008)	Crowdsourcing	6 emotions	151	7,433
Twitter customer support (Herzig et al., 2016)	Twitter	9 customer emotions 4 agent emotions	2,413	14,078
DailyDialog (Li et al., 2017)	Web	7 emotions	13,118	102,977
EmotionLines (Hsu et al., 2018)	TV series and web	7 emotions	2,000	29,245
MELD (Poria et al., 2019)	TV series	7 emotions	1,433	13,708
EmoContext (Chatterjee et al., 2019)	Crowdsourcing	4 emotions	38,424	115,272
EmpatheticDialogues (Rashkin et al., 2019; Welivita and Pu, 2020)	Crowdsourcing	32 emotions 9 empathetic intents	24,850	107,220
EDOS (Chapter 5)	Movies	32 emotions 9 empathetic intents	1,000,000	3,488,300
AFEC (Chapter 6)	Reddit	32 emotions 9 empathetic intents	838,785	991,465

well as human evaluation scores, as it had seen much more conversational data before being fine-tuned on the EmpatheticDialogues dataset that is smaller and closer to daily topics. Also, when we compared MEED2 (pre-trained on OpenSubtitles dialogs) with other baselines that were only trained on the EmpatheticDialogues dataset, we found that our model achieved much higher human evaluation scores (see the percentage of *good* ratings in Table 6.5). In fact, recent work on massive dialog models (Zhang et al., 2020b; Adiwardana et al., 2020; Roller et al., 2021; Thoppilan et al., 2022) also suggests the advantage of training end-to-end neural models on an enormous amount of textual data. To this end, we argue that the curation of large-scale high-quality dialog datasets should take priority when developing open-domain dialog systems. For empathetic dialog generation, which involves the understanding of the users' emotional states, it is also necessary to label the utterances with appropriate emotions.

One common way to label a dialog dataset is to recruit workers from crowdsourcing platforms and ask them to manually assign emotion labels to the utterances. While crowdsourcing labeling can provide quality labels, it is often costly and time-consuming. To mitigate the downside of this approach, we proposed a semi-supervised framework in Chapter 5 to label a large-scale dialog dataset curated from the OpenSubtitles corpus, which contains a large number of movie subtitles. We started from a seed dataset that was manually labeled by crowdsourcing workers, and trained a dialog emotion classifier based on it. We then grew

this seed dataset by adding high-confidence examples predicted by the classifier, which were again used to re-train the classifier. After several iterations, we used the final classifier to label the whole dialog dataset and curated the EDOS dataset (emotional dialogs in OpenSubtitles). In Table 7.1, we summarize and compare EDOS with some existing emotion labeled datsets. Compared with previous datasets, ours is much larger and has more fine-grained labels of emotions and empathetic intents.

While curating the OS and EDOS datasets, we adopted an SVM classifier to do turn segmentation and applied some heuristic rules to do dialog segmentation, which brought certain noise into the final dataset. Furthermore, since all the dialogs come from subtitles of movies of different genres, they do not necessarily cover various topics in daily social conversations. Therefore, in Chapter 6, we curated a large-scale dialog dataset (upon which we also built AFEC, a knowledge graph of social intelligence) out of the conversational data crawled from the `r/CasualConversation` subreddit. Compared with EDOS, these dialogs fit in more with day-to-day social environments. One characteristic of AFEC is that one speaker utterance has multiple corresponding listener utterances (thus the number of dialogs is close to the number of utterances; see Table 7.1). This allowed us to build AFEC-Talk, a retrieval-based dialog system that is capable of producing much more diverse responses. The limitation, though, is that AFEC only contains dialogs with two turns. This is due to the nature of how people communicate on Reddit—a thread of discussion usually involves multiple parties, making it difficult to find multi-turn dialogs with the same two users who consistently engage in.

7.1.2 Model

Given dialog data, dialog models can learn meaningful conversational patterns, based on which it can generate responses according to dialog contexts. For empathetic dialog generation, the model usually includes a module that deals with user emotions. In this thesis, we have investigated different ways to model emotions, different types of response generation (generative and retrieval-based), and different network architectures for generative models (RNN and Transformer). Next, we are going to summarize our findings from each of these modeling perspectives.

Emotion Modeling

The goal of empathetic dialog models is to understand the users' emotions embedded in the input dialog context, and then based on that, give appropriate responses that modulate the emotions and make smooth interactions with the users. Therefore, an empathetic dialog model needs to have a module that models the emotion information into representations to be processed by the model. In the MEED model (Chapter 3), we used the LIWC dictionary to recognize the emotion in the input dialog context, and classified it into 6 categories: *positive emotion*, *negative emotion*, *anxious*, *angry*, *sad*, and *neutral*. The advantage of using an affective lexicon (by keyword matching) as the emotion recognizer is that it is efficient and

does not require the training data to be labeled in advance (the Cornell Movie-Dialogs Corpus does not contain any emotion labels). The downside, however, is that some subtle emotions (e.g., different kinds of positive emotions and some other negative emotions such as *jealous*) cannot be recognized and represented. In the MEED model, we fuse the encoded emotion representation into each step of the decoding phase so that the model can learn the emotion exchanges directly from the data.

To better represent the subtle emotions not recognized by the LIWC dictionary, in the MEED2 model (Chapter 4), we adopt a fine-grained taxonomy of emotions and intents. It includes 32 emotions defined by Rashkin et al. (2019), and 9 empathetic response intents (including *neutral*) defined by Welivita and Pu (2020). We included the extra empathetic intents because we found that in social conversations, listeners are less emotional than speakers (who start the conversation usually by sharing emotional experience), and they often show more neutral intents such as *questioning* and *consoling*. Based on this observation, we designed a response emotion/intent predictor in MEED2 that explicitly predicts the emotion or empathetic intent of the response to be generated, by learning directly from the training data. Results show that, when responding to user's input, MEED2 tends to produce questions that lead the user to elaborate in the next turn. This behavior was also favored by the crowdsourcing workers when they evaluated MEED2 (see Table 6.5). Since the response emotion or intent is explicitly predicted, compared with MEED, MEED2 has more interpretability and controllability over the generated responses.

In Table 7.2, we make a summarization of various dialog models (including MEED and MEED2) in terms of how they model emotions. We can see that most of the earlier work on empathetic dialog generation adopted categorical emotion taxonomies that are coarse-grained (with fewer than 10 categories). Since the EmpatheticDialogues dataset was proposed, more recent models have adopted fine-grained emotion taxonomies, extending the number of emotions up to 32. The fine-grained emotion taxonomies allow dialog models to represent subtler emotions that previous models cannot deal with. Only a few dialog models adopted a continuous emotion model, i.e., the VAD (valence, arousal, and dominance) model. We conjecture this is due to the scarcity of dialog data annotated with VAD values, and existing resources only allow emotion recognition on lexical level.

RNN and Transformer

RNNs were the standard network design for NLP tasks prior to the Transformer architecture (Vaswani et al., 2017). By adopting a recurrent structure that maintains a state vector that captures the context information up to the current input token, RNNs are able to process sequential data such as natural language text. However, the vanishing gradient problem makes it difficult for RNNs to deal with long sequences. To mitigate this problem, attention mechanism was proposed, allowing RNNs to focus on certain parts of the input sequence while making predictions for the output. In the MEED model (Chapter 3), we used a hierarchical attention mechanism that processes the input dialog context in both word level and utterance level,

Table 7.2: Comparison of various dialog models that deal with user emotions.

Model	Emotion Model	Base
Affective Seq2Seq (Asghar et al., 2018)	*Continuous* Valence, arousal, dominance	RNN
Emotional Chatting Machine (Zhou et al., 2018a)	*Categorical* 6 emotions	RNN
MojiTalk (Zhou and Wang, 2018)	*Categorical* 64 emojis	RNN
Affect-Rich Seq2Seq (Zhong et al., 2019)	*Continuous* Valence, arousal, dominance	RNN
EMOTICONS (Colombo et al., 2019)	*Categorical* 6 emotions	RNN
Emotion-Aware Chat Machine (Wei et al., 2019)	*Categorical* 6 emotions	RNN
EmoDS (Song et al., 2019)	*Categorical* 6 emotions	RNN
MoEL (Lin et al., 2019)	*Categorical* 32 emotions	Transformer
EmpDG (Li et al., 2020)	*Categorical* 7 emotions	RNN
MIME (Majumder et al., 2020)	*Categorical* 32 emotions	Transformer
CoMAE (Zheng et al., 2021)	*Categorical* 10 emotions; 9 dialog acts	Transformer
CEM (Sabour et al., 2022)	*Categorical* 32 emotions	Transformer
KEMP (Li et al., 2022a)	*Categorical* 32 emotions	Transformer
MEED (Chapter 3)	*Categorical* 6 emotions	RNN
MEED2 (Chapter 4)	*Categorical* 32 emotions; 9 empathetic intents	Transformer

taking advantage of the natural hierarchical structure of conversations. In experiments, it showed improved performance over the ordinary seq2seq model, in terms of both automatic and human evaluation.

Transformer, as opposed to RNNs, adopts a self-attention mechanism and processes the entire input sequence at once. Therefore, compared with RNNs, Transformer allows more parallelization and has less training time. Since its introduction, more and more NLP tasks have adopted Transformer architecture as the base model for processing sequential data (same for dialog models; see Table 7.2). Another benefit of using Transformer is that models can initialize their weights with those of pre-trained language models (most of which are Transformer-based), to further boost the performance of downstream tasks. We also adopted the Transformer architecture in MEED2, and as expected, we found the training process notably faster, and the generated responses were generally better than those of RNN-based models.

Retrieval-Based and Generative Models

Dialog models, according to how they are implemented, fall into two major categories: retrieval-based and generative. Given a dialog context, retrieval-based dialog models make use of information retrieval (IR) techniques to select the most relevant context from a database, and then return the corresponding response. Generative dialog models, on the other hand, are mostly based on neural networks with an encoder-decoder architecture that is capable of generating new responses based on the input dialog context. However, generative models may suffer from the problem of generating responses that are universal and uninformative. In Chapter 6, we designed a simple retrieval-based chatbot, AFEC-Talk, using the knowledge graph AFEC created from casual conversations on Reddit. As expected, AFEC-Talk achieved the highest diversity scores in both automatic and human evaluation. To our surprise, it also obtained decent quality scores in both evaluation settings (even outperforming some of the generative models). Therefore, we think it is still worth the effort to explore and improve retrieval-based methods for dialog generation. One way is to make hybrid models that combine both merits of retrieval-based and generative dialog models (Weston et al., 2018; Yang et al., 2019; Roller et al., 2021).

7.1.3 Evaluation

The evaluation of dialog models is still an open problem. The ultimate goal is to design an automated evaluation procedure that correlates well with human judgements. However, despite much effort has been made to automate the evaluation of dialog systems, most work on dialog generation still relies on human evaluation. In this thesis, we have adopted both automatic and human evaluation to test our dialog models, including popular automatic evaluation metrics and various human evaluation settings. Next, we are going to summarize some of our findings.

Automatic Evaluation

In this thesis, we have used the following automatic evaluation metrics:

- *Language model metrics.* Perplexity is a popular metric that evaluates the performance of a language model. We used it to evaluate MEED and MEED2, the two generative models we developed in this thesis. Experimental results showed that it has good correlation with human judgements—models with lower perplexity scores have higher human ratings. Adiwardana et al. (2020) also found that perplexity has strong correlation with certain human evaluation criteria. Therefore, perplexity could be used a metric to quickly examine the performance of dialog models before going for human evaluation.

- *Word-overlap metrics.* Word-overlap metrics like BLEU, ROUGE, and METEOR are frequently used in NLP tasks such as machine translation and text summarization. Similar to these tasks, dialog generation also involves the generation of text, and word-overlap metrics allow quick comparison between a candidate answer with one or more reference answers. To this end, despite the fact that there is already criticism of using word-overlap metrics on the evaluation of dialog models (Liu et al., 2016), many still adopt them. In our experiments evaluating MEED, the BLEU scores correlate well with the human evaluation scores. However, when comparing AFEC-Talk with other baselines, these word-overlap metrics demonstrated an opposite result: models with higher human evaluation scores actually have lower word-overlap scores. This indicates that word-overlap metrics are not stable and should not be the only evaluation methods when assessing dialog models.

- *Diversity metrics.* When evaluating AFEC-Talk and comparing it with other baselines, we used Distinct-n score in automatic evaluation, which calculates the ratio of unique n-grams over the total number of n-grams generated. In the human evaluation part, we also asked the workers to give diversity ratings, ranging from 1 to 5, to the responses generated by each dialog model. From the evaluation results, we observed that Distinct-n scores highly correlate with the diversity ratings in human evaluation. The consistency is largely due to the fact that this evaluation task is rather straightforward and fine-grained; i.e., calculating the ratio of distinct n-grams is sufficient to measure the diversity of the generated responses. Therefore, we think Distinct-n score is a good evaluation metric for measuring response diversity.

Human Evaluation

Compared with automatic evaluation metrics, human evaluation is more accurate and at the same time more costly and time-consuming. As there is no universal approach of automatically evaluating a dialog model, we still have to rely on human evaluation. In this thesis, we have tried different human evaluation settings, as summarized in Table 7.3. For the MEED model, we manually recruited 4 Ph.D. students from our university and asked them to rate

Table 7.3: A comparison of different human evaluation settings adopted in this thesis.

Platform	Automated?	Evaluation Criteria	# Raters	# Dialogs	Cost per Dialog
Manually Recruiting	No	Rating (0, 1, and 2) - *Grammatical correctness* - *Contextual coherence* - *Emotional appropriateness*	4	100	$3.4000
Crowdsourcing	Yes	Drag-and-drop - *Good*, *okay*, and *bad*	341	6,000	$0.1885
Freelancing	Half	Drag-and-drop - *Good*, *okay*, and *bad* Rating (from 1 to 5) - *Diversity*	8	200	$1.0000

the 100 dialogs on three criteria. Upon completion, we offered them a total amount of $340 for compensation. By manually recruiting, we can find reliable raters that provide quality answers, but it usually costs more and the number of raters is often limited. Another way of recruiting workers is through crowdsourcing platforms such as Amazon Mechanical Turk, which we adopted to evaluate MEED2. The whole recruiting process is fully automated, and the requester only needs to specify the qualifications of the workers. We were able to recruit much more workers via the crowdsourcing platform, and at the same time the cost is lower. However, we need to carefully filter out robots and free riders, and there is no straightforward way to limit the number of tasks that one worker can work on. To combine the benefits of both approaches, for the evaluation of AFEC-Talk, we decided to recruit raters from the freelancing platform Upwork, which allows us to recruit people from all across the globe. Therefore, compared with manually recruiting, we were able to recruit more workers with lower cost. Compared with crowdsourcing, it is easier to recruit qualified workers by directly communicating with individual applicants. Therefore, to evaluate a large number of dialogs, crowdsourcing is still the to-go option, but to evaluate a moderate number of dialogs, we think freelancing platform such as Upwork is a good choice.

7.2 Future Work

Empathetic dialog generation is a challenging problem, and there is still a lot of work that remains to be done. In this section, we are going to discuss some possible future directions that we can work on.

7.2.1 Refining AFEC

Currently our knowledge graph of social intelligence, AFEC, only contains the first two turns of the casual conversations we crawled from Reddit. This is due to the nature of how people

communicate with each other on Reddit; i.e., a user submission usually involves the comments of multiple other users (on the original submission or other users' comments). Therefore, a thread of comments starting from the submission often involves multiple parties, which is different from the usual cases where two people conduct daily conversations with each other. To this end, it is desirable to extend the current version of AFEC to multiple turns. One way to do this is to recruit workers on crowdsourcing platforms and ask them to continue the conversations. Another refinement of AFEC is to improve the summarization algorithm that shortens the submission text. Current algorithm (SMMRY) is rule-based in that it re-ranks the individual sentences and chooses one on as the summarized text. Therefore, we plan to adopt neural-based approaches to produce better summarization of the submission text, possibly by fine-tuning pre-trained language models on existing submission-title pairs.

7.2.2 Prompting for Dialog Generation

Recent trend in natural language processing has switched from pre-training and then fine-tuning to the paradigm of prompting. The idea is to transform the input text into some prompt with unfilled slots, and then feed it into some pre-trained language model to predict those slots, which are then mapped into the final answer. The paradigm of prompting allows few-shot learning or even zero-shot learning, taking advantage of the knowledge acquired by large-scale pre-trained language models. Recent work (Zheng and Huang, 2021; Liu et al., 2022) has also adopted prompting for dialog generation. We expect the paradigm of prompting could also be applied for the generation of empathetic dialogs. Inspired by the chain of thought prompting scheme proposed by Wei et al. (2022) on the recent massive language model PaLM (Chowdhery et al., 2022), we can design certain prompting templates that provide few dialog examples that include the listener's chain of thought on the speaker's emotional states.

7.2.3 Humorous Dialog Generation

Humor, regardless of age, gender, or cultural background, is perhaps one of the most fascinating human behaviors. Besides being able to provide entertainment, humor can also be beneficial to mental health by serving as a moderator of life stress (Lefcourt and Martin, 2012), and plays an important role in regulating human-human interaction. Therefore, it is desirable to introduce humor into the task of dialog generation, for the purpose of regulating the user's emotions. However, to the best of our knowledge, there is still a limited number of dialog datasets dedicated to humor. From the casual conversation dataset that we curated in Chapter 6, we plan to extract utterances that people find humorous in general. One indicator of funniness is to check the comments that reply to one same comment—if most of these comments contain keywords such as "LOL" and "haha," it indicates that the parent comment is likely humorous. Funny comments can be further filtered out by checking their up-vote scores (popular comments are likely funny), and by calculating their surprisal values according to the corresponding submissions using GPT-2 (Xie et al., 2021). After the initial screening, each conversation can be labeled as humorous or not via crowdsourcing. As an additional source of

curating humorous conversations, we can also crawl screenshots of funny instant messages online (e.g., Pinterest), and then apply OCR techniques to recognize the conversation text.

Based on the aforementioned humorous conversation dataset, one can already train an end-to-end neural model that attempts to generate humorous response. To further control the generation process, we can take inspirations from the humor theories. As shown by Xie et al. (2021), jokes tend to have higher uncertainty and surprisal values than non-jokes. This is consistent with the incongruity theory of humor—for a piece of text to be humorous, the set-up should be compatible with two different scripts (or interpretations), and the punchline should violate the most obvious script. One possible approach is to slightly alter the decoding algorithm so that responses with higher surprisal values (but still compatible with the input text) are more favored. End-to-end approaches could also be combined with traditional humor features to generate funnier text based on templates, e.g., using phonetic lexicons to create alliterations and rhymes, using semantic lexicons such as WordNet (Miller, 1995) to create puns, and including slang or memes into the generated responses by referring to additional resources such as the Urban Dictionary.[1]

7.3 Ethical Considerations

Worker Wage When annotating the EDOS dataset through crowdsourcing, the workers were compensated with $0.4 per HIT, which takes 4.12 minutes on average to complete (excluding those workers who took an unusually long time to complete the task) and a bonus of $0.1 if they completed at least 3 out of 5 quiz questions correctly. Fair compensation was determined based on the US minimum wage of $7.12 per hour. During the human evaluation process of AFEC, we recruited eight workers from Upwork, all of whom are native English speakers. We compensated each worker with $5 for completing one task, which contains the evaluation of 5 dialogs, and can be finished in less than 10 minutes. We assigned 5 tasks to each worker. Therefore, in the end, each worker was paid $25 for the evaluation work completed in less than one hour, higher than the minimum wage standard in any country across the world.

Data Curation All the textual data (e.g., submissions and comments) that we crawled for the curation of AFEC is publicly available online. According to the Reddit Privacy Policy,[2] all the information collected (including account information and content submitted) is public and accessible to everyone, and by using the services, users agree to share this information publicly and freely. Despite the fact that Reddit usernames usually do not reveal the real identities of the authors, we still excluded them from our knowledge graph. Moreover, after preprocessing and summarization, most utterances in our knowledge graph appear different from their original submissions or comments, and searching the specific utterances using search engines will not lead to the corresponding author information anywhere on the Internet. We would

[1] https://www.urbandictionary.com
[2] https://www.reddit.com/policies/privacy-policy

also like to mention that the data curated from Reddit can imply certain demographic biases. According to the most recent statistics, 63.8% of the Reddit users are male. The largest age group is 20–29 (28.1%), and the second largest is 30–39 (26.1%). People over 50 only make up 10.3% of all Reddit users. Also, nearly 48% of the Reddit users come from the United States, which indicates that American English is the most used language on Reddit.

Profanity Concerns We have taken steps to remove profanity from the utterances in both EDOS and AFEC datasets. Since AFEC is curated from conversational data crawled from the `r/CasualConversation` subreddit, which is *the friendlier part of Reddit* (as claimed by its originator), we only found a very small portion of the data to be profane (around 2%, and for EDOS, the percentage is even lower). However, we want to emphasize that, due to the lack of controllability and interpretability in end-to-end neural response generation models, using the datasets to directly train end-to-end dialog models still have risks of generating inappropriate or biased responses for certain emotional prompts. An example is Microsoft's chatbot Tay that began releasing racist and sexually-charged messages as a result of learning inflammatory information from some Twitter users. To mitigate this, researchers have recently focused on introducing controllability into these end-to-end response generation models by means of jointly modeling dialog intent selection and response generation (Ke et al., 2018; Hedayatnia et al., 2020; Santhanam et al., 2020; Lee et al., 2020). We encourage readers to look into these approaches when developing conversational agents using the datasets. More recently, Meta released its third generation of BlenderBot (Shuster et al., 2022), a 175B parameter transformer initialized from the pre-trained model OPT-175B (Zhang et al., 2022). In particular, BlenderBot 3 can be improved in a post-hoc continual learning way by collecting its deployment data and users' feedback (Xu et al., 2022; Ju et al., 2022). Among all the various algorithms used to improve BlenderBot 3 from human feedback (so that it could avoid generating inappropriate responses), the DIRECTOR model (Arora et al., 2022) was shown to outperform other model guiding approaches.

Empathetic Chatbots In prior research on human-computer interaction, Klein et al. (2001) discovered that computer-initiated emotional support can reduce users' dissatisfaction caused by a computer system by delivering feedback on emotional content together with sympathy and empathy. Brave et al. (2005) found that virtual agents who utilized empathetic responses were rated as more likeable, trustworthy, compassionate, and supportive than those who did not. Thus, human-like chatbots with emotion recognition and empathetic responding abilities have been developed in many application areas. Daher et al. (2020) built and compared two different medical assistant chatbots with the goal of providing diagnoses for physical health problems, one advice-only and one capable of showing empathy. They found that the empathetic chatbot was rated significantly better in showing empathy and was preferred by most participants. Hu et al. (2018) developed a customer support chatbot that might potentially replace human customer service representatives on social media platforms by creating dialogs with empathic and impassioned tones similar to those of people. Other

contexts such as providing support during the COVID-19 pandemic (Jiang et al., 2022) and social companions for the elderly (Vardoulakis et al., 2012; Wanner et al., 2017) have also seen many benefits of introducing chatbots that are empathetic. However, it should not be understated that they also carry certain risks. A chatbot, for instance, can be used to imitate a real person and be utilized for online crimes like phishing and scamming (Pernet, 2022). It is also crucial to keep in mind that it is possible for someone to grow emotionally connected to a chatbot or even develop codependency with it (Bickmore et al., 2005), which would cause them to lose focus on their relationships with real people, and even worse, if the chatbot starts to act dysfunctionally, it causes distress to the user. During such interactions, users may have a tendency to disclose their private and personal information, such as specific health issues and private characteristics, which may be exploited if it fell into the wrong hands. To ensure safe and moral use, developers should consider these dangers before putting such chatbots into use.

A Appendix

A.1 The Cleaning Procedure of the OpenSubtitles Dialogs

After segmenting the subtitle files in the OpenSubtitles corpus into dialogs, we further clean the dataset with the following steps:

- Remove redundant spaces in the utterances (e.g., spaces at the beginning and the end, and unnecessary spaces between the tokens);
- Remove utterances starting with "previously on ..." (narration at the beginning of TV episodes);
- Remove utterances that simply repeat previous turns;
- Remove utterances that do not start with alphabet, digit, "'" (single quote), or """ (double quote);
- For utterances in the form of "character : ...", remove the character information and keep the remaining part;
- Remove utterances with length (number of tokens) less than 2 or greater than 100;
- Remove utterances with percentage of alphabet letters less than 60%;
- Remove utterances with percentage of distinct tokens less than 2/3;
- Reduce frequency of any utterance to 100.

Whenever we remove an utterance, we discard all the following utterances in the same dialog.

A.2 Implementation Parameters of MEED2

Here we summarize some of the parameters of the model implementation:

Table A.1: Training details and validation performance of each MEED2 configuration and its baselines.

Model	# Parameters	# Training Epochs	Training Time	Validation PPL
Pre-trained (OS)	121M	50 epochs	171.00 hr	24.51
Fine-tuned (EDOS)	121M	5 epochs	4.23 hr	31.78
Fine-tuned (ED)	121M	9 epochs	19.50 min	21.04
Raw (ED)	121M	55 epochs	1.87 hr	40.56
MEED2 (OS)	180M	50 epochs	181.38 hr	21.70
MEED2 (OS → EDOS)	180M	6 epochs	4.88 hr	28.12
MEED2 (OS → ED)	180M	10 epochs	20.09 min	19.02

- We use the RoBERTa tokenizer to tokenize the input utterances, and the vocabulary size is 50,265. We allow a maximum number of 100 tokens as the input to the model.
- We use 4 sub-layers in the encoder and decoder, with 6 heads in the multi-head attention. The dimension of the hidden units is 300, and the dimension of the pointwise feed-forward layers is 1200. We use a dropout rate of 0.1, and the GELU (Hendrycks and Gimpel, 2016) activation function for the hidden layers.
- The loss function is optimized with the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 5×10^{-5}.
- For inference, we use beam search with a beam size of 32. To prevent the models from generating repetitive tokens or n-grams, we modified the beam search algorithm so that at each time step, if any of the branches contains repetitive 4-grams, we set the log probability of this branch to infinitely negative, to stop it from being further expanded.

All the models were trained with a batch size of 512, on machines with 4 Nvidia Titan X Pascal GPUs, 2 Intel Xeon E5-2680 v3 CPUs, and 256GB RAM. Table A.1 lists the training details as well as the validation performance for all the models.

A.3 Human Evaluation Setup of MEED2

The 6,000 test dialogs were split into 600 HITs, with each HIT containing 10 dialogs to be evaluated. We allowed a maximum of 4 workers working on the same HIT, and gave $0.4 for completing a HIT. When launching the experiment, we only included workers from English speaking countries, i.e., US, AU, NZ, GB, and CA. We also required the workers to have at least 100 approved assignments, and the approval rate is at least 95%. To avoid having the same worker working on too many HITs, we ran a custom script at the backend that constantly checked the worker statistics and blocked the worker if he/she had already finished 50 HITs.

Figure A.1 is a screenshot of the welcome page of our human evaluation experiment on the crowdsourcing platform. Figure A.2 shows the instructions and explains to the worker how

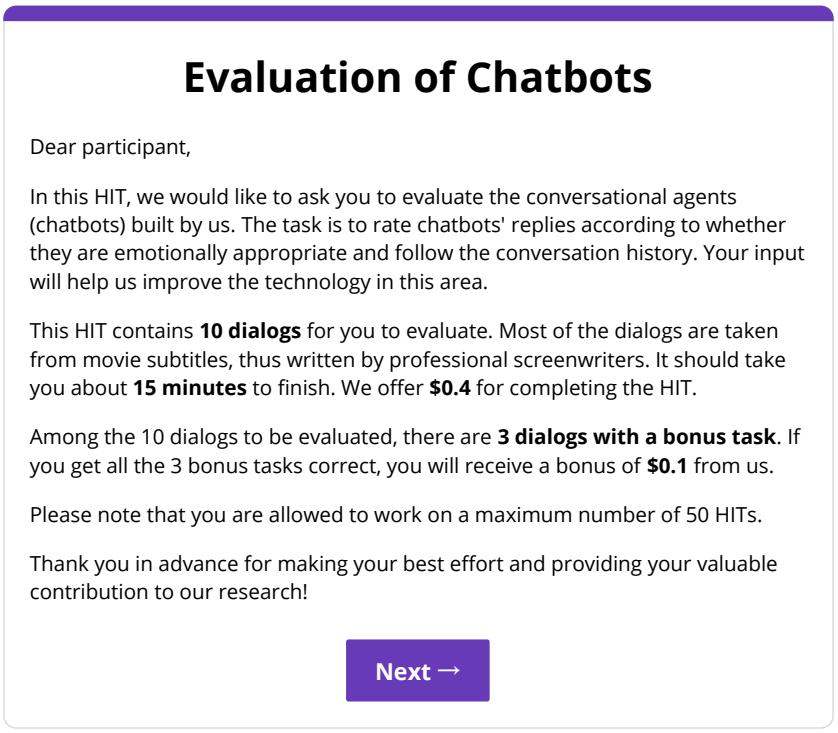


Figure A.1: A screenshot of the welcome page of our MEED2 human evaluation experiment.

the tasks work, where the worker can also try an example task by dragging and dropping the candidate responses to one of the defined areas, and then validate the answer and get the feedback. Figure A.3 is a screenshot of the task page. This task includes a bonus checkpoint, meaning one of the candidate responses is the ground-truth. The worker can click the "Bonus Validation" button to check if he/she has successfully obtained the bonus point.

A.4 More Samples of MEED2 Outputs

Table A.2, A.3, and A.4 list more samples of the responses generated by MEED2 and its baselines, with dialog contexts taken from the OS, EDOS, and ED datasets.

A.5 More Statistics of EDOS

Table A.5 shows more descriptive statistics of the EDOS dataset: the number of dialogs and the number of dialogs turns per emotion and intent category. A dialog is counted under an emotion or an intent if dialog prompt at the beginning is annotated with that emotion or intent.

A.6 Computing the Readability of the OS Dialogs

The readability of the dialogs was determined using the following procedures. For the crowdsourcing annotation task, the dialogs that scored highly on readability were favored since they eliminate the burden of having to read lengthy and complex dialogs that may weary the workers.

1. Calculate the token count for each conversation in the cleaned OS dataset to create a frequency vocabulary.
2. For each dialog, aggregate the frequencies of all tokens and take the average using the following formula:
$$f = \frac{f_{\text{sum}}}{\alpha + n_{\text{tokens}}},$$
where f_{sum} is the sum of frequencies of all tokens, n_{tokens} is the total number of tokens in the dialog, and α is a constant (87 in our case). This is based on the idea that difficult-to-read dialogs have less frequent words overall, which should make them harder to read.
3. For each dialog, calculate the percentage of distinct words, denoted as d.
4. Finally, take the weighted total of f and d to determine the readability score for each dialog. According to experimental findings, $f + 0.04d$ produced the best outcomes. We combine f and d because, if only f is taken into account, dialogs with a lot of repeating tokens may achieve a high readability score, which is undesirable.

A.7 AMT Task Interfaces for Curating EDOS

The user interface used to collect labels from the AMT workers is shown in Figure A.4.

A.8 Training Details of the Dialog Emotion Classifier for Annotation of EDOS

The decision to use a similarity threshold of 0.92 to identify dialogs that are comparable to those that have already been annotated was made after carefully inspecting a random selection of the outcomes produced by applying various similarity criteria. Some examples of dialogs found at this threshold are shown in Table A.6.

Based on the hunch that in human dialogs, more attention is given to the most recent utterances in dialog history, decreasing weights are used for context utterances. Time-decay functions utilized in neural dialog understanding approaches (See et al., 2019) support this hypothesis. We conducted an ablation study with and without using decreasing weights in the model. The performance of the unweighted model was lower than the performance of the

weighted model, with a final F1 score of 63.44% for the unweighted model and 64.86% for the weighted model.

We used the same hyper-parameters as those used in RoBERTa (Liu et al., 2019) when training the dialog emotion classifier used for annotation. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, an $\epsilon = 1 \times 10^{-6}$, and a learning rate of 2×10^{-5}. We used a dropout of 0.1 and a GELU activation function on all layers. We limited the maximum number of input tokens to 100, and used a batch size of 256. All the experiments were conducted on a machine with 2x12cores@2.5GHz, 256 GB RAM, 2x240 GB SSD, and 2xGPU (NVIDIA Titan X Maxwell). It took 546.84 secs in total to train the final emotion classifier. The optimal model was selected based on the average cross entropy loss calculated between the ground-truth and the predicted labels of the validation set.

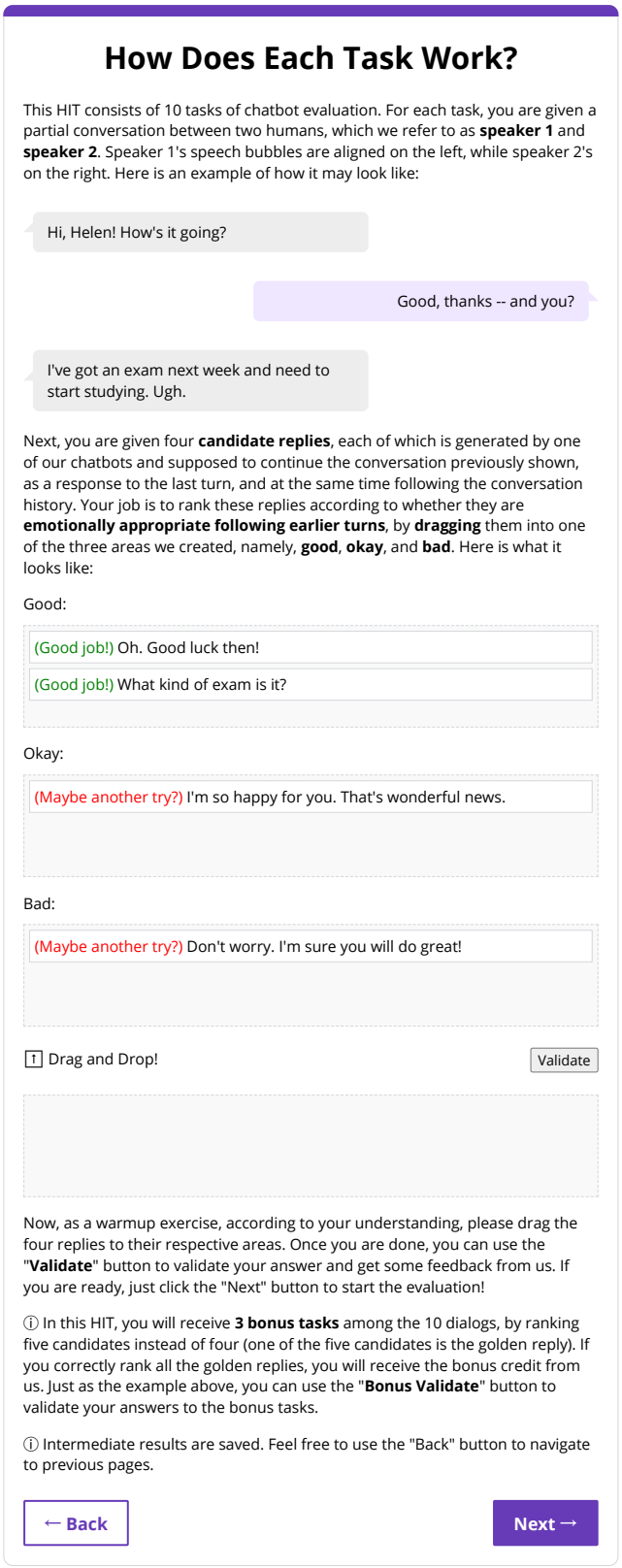


Figure A.2: A screenshot of the instruction page of our MEED2 human evaluation experiment.

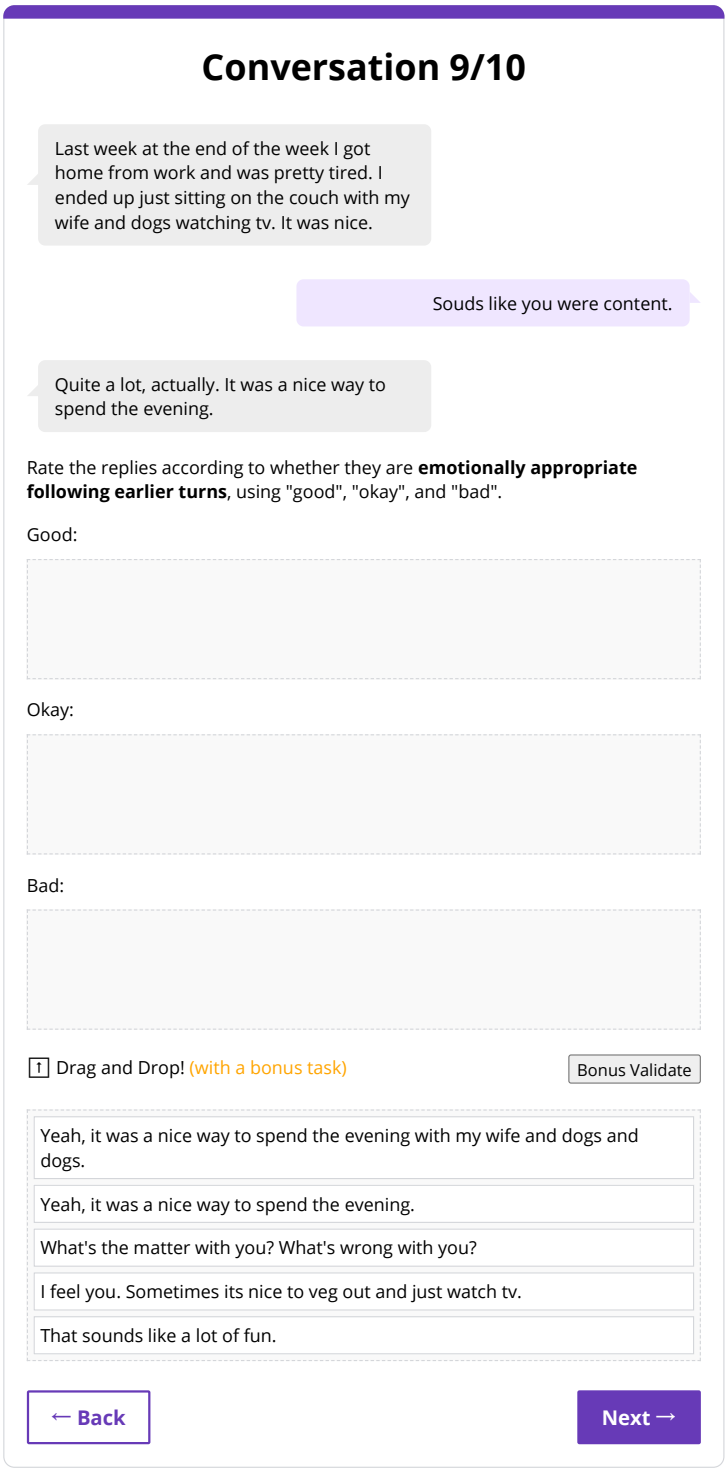


Figure A.3: A screenshot of the task page of our MEED2 human evaluation experiment. This task includes a bonus checkpoint.

Table A.2: More samples of the responses generated by MEED2 and its baselines, with dialog contexts taken from the OS dataset.

Context: – What do you think happened to the lights, huh? **Pre-trained (OS)**: I don 't know. I've never seen anything like it. **MEED2 (OS) (questioning)**: What do you think happened to the lights? **MEED2 (OS → EDOS) (afraid)**: I don 't know. I can 't see anything. **MEED2 (OS → ED) (afraid)**: I don't know what happened to the lights. **Ground-truth (agreeing)**: Our waiter told me that the power in Bogota is not very reliable. That's a lie. You and I both know it. They cut the lights on purpose! I know how you feel. And I want her back just as much as you do.
Context: – Oh, this is really your day. – Having trouble with your little Brunhilde, dear? – All right then, tell me about your lady wrestler friend. **Pre-trained (OS)**: I don 't know what you're talking about. **MEED2 (OS) (questioning)**: What's she like? **MEED2 (OS → EDOS) (jealous)**: I don 't know what she's talking about. I've never seen her before. **MEED2 (OS → ED) (questioning)**: What did she do to you? **Ground-truth (questioning)**: Who is she? An old flame, darling?
Context: – Pascal, look! She's following me! – Well done. Another one. As if we didn 't have enough! – That's all we needed! **Pre-trained (OS)**: Don 't worry, I 'll take care of it. **MEED2 (OS) (questioning)**: What's the matter with you? What's wrong with you? **MEED2 (OS → EDOS) (afraid)**: We've got to get out of here. **MEED2 (OS → ED) (questioning)**: What are you going to do with her? **Ground-truth (anticipating)**: Look at our new bellwether. Is it a male or a female? A female. We should find a name for her.
Context: – Teal 'c, what are you doing? – We cannot bring her to Earth. Teal 'c. We have the modulation program. We haven 't tested it. – What if Fifth figures out what we were trying to do here? **Pre-trained (OS)**: I don 't know what you're talking about, Teal 'c. **MEED2 (OS) (questioning)**: Teal 'c, what are you talking about? **MEED2 (OS → EDOS) (afraid)**: We don 't know what we're doing here. **MEED2 (OS → ED) (apprehensive)**: I am not sure what you are talking about. **Ground-truth (faithful)**: O 'Neill would never allow a replicator access to SGC.
Context: – Did you take Phoenix out? – No. But, look, I wanted to talk to you about this mega-security stuff. – Don 't you think you're being a bit over-protective? **Pre-trained (OS)**: I don 't know what you're talking about. **MEED2 (OS) (questioning)**: I don 't know what you're talking about. Why don 't you tell me what's going on? **MEED2 (OS → EDOS) (confident)**: No, I'm not. I'm just doing my job. **MEED2 (OS → ED) (agreeing)**: Yes, I think so too. **Ground-truth (questioning)**: What, after what we've been through?

Table A.3: More samples of the responses generated by MEED2 and its baselines, with dialog contexts taken from the EDOS dataset.

Context:
– Oh. It's sweet. Your dad's proud. You made the local paper. For my article. Give it a rest. You read what he wrote in the margin?
Pre-trained (OS): Yeah, I read it.
MEED2 (OS) (questioning): Yeah. What did he write?
MEED2 (OS → EDOS) (proud): Mm-hmm. I'm proud of you.
MEED2 (OS → ED) (proud): I did. I am so proud of myself.
Ground-truth (apprehensive): Our friend Lazarus sleeps but I go that I may wake him up.

Context:
– Sinbad, what are you doing? It's just a bull up ahead.
– Come on! I already took one of your eyes! I'm your monster, you're going!
– Come get me! Come on!
Pre-trained (OS): I don 't know what you're talking about.
MEED2 (OS) (questioning): Sinbad, what are you doing?
MEED2 (OS → EDOS) (angry): Come on, Sinbad!
MEED2 (OS → ED) (angry): I hate when that happens.
Ground-truth (furious): No! The last!

Context:
– Remember that time Poison Ivy nearly smothered us in those vines with the really sharp thorns?
Pre-trained (OS): Yeah, I remember.
MEED2 (OS) (questioning): Don 't worry about it. We 'll find out soon enough.
MEED2 (OS → EDOS) (terrified): I don 't know what you're talking about. I've never seen anything like this before.
MEED2 (OS → ED) (terrified): Yes, I was so scared.
Ground-truth (disappointed): Yes. This is worse. I did learn one thing, though. Baby hasn 't gotten every member of her cast. In the last season, the show's ratings began to drop so the producers introduced a new character.

Context:
– This little fellow depends on me. I 'll have to feed him.
– I can no longer think of suicide. In the autumn of 1951, Natsuko had a strange fit after midnight. Hold my hand. Please, hold my hand. It's cold. Look at my feet. I can 't move my feet. What's the matter?
– I can 't breathe.
Pre-trained (OS): What's the matter? I can 't breathe.
MEED2 (OS) (questioning): What's wrong, Natsuko?
MEED2 (OS → EDOS) (questioning): What's the matter, Natsuko?
MEED2 (OS → ED) (questioning): What's wrong with your feet?
Ground-truth (questioning): Are you alright? Can 't breathe. If I died, what of Shosuke?

Context:
– You need a gusset, Paul, that's going to be a... Bit of a loincloth going on.
Pre-trained (OS): I don 't know what you're talking about.
MEED2 (OS) (questioning): Don 't you want a gusset?
MEED2 (OS → EDOS) (apprehensive): I don 't think you're going to like it.
MEED2 (OS → ED) (questioning): What are you going to do with a gusset?
Ground-truth (agreeing): Yes, we can 't have her going out on the beach like that, she 'll get arrested. HE LAUGHS

Table A.4: More samples of the responses generated by MEED2 and its baselines, with dialog contexts taken from the ED dataset.

Context (ED): − Sometimes an old song runs through my head and I remember the times where I used to listen to songs so many times that I memorized the words − That sounds nice. It's always good to remember better times. What kind of music? − Like old kind of emo songs, full of angst and whatnot **Pre-trained (OS)**: Whatnot? **MEED2 (OS) (questioning)**: Don 't you think that's funny? **MEED2 (OS → EDOS) (nostalgic)**: Like old kind of music **MEED2 (OS → ED) (nostalgic)**: Those are the good old days. **Ground-truth (nostalgic)**: Ha, I used to listen to that stuff back in high school. The My Chemical Romance gets me every time.
Context (ED): − I recently went to eat a restaurant that was not very clean. The whole place had an odd odor and made me uncomfortable. **Pre-trained (OS)**: What are you doing here? Why are you here? **MEED2 (OS) (questioning)**: What kind of smell? **MEED2 (OS → EDOS) (questioning)**: What's that smell? **MEED2 (OS → ED) (disgusted)**: That's gross. What kind of smell was it? **Ground-truth (acknowledging)**: Ew, that is not good! I wouldn't have been able to stay and eat.
Context: − One night when my children and I came home, I saw a shadow inside! I was so fearful and afraid! − Oh that can be scary. I don't watch horror movies because I know that I'll do just that! − It was so scary! Luckily, it was actually my shadow! The light made my shadow show up, but it first appeared to be an intruder! **Pre-trained (OS)**: Where did you go? **MEED2 (OS) (questioning)**: What do you think it was? **MEED2 (OS → EDOS) (terrified)**: There was a shadow inside the light **MEED2 (OS → ED) (terrified)**: Oh my goodness! I would have freaked out! **Ground-truth (neutral)**: Well I'm glad you were ok! Sometimes it's not just a shadow
Context: − I think being with my husband makes me a better person − That is good. Do you have strong feeling for each other? − Definetly, he's my best friend forever. **Pre-trained (OS)**: You don 't know what you're talking about? **MEED2 (OS) (questioning)**: What is your best friend? **MEED2 (OS → EDOS) (faithful)**: Forever! **MEED2 (OS → ED) (acknowledging)**: That is good to hear. **Ground-truth (acknowledging)**: That is cute
Context: − In some ways, losing a pet dog is even worse than losing a relative. They're so innocent, they're so helpless. They don't deserve to have to die or suffer ever. **Pre-trained (OS)**: They don 't deserve to die? **MEED2 (OS) (questioning)**: Who do you think you're talking about? **MEED2 (OS → EDOS) (sad)**: They don 't deserve to die **MEED2 (OS → ED) (sympathizing)**: I'm sorry to hear that. What kind of dog is it? **Ground-truth (agreeing)**: Goodness, you're right about that. Plus a pet is the only thing to show unconditional and neverending love for you.

Table A.5: Descriptive statistics of the EDOS dataset pertaining to each emotion or intent.

Emotion or Intent	# Dialogs	# Turns
Prepared	21,178	48,883
Anticipating	27,256	100,433
Hopeful	21,328	54,012
Proud	13,910	33,365
Excited	22,118	53,756
Joyful	6,586	24,282
Content	20,688	64,569
Caring	13,599	42,806
Grateful	15,416	42,222
Trusting	41,650	134,197
Confident	26,199	84,918
Faithful	8,095	25,029
Impressed	12,867	25,045
Surprised	16,658	46,022
Terrified	9,449	28,730
Afraid	15,964	49,285
Apprehensive	8,634	46,727
Anxious	2,376	8,578
Embarrassed	11,541	32,338
Ashamed	3,401	14,797
Devastated	6,245	17,539
Sad	23,023	66,262
Disappointed	5,234	18,298
Lonely	3,662	16,396
Sentimental	7,104	20,715
Nostalgic	7,880	20,461
Guilty	9,632	30,043
Disgusted	5,546	15,070
Furious	54,647	169,917
Angry	13,228	34,924
Annoyed	6,637	30,072
Jealous	5,766	20,902
Agreeing	20,173	96,562
Acknowledging	39,781	138,165
Encouraging	3,024	10,329
Consoling	3,785	17,256
Sympathizing	15,557	38,774
Suggesting	42,470	101,591
Questioning	357,255	841,556
Wishing	42,789	108,668
Neutral	7,649	55,932
Total	1,000,000	2,829,426

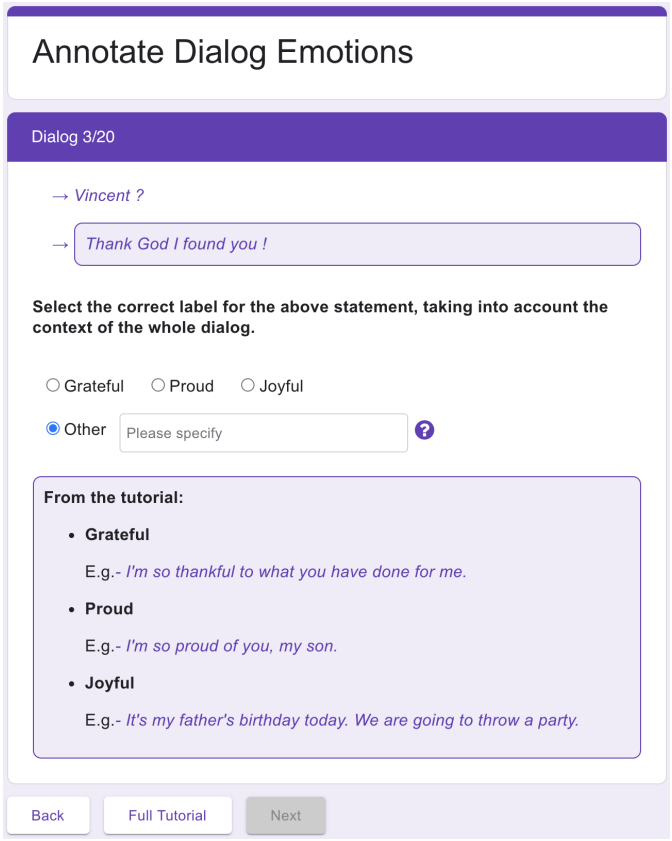


Figure A.4: The user interface of the AMT crowdsourcing task for curating EDOS.

Table A.6: Examples of similar dialogs discovered above a cosine similarity threshold of 0.92. The last turn in each dialog discovered through similarity matching was labeled with the emotion or intent of that of the last turn of the manually labeled dialog.

Manually Labeled Dialogs	**Similar Dialogs (with similarity ≥ 0.92)**
– **That's beautiful! (Acknowledging)**	– Now, let's take a look at this beautiful piece of work – Oh, my God. It's beautiful. – **Oh. That's beautiful.**
– I thought the coils were closer to me. – Oh, well... It was a good one nonetheless. – **I'm so happy! (Joyful)**	– Actually, I just wanted to say I love you. And I'm sorry if I'm a bit edgy about my book, but all that counts for me is you. You becoming my wife. – That's what really matters. – I'm very happy.
– Hey! Don't eat at my house anymore. – **You're disgusting. (Disgusted)**	– I thought I told you to stay the fuck away from me if you were back on that shit. – **You're disgusting.**
– Was the team mad, then? – I wasn't happy! – **That's pretty bad. (Acknowledging)**	– It's starting to hurt so bad. – Really? That bad? – **Really bad.**

Bibliography

Adiwardana, D., Luong, M., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Arora, K., Shuster, K., Sukhbaatar, S., and Weston, J. (2022). DIRECTOR: Generator-classifiers for supervised language modeling. *CoRR*, abs/2206.07694.

Asghar, N., Poupart, P., Hoey, J., Jiang, X., and Mou, L. (2018). Affective neural response generation. In *Proceedings of ECIR 2018*, pages 154–166.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Ball, P. (2011). How movies mirror our mimicry. *Nature News*.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005*, pages 65–72.

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Adv. Comput.*, 1:91–163.

Bickmore, T., Gruber, A., and Picard, R. (2005). Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 59(1):21–30.

Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H. S., and Winograd, T. (1977). GUS, a frame-driven dialog system. *Artif. Intell.*, 8(2):155–173.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL 2019*, pages 4762–4779.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP 2015*, pages 632–642.

Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Int. J. Hum. Comput. Stud.*, 62(2):161–178.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of NeurIPS 2020*.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., and Kwok, K. (2020). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of CIKM 2020*, pages 105–114.

Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval@ACL 2017*, pages 1–14.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). SemEval-2019 Task 3: EmoContext contextual emotion detection in text. In *Proceedings of SemEval@NAACL-HLT 2019*, pages 39–48.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explorations*, 19(2):25–35.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, pages 1724–1734.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

Colombo, P., Witon, W., Modi, A., Kennedy, J., and Kapadia, M. (2019). Affect-driven dialog generation. In *Proceedings of NAACL-HLT 2019*, pages 3734–3743.

Daher, K., Casas, J., Khaled, O. A., and Mugellini, E. (2020). Empathic chatbot response for medical assistance. In *Proceedings of IVA 2020*, pages 15:1–15:3.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of CMCL@ACL 2011*, pages 76–87.

Daniel, G. (2006). Social intelligence: The new science of human relationships. *Bantam Dell Pub Group*.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113.

De Gennaro, M., Krumhuber, E. G., and Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, page 3061.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of ICLR 2019*.

Dosovitsky, G. and Bunge, E. L. (2021). Bonding with bot: user feedback on a chatbot for social isolation. *Frontiers in Digital Health*, 3.

Du, J., Li, W., He, Y., Xu, R., Bing, L., and Wang, X. (2018). Variational autoregressive decoder for neural response generation. In *Proceedings of EMNLP 2018*, pages 3154–3163.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Fazel-Zarandi, M., Li, S., Cao, J., Casale, J., Henderson, P., Whitney, D., and Geramifard, A. (2017). Learning robust dialog policies in noisy environments. In *Proceedings of NIPS 2017 Workshop on Conversational AI: Today's Practice and Tomorrow's Potential*.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of EMNLP 2017*, pages 1615–1625.

Filannino, M. and Di Bari, M. (2015). Gold standard vs. silver standard: the case of dependency parsing for Italian. *CLiC it*, page 141.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 30(1):71–76.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Ganaie, M. and Mudasir, H. (2015). A study of social intelligence & academic achievement of college students of district Srinagar, J&K, India. *Journal of American Science*, 11(3):23–27.

Ghazvininejad, M., Brockett, C., Chang, M., Dolan, B., Gao, J., Yih, W., and Galley, M. (2018). A knowledge-grounded neural conversation model. In *Proceedings of AAAI 2018*, pages 5110–5117.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. F. (2019). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 154–164.

Ghosh, S., Chollet, M., Laksana, E., Morency, L., and Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of ACL 2017*, pages 634–642.

Giannakopoulos, T., Pikrakis, A., and Theodoridis, S. (2009). A dimensional approach to emotion recognition of speech from movies. In *Proceedings of ICASSP 2009*, pages 65–68.

Gu, J., Li, T., Liu, Q., Ling, Z., Su, Z., Wei, S., and Zhu, X. (2020). Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of CIKM 2020*, pages 2041–2044.

Hall, J. A. and Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology*, 159(3):225–243.

Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., and Zimmermann, R. (2018a). ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of EMNLP 2018*, pages 2594–2604.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L., and Zimmermann, R. (2018b). Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of NAACL-HLT 2018*, pages 2122–2132.

Hedayatnia, B., Gopalakrishnan, K., Kim, S., Liu, Y., Eric, M., and Hakkani-Tür, D. (2020). Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of INLG 2020*, pages 412–421.

Henderson, M., Vulic, I., Gerz, D., Casanueva, I., Budzianowski, P., Coope, S., Spithourakis, G., Wen, T., Mrksic, N., and Su, P. (2019). Training neural response selection for task-oriented dialogue systems. In *Proceedings of ACL 2019*, pages 5392–5404.

Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (GELUs). *CoRR*, abs/1606.08415.

Herzig, J., Feigenblat, G., Shmueli-Scheuer, M., Konopnicki, D., Rafaeli, A., Altman, D., and Spivak, D. (2016). Classifying emotions in customer support dialogues in social media. In *Proceedings of SIGDIAL 2016*, pages 64–73.

Howell, D. C. (2016). *Fundamental Statistics for the Behavioral Sciences*. Nelson Education.

Hripcsak, G. and Heitjan, D. F. (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110.

Hsu, C., Chen, S., Kuo, C., Huang, T. K., and Ku, L. (2018). EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of LREC 2018*.

Hu, B., Lu, Z., Li, H., and Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Proceedings of NIPS 2014*, pages 2042–2050.

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., and Akkiraju, R. (2018). Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of CHI 2018*, page 415.

Huang, C., Zaïane, O. R., Trabelsi, A., and Dziri, N. (2018). Automatic dialogue generation with expressed emotions. In *Proceedings of NAACL-HLT 2018*, pages 49–54.

Huang, M., Zhu, X., and Gao, J. (2020). Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3):21:1–21:32.

Humeau, S., Shuster, K., Lachaux, M., and Weston, J. (2020). Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of ICLR 2020*.

Izard, C. E., Libero, D. Z., Putnam, P., and Haynes, O. M. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology*, 64(5):847.

Jiang, Q., Zhang, Y., and Pian, W. (2022). Chatbot as an emergency exist: Mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. *Information Processing & Management*, page 103074.

Ju, D., Xu, J., Boureau, Y., and Weston, J. (2022). Learning from data in the mixed adversarial non-adversarial case: Finding the helpers and ignoring the trolls. *CoRR*, abs/2208.03295.

Kayhani, A. K., Meziane, F., and Chiky, R. (2020). Movies emotional analysis using textual contents. In *Proceedings of NLDB 2020*, volume 12089 of *Lecture Notes in Computer Science*, pages 205–212.

Ke, P., Guan, J., Huang, M., and Zhu, X. (2018). Generating informative responses with controlled sentence function. In *Proceedings of ACL 2018*, pages 1499–1508.

Khosla, S. (2018). EmotionX-AR: CNN-DCNN autoencoder based emotion classifier. In Ku, L. and Li, C., editors, *Proceedings of SocialNLP@ACL 2018*, pages 37–44.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Klein, J., Moon, Y., and Picard, R. W. (2001). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2):119–140.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lanzoni, S. (2018). *Empathy: A history*. Yale University Press.

Lee, H., Ho, C., Lin, C., Chang, C., Lee, C., Wang, Y., Hsu, T., and Chen, K. (2020). Investigation of sentiment controllable chatbot. *CoRR*, abs/2007.07196.

Lefcourt, H. M. and Martin, R. A. (2012). *Humor and life stress: Antidote to adversity*. Springer Science & Business Media.

Lenat, D. B. and Guha, R. V. (1989). *Building large knowledge-based systems: Representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016a). A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*, pages 110–119.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, W. B. (2016b). A persona-based neural conversation model. In *Proceedings of ACL 2016*.

Li, J., Monroe, W., and Jurafsky, D. (2016c). A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., and Gao, J. (2016d). Deep reinforcement learning for dialogue generation. In *Proceedings of EMNLP 2016*, pages 1192–1202.

Li, J. and Sun, X. (2018). A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. In *Proceedings of EMNLP 2018*, pages 678–683.

Li, Q., Chen, H., Ren, Z., Ren, P., Tu, Z., and Chen, Z. (2020). EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of COLING 2020*, pages 4454–4466.

Li, Q., Li, P., Ren, Z., Ren, P., and Chen, Z. (2022a). Knowledge bridging for empathetic dialogue generation. In *Proceedings of AAAI 2022*, pages 10993–11001.

Li, Y., Peng, B., Shen, Y., Mao, Y., Liden, L., Yu, Z., and Gao, J. (2022b). Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of NAACL 2022*, pages 206–218.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP 2017*, pages 986–995.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out@ACL 2004*, pages 74–81.

Lin, Z., Madotto, A., Bang, Y., and Fung, P. (2021). The Adapter-Bot: All-in-one controllable conversational model. In *Proceedings of AAAI 2021*, pages 16081–16083.

Lin, Z., Madotto, A., Shin, J., Xu, P., and Fung, P. (2019). MoEL: Mixture of empathetic listeners. In *Proceedings of EMNLP-IJCNLP 2019*, pages 121–132.

Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., and Fung, P. (2020). CAiRE: An end-to-end empathetic chatbot. In *Proceedings of AAAI 2020*, pages 13622–13623.

Lison, P. and Meena, R. (2016). Automatic turn segmentation for movie & TV subtitles. In *2016 IEEE Spoken Language Technology Workshop, SLT 2016*, pages 245–252.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of LREC 2016*.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of LREC 2018*.

Liu, C., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP 2016*, pages 2122–2132.

Liu, K. and Picard, R. W. (2005). Embedded empathy in continuous, interactive health assessment. In *CHI Workshop on HCI Challenges in Health Assessment*, volume 1, page 3. Citeseer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Liu, Z., Patwary, M., Prenger, R., Prabhumoye, S., Ping, W., Shoeybi, M., and Catanzaro, B. (2022). Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (2017). Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings ACL 2017*, pages 1116–1126.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of SIGDIAL 2015*, pages 285–294.

Lubis, N., Sakti, S., Yoshino, K., and Nakamura, S. (2018). Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of AAAI 2018*, pages 5293–5300.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A. F., Mihalcea, R., and Poria, S. (2020). MIME: MIMicking emotions for empathetic response generation. In *Proceedings of EMNLP 2020*, pages 8968–8979.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A. F., and Cambria, E. (2019). DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of AAAI 2019*, pages 6818–6825.

Mehrabian, A. (1980). Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies.

Mehrabian, A. (2017). *Nonverbal communication*. Routledge.

Mehrabian, A. and Epstein, N. (1972). A measure of emotional empathy. *Journal of personality*.

Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., and Geist, M. (2020). Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR 2013 Workshop*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. Number 47. University of Illinois press.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pernet, C. (2022). New phishing technique lures users with fake chatbot. *TechRepublic*.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to Emotion*, 1984(197-219):2–4.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL 2019*, pages 527–536.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734.

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of IJCAI 2018*, pages 4279–4285.

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

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

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL 2019*, pages 5370–5381.

Reeves, B. and Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3980–3990.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of EMNLP 2011*, pages 583–593.

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., and Convit, A. (2007). Who cares? revisiting empathy in Asperger syndrome. *Journal of Autism and Developmental Disorders*, 37(4):709–715.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y., and Weston, J. (2021). Recipes for building an open-domain chatbot. In *Proceedings of EACL 2021*, pages 300–325.

Roth-Hanania, R., Davidov, M., and Zahn-Waxler, C. (2011). Empathy development from 8 to 16 months: Early signs of concern for others. *Infant Behavior and Development*, 34(3):447–458.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805.

Sabour, S., Zheng, C., and Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. In *Proceedings of AAAI 2022*, pages 11229–11237.

Santhanam, S., Cheng, Z., Mather, B., Dorr, B. J., Bhatia, A., Hebenstreit, B., Zemel, A., Dalton, A., Strzalkowski, T., and Shaikh, S. (2020). Learning to plan and realize separately for open-ended dialogue systems. In *Findings of ACL: EMNLP 2020*, pages 2736–2750.

Sap, M., Bras, R. L., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2019). ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of AAAI 2019*, pages 3027–3035.

Scherer, K. R. (1987). Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communication*, 1:1–98.

Scherer, K. R. (2000). Psychological models of emotion. *The Neuropsychology of Emotion*, 137(3):137–162.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal Processes in Emotion: Theory, Methods, Research*, 92(120):57.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.

See, A., Roller, S., Kiela, D., and Weston, J. (2019). What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of NAACL-HLT 2019*, pages 1702–1723.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI 2016*, pages 3776–3784.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A. C., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of AAAI 2017*, pages 3295–3301.

Shamay-Tsoory, S. G., Aharon-Peretz, J., and Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, 132(3):617–627.

Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. In *Proceedings of ACL-IJCNLP 2015*, pages 1577–1586.

Shen, L. and Feng, Y. (2020). CDL: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of ACL 2020*, pages 556–566.

Shin, J., Xu, P., Madotto, A., and Fung, P. (2020). Generating empathetic responses by looking ahead the user's sentiment. In *Proceedings of ICASSP 2020*, pages 7989–7993.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., Behrooz, M., Ngan, W., Poff, S., Goyal, N., Szlam, A., Boureau, Y., Kambadur, M., and Weston, J. (2022). BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage. *CoRR*, abs/2208.03188.

Skerry, A. E. and Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology*, 25(15):1945–1954.

Smith, E. M., Williamson, M., Shuster, K., Weston, J., and Boureau, Y. (2020). Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of ACL 2020*, pages 2021–2030.

Song, Z., Zheng, X., Liu, L., Xu, M., and Huang, X. (2019). Generating responses with a specific emotion in dialog. In *Proceedings of ACL 2019*, pages 3685–3695.

Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Simonsen, J. G., and Nie, J. (2015a). A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of CIKM 2015*, pages 553–562.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., and Dolan, B. (2015b). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT 2015*, pages 196–205.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI 2017*, pages 4444–4451.

Sun, K., Moon, S., Crook, P. A., Roller, S., Silvert, B., Liu, B., Wang, Z., Liu, H., Cho, E., and Cardie, C. (2021). Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of NAACL-HLT 2021*, pages 1570–1583.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of NIPS 2014*, pages 3104–3112.

Tandon, N., de Melo, G., and Weikum, G. (2017). WebChild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017*, pages 115–120.

Tao, C., Mou, L., Zhao, D., and Yan, R. (2018). RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of AAAI 2018*, pages 722–729.

Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., and Yan, R. (2019). One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of ACL 2019*, pages 1–11.

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhou, Y., Chang, C., Krivokon, I., Rusch, W., Pickett, M., Meier-Hellstern, K. S., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E. H., and Le, Q. (2022). LaMDA: Language models for dialog applications. *CoRR*, abs/2201.08239.

Tillmann, C. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.

Tinsley, H. E. and Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22(4):358.

Triguero, I., García, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowl. Inf. Syst.*, 42(2):245–284.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX(236):433–460.

Valente, F. (2016). Empathy and communication: A model of empathy development. *Journal of New Media and Mass Communication*, 3(1):1–24.

Vardoulakis, L. P., Ring, L., Barry, B., Sidner, C. L., and Bickmore, T. W. (2012). Designing relational agents as long term social companions for older adults. In *Proceedings of IVA 2012*, volume 7502 of *Lecture Notes in Computer Science*, pages 289–302.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of NIPS 2017*, pages 5998–6008.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., and Batra, D. (2018). Diverse beam search for improved description of complex scenes. In *Proceedings of AAAI 2018*, pages 7371–7379.

Vinyals, O. and Le, Q. V. (2015). A neural conversational model. *CoRR*, abs/1506.05869.

Wallace, R. S. (2009). The anatomy of A.L.I.C.E. In *Parsing the Turing Test*, pages 181–210. Springer.

Wanner, L., André, E., Blat, J., Dasiopoulou, S., Farrús, M., Fraga, T., Kamateri, E., Lingenfelser, F., Llorach, G., Martínez, O., Meditskos, G., Mille, S., Minker, W., Pragst, L., Schiller, D., Stam, A., Stellingwerff, L., Sukno, F., Vieru, B., and Vrochidis, S. (2017). Design of a knowledge-based agent as a social companion. *Procedia Computer Science*, 121:920–926.

Ward, W. H. and Issar, S. (1994). Recent improvements in the CMU spoken language understanding system. In *Proceedings of HLT 1994*.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Wei, W., Liu, J., Mao, X., Guo, G., Zhu, F., Zhou, P., and Hu, Y. (2019). Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of CIKM 2019*, pages 1401–1410.

Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Welivita, A. and Pu, P. (2020). A taxonomy of empathetic response intents in human social conversations. In *Proceedings of COLING 2020*, pages 4886–4899.

Welivita, A., Xie, Y., and Pu, P. (2021). A large-scale dataset for empathetic response generation. In *Proceedings of EMNLP 2021*, pages 1251–1264.

Weston, J., Dinan, E., and Miller, A. H. (2018). Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of SCAI@EMNLP 2018*, pages 87–92.

Williams, J. D., Raux, A., and Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue Discourse*, 7(3):4–33.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019). TransferTransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Wu, Y., Wu, W., Xing, C., Zhou, M., and Li, Z. (2017). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of ACL 2017*, pages 496–505.

Wundt, W. M. and Judd, C. H. (1902). *Outlines of psychology*. W. Engelmann.

Xie, Y., Li, J., and Pu, P. (2021). Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. In *Proceedings of ACL/IJCNLP 2021, (Volume 2: Short Papers)*, pages 33–39.

Xie, Y. and Pu, P. (2021). Empathetic dialog generation with fine-grained intents. In *Proceedings of CoNLL 2021*, pages 133–147.

Xie, Y., Svikhnushina, E., and Pu, P. (2020). A multi-turn emotionally engaging dialog model. In *Joint Proceedings of the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents co-located with IUI 2020*, volume 2848 of *CEUR Workshop Proceedings*.

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W. (2017). Topic aware neural response generation. In *Proceedings of AAAI 2017*, pages 3351–3357.

Xing, C., Wu, Y., Wu, W., Huang, Y., and Zhou, M. (2018). Hierarchical recurrent attention network for response generation. In *Proceedings of AAAI 2018*, pages 5610–5617.

Xu, A., Liu, Z., Guo, Y., Sinha, V., and Akkiraju, R. (2017). A new chatbot for customer service on social media. In *Proceedings of CHI 2017*, pages 3506–3510.

Xu, C., Wu, W., and Wu, Y. (2018). Towards explainable and controllable open domain dialogue generation with dialogue acts. *CoRR*, abs/1807.07255.

Xu, J., Ung, M., Komeili, M., Arora, K., Boureau, Y., and Weston, J. (2022). Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. *CoRR*, abs/2208.03270.

Xu, R., Tao, C., Jiang, D., Zhao, X., Zhao, D., and Yan, R. (2021). Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of AAAI 2021*, pages 14158–14166.

Yang, L., Hu, J., Qiu, M., Qu, C., Gao, J., Croft, W. B., Liu, X., Shen, Y., and Liu, J. (2019). A hybrid retrieval-generation neural conversation model. In *Proceedings of CIKM 2019*, pages 1341–1350.

Yeh, S., Lin, Y., and Lee, C. (2019). An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *Proceedings of ICASSP 2019*, pages 6685–6689.

Young, T., Xing, F., Pandelea, V., Ni, J., and Cambria, E. (2022). Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of AAAI 2022*, pages 11622–11629.

Yuan, C., Zhou, W., Li, M., Lv, S., Zhu, F., Han, J., and Hu, S. (2019). Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of EMNLP-IJCNLP 2019*, pages 111–120.

Zamora, J. (2017). I'm sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations. In *Proceedings of HAI 2017*, pages 253–260.

Zhang, H., Liu, X., Pan, H., Song, Y., and Leung, C. W. (2020a). ASER: A large-scale eventuality knowledge graph. In *Proceedings of WWW 2020*, pages 201–211.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL 2018*, pages 2204–2213.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Zhang, W., Zhu, Q., Wang, Y., Zhao, Y., and Liu, T. (2019a). Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446.

Zhang, Y., Gao, X., Lee, S., Brockett, C., Galley, M., Gao, J., and Dolan, B. (2019b). Consistent dialogue generation with self-supervised feature learning. *CoRR*, abs/1903.05759.

Zhang, Y., Sun, S., Galley, M., Chen, Y., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020b). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of ACL 2020: System Demonstrations*, pages 270–278.

Zhang, Z., Takanobu, R., Zhu, Q., Huang, M., and Zhu, X. (2020c). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Zhao, T., Lu, A., Lee, K., and Eskénazi, M. (2017a). Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. In *Proceedings of SIGDIAL 2017*, pages 27–36.

Zhao, T., Zhao, R., and Eskénazi, M. (2017b). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of ACL 2017*, pages 654–664.

Zhao, X., Wu, W., Xu, C., Tao, C., Zhao, D., and Yan, R. (2020). Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of EMNLP 2020*, pages 3377–3390.

Zheng, C. and Huang, M. (2021). Exploring prompt-based few-shot learning for grounded dialog generation. *CoRR*, abs/2109.06513.

Zheng, C., Liu, Y., Chen, W., Leng, Y., and Huang, M. (2021). CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824.

Zhong, P., Wang, D., and Miao, C. (2019). An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of AAAI 2019*, pages 7492–7500.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018a). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of AAAI 2018*.

Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018b). Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of IJCAI 2018*, pages 4623–4629.

Zhou, L., Gao, J., Li, D., and Shum, H. (2020). The design and implementation of XiaoIce, an empathetic social chatbot. *Comput. Linguistics*, 46(1):53–93.

Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W. X., Yu, D., and Wu, H. (2018c). Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of ACL 2018*, pages 1118–1127.

Zhou, X. and Wang, W. Y. (2018). MojiTalk: Generating emotional responses at scale. In *Proceedings of ACL 2018*, pages 1128–1137.

Yubo XIE

yuboxie@hotmail.com +41 76 638 41 36 Aug 14, 1993
yuboxie.github.io github.com/yuboxie linkedin.com/in/yuboxie
Route Cantonale 37, 1025 St-Sulpice VD, Switzerland

RESEARCH INTERESTS

Human-Computer Interaction · Natural Language Processing · Dialog Systems

EDUCATION

École Polytechnique Fédérale de Lausanne *Sept 2017 – Nov 2022*
Ph.D. in Computer and Communication Sciences

University of California, San Diego *Sept 2015 – Dec 2016*
M.S. in Computer Science

Shanghai Jiao Tong University *Sept 2011 – Jul 2015*
B.E. in Computer Science and Technology (IEEE Honor Class)

EXPERIENCE

Human-Computer Interaction Group *École Polytechnique Fédérale de Lausanne*
Research Assistant *Sept 2017 – Present*

- Supervisor: Dr. Pearl Pu
- Research areas: dialog systems, affective computing, computational humor
- Emotion recognition in tweets using convolutional neural networks with attention
- Empathetic response generation and its evaluation on crowdsourcing platform
- Verbal humor recognition using large-scale pre-trained language models

Apple Inc. *Lausanne, Switzerland*
Siri Prototyping Intern *Jul 2021 – Dec 2021*

- Mentor: Dr. Didier Guzzoni
- Explore ways to improve Siri experience by making various prototypes
- Help with the development of internal tools

Learning and Optimization Group *Shanghai Jiao Tong University*
Undergraduate Research Assistant *Sept 2013 – Jul 2015*

- Supervisor: Prof. Zhihua Zhang
- Research areas: approximate inference, topic modeling
- Topic modeling using latent Dirichlet allocation
- Electroencephalogram emotion analysis using support matrix machines

PUBLICATIONS

- **Yubo Xie**, Junze Li, and Pearl Pu. AFEC: A Knowledge Graph Capturing Social Intelligence in Casual Conversations. *arXiv preprint arXiv:2205.10850*.

· **Yubo Xie** and Pearl Pu. Empathetic Dialog Generation with Fine-Grained Intents. *CoNLL 2021*.

· Anuradha Welivita, **Yubo Xie**, and Pearl Pu. A Large-Scale Dataset for Empathetic Response Generation. *EMNLP 2021*.

· **Yubo Xie** and Pearl Pu. How Commonsense Knowledge Helps with Natural Language Tasks: A Survey of Recent Resources and Methodologies. *arXiv preprint arXiv:2108.04674*.

· **Yubo Xie**, Junze Li, and Pearl Pu. HumorHunter at SemEval-2021 Task 7: Humor and Offense Recognition with Disentangled Attention. *SemEval 2021*.

· **Yubo Xie**, Junze Li, and Pearl Pu. Uncertainty and Surprisal Jointly Deliver the Punchline: Exploiting Incongruity-Based Features for Humor Recognition. *ACL-IJCNLP 2021*.

· **Yubo Xie**, Ekaterina Svikhnushina, and Pearl Pu. A Multi-Turn Emotionally Engaging Dialog Model. *IUI 2020: Workshop on User-Aware Conversational Agents*.

· Luo Luo, **Yubo Xie**, Zhihua Zhang, and Wu-Jun Li. Support Matrix Machines. *ICML 2015*.

ACADEMIC ACTIVITIES

· Reviewer of EMNLP 2021/2022

· Reviewer of IEEE Transactions on Audio, Speech and Language Processing 2022

· Oral presentation at ACL-IJCNLP 2021, "Uncertainty and Surprisal Jointly Deliver the Punchline: Exploiting Incongruity-Based Features for Humor Recognition" (video link)

· Tutorial at LauzHack 2020, "Learn How to Prototype Interactive Apps Using FluidUI" (video link)

· Seminar talk at EPFL with Google (2019), "A Multi-Turn Emotionally Engaging Dialog Model"

TEACHING ASSISTANTSHIP

CS-486 Interaction Design, EPFL	*Spring 2021*
CS-431 Introduction to Natural Language Processing, EPFL	*Fall 2020*
CS-486 Human Computer Interaction, EPFL	*Spring 2018 / 2019 / 2020*
CS-433 Machine Learning, EPFL	*Fall 2018 / 2019*

AWARDS

Meritorious Winner of Mathematical Contest in Modeling	*2014*
Academic Excellence Scholarship of Shanghai Jiao Tong University	*2012, 2013, 2014*

SKILLS

Languages	Chinese (native), English (working proficiency)
Programming	Python, HTML/CSS, JavaScript, LaTeX
Packages & Tools	TensorFlow, PyTorch, Pandas, NumPy, Flask