Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Electronic & Electrical Engineering

Emotional Dialogue System in Open-domain Human-Machine Interactions

Shiqi Shen
Supervisor: Prof. Vincent Vade

April 24, 2022

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: Shiqi Shen Date: 4/24/2022

Contents

List of Figures

List of Tables

1 Introduction

Increasing importance to allow natural interface between humans and computers is shown in the use of Intelligent agents such as Alexa, OkGoogle, as well as embedded chat bots in different customer service websites. In the recent two years, COVID 19 pandemic shook the world in every aspect of it. However, it largely accelerates the process of dialogue generation technology adoption as the need for online communication grows. Conversation automation is booming more than ever and according to a survey from Landbot [1]: from 2018 to 2020, the total usage of online chat bots has increased by 67%.

A particular challenge in intelligent agent of chat bots is the problems around natural language understanding and natural language generation. Dialogue systems focus on two broad categories: task-oriented systems where the system is designed to solve a specific problem for users belonging to a concrete domain; open-domain conversations with casual chit-chat. For a long time, dialogue applications are based on simple rule-based rigid agents that can only react to limited instructions and dialogue templates. With the recent advances in deep learning techniques and especially in the field of natural language understanding, they have evolved greatly to become much more sophisticated intelligent systems that can converse across rich topics and generate human-like utterances.

The implementation of intelligent chat bots or dialogue systems is of great importance to both industrial and academic communities. For example, a well-designed conversational agent enables enterprises to offer automatic customer service interaction and responses, and thus significantly reduces human labour resources. For academia, it is appealing yet challenging to build an intelligent agent where the fundamental need is to capture the semantics of the dialogue context and use it for selecting the appropriate response. Such processes involve a series of high-level natural

[1] https://landbot.io/blog/chatbot-statistics-compilation

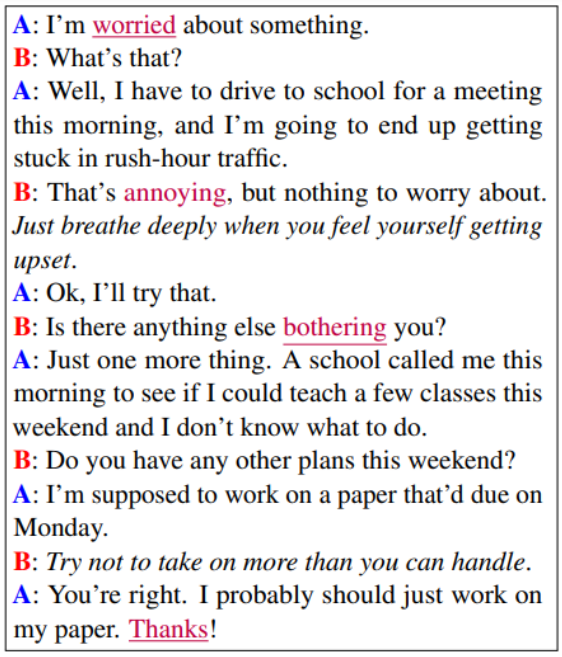


Figure 1.1: An example of a bi-turn dialogue with speakers emotions labelled in purple.

language processing techniques, such as understanding the underlying semantics of user input utterances and generating coherent and relevant responses.

Most existing dialogue data sets like (3, 4, 5, 6, 7) follow a natural bi-turn flow where the speakers communicate iteratively. One example of them is presented in Figure 1.1. Occasionally, in addition to the content of the dialogue, there are contextual resources available such as dialogue act (intention of the utterance like "inform" or "question"), speakers' personalities or conversation sentiment. These complementary context information could provide essential clues as to what could be a follow-up response. In this project, we aim to explore the potential of this auxiliary information and how to leverage them in the improvement of dialogue systems.

1.1 Motivation

One of the common objectives of dialogue systems is to achieve human-like communication between humans and the dialogue agent. Despite the rapid development of deep learning technologies, human-machine communications are still not comparable to human-human dialogues. One particular reason underlying this gap is that human-human interaction usually involves exchanging not only explicit linguistic information, but also implicit emotional states. The analysis of such

emotional contexts can cover the areas of emotion recognition, understanding, and most importantly in dialogue systems, generation.

In real-life conversations, the ability to read and express specific sentiments and emotions often leads to effective communication. For dialogue systems, such ability is also desirable and can strengthen the communication in a positive direction (8). There is a growing need for dialogue systems to perceive and understand users' emotions correctly, hence making the generated responses more accurate. In fact, dialogue systems that are able to accurately recognize and convey emotions can communicate with the users at the human level. This feature can further enhance user satisfaction (9) and lead to coherent and longer conversations (10).

These findings motivate us in integrating emotional contexts into open-domain dialogue systems, hoping to increase the level of user engagement and satisfaction.

1.2 Research Questions and Objectives

The overall objective of this research project is to investigate the effectiveness of integrating auxiliary emotional information in dialogue systems. Towards the achievement of this objective, we break the overall goal into several research questions:

RQ1: How can we make the generated response express an assigned emotion? A particular aspect that we care about when building an emotional dialogue system is the ability to express desired emotions. Therefore, we aim to explore the possibility of controlling the dialogue agent with specific emotions.

RQ2: To what extent does integrating emotional responses make an effect on automated metrics? The ability to recognize user emotions also contributes to the process of natural language understanding. For this research question, we will examine whether the integration of emotional contexts improves model performance during automatic evaluation.

RQ3: Does emotional content make the conversation more engaging? In the last research question, we propose to investigate whether the integration of emotional contexts improves overall user experience.

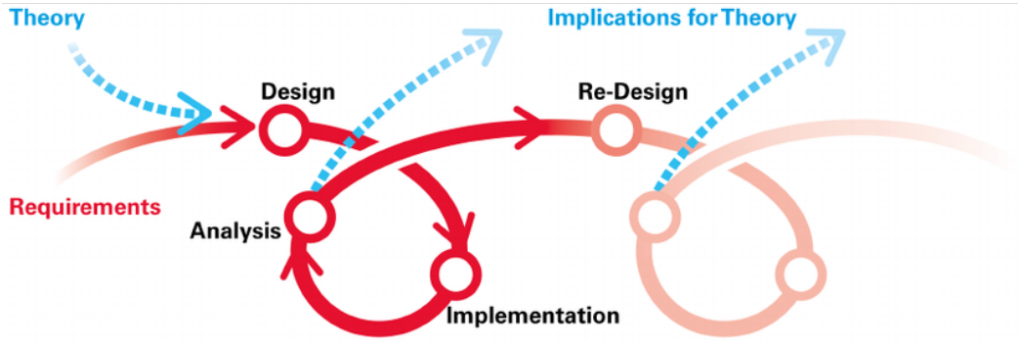


Figure 1.2: The workflow of a typical design-based research project.

1.3 Research Methodology

Design-based research is a kind of research methodology used by researchers in education. The basic process of Design-based research involves analyzing the theory and requirements of a particular problem and then developing an innovative solution. The innovation is so-called "interventions", to problems. The innovative solutions are rigorously evaluated (against requirements & where applicable end-users) and then implications are assessed. The evaluations and improvements are iterated in a design cycle. The purpose of this research approach is to inspire new theoretical and applicable methods. For example in educational uses of Design-based research, the methodology has been used to inform conceptualizing learning, and instruction (11). Design-based research methodology can improve designs for ICT based programs as well as the understanding of application-specific logic. Researchers first evaluate the designs in authentic settings and collect information about the effects of the design/innovation. Then the researcher(s) can analyze the results of the investigation, reformulates designs/innovations, and finally refactor the theoretical framework that guided the research (see Figure 1.2).

Design-based research characterizes an ongoing process of innovation. There is always multiple research cycle existing in one research project. Therefore, the research process is always stimulated by the intended improvement in the design. Compared with other design methods and simple evaluation in traditional predictive scientific research, design-based research is characterized by the following specific benefits:

- With innovation at the cutting edge of theoretical knowledge.
- Can actively customize the design/re-design procedure within a research cycle.

- Results are continuously analyzed in quantitative and qualitative aspects which allows for the prompt adjustments being adapted to the later implementation.
- The concept of the research design and the theoretical foundations are continuously implemented and revised.

1.4 Contribution

Our main contributions are list as follows:

- We successfully implemented a Transformer-based emotional dialogue system which makes use of emotional tags and previous dialogue history to generate more emotion conveying responses.
- The experimental results show that emotional contexts affect the behaviour of the dialogue system to some extent.

2 Literature Review and state-of-the-art

Dialogue generation is an increasingly important task in the area of NLP and has received considerable critical attention across a significant number of researchers in recent years. In this section we present the techniques that are related to our research objective. We start from an overview of three different types of existing technologies for building dialogue systems. Then, we look at state-of-the-art in emotional dialogue systems. For the training of dialogue systems, we consider a number of dialogue datasets and report their sizes and dimensions. In the end of this section, we investigate evaluation metrics of the dialogue response generation task.

2.1 Dialogue System Architectures

The field of dialogue response generation is considered one of the most challenging tasks in Natural Language Processing (NLP) and has been attracting enormous attention of NLP researchers since the 1960s. With a general goal of producing meaningful and fluent natural language from various dialogue context, the technologies applied in building dialogue systems are constantly evolving.

Rule-based Systems Many of the early successes in the field of Dialogue Response Generation were largely reliant on rule-based systems. ELIZA was developed by Weizenbaum (12) in 1966, which is a conversational program and was created to demonstrate the superficiality of communication between humans and machines. This system enables human-computer conversation by applying pattern matching and substitution methodologies. In 1975, Colby (13) proposed PARRY, this dialogue system implements a rule-based model to simulate the behaviour of a paranoid schizophrenic. It was considered as a much more advanced program than ELIZA

because it can generate utterances with evident and strong human emotions, a conversation between these two early dialogue systems is presented in Figure 2.1. Where ELIZA tend to generate neutral and general content and keep asking random questions. As a contrast, PARRY converses with strong negative emotions and behave like an aggressive and paranoid person. These earliest attempts are able to generate human-like responses, and rule-based dialogue systems are still widely used in practical applications (14). The main advantages of rule-based dialogue systems is that, the artificial output produced from rule-based approaches is safe and controlled. Because each development process of a rule-based dialogue system is fully under control, thus the system is eventually highly controllable and explainable. This is especially important in building commercial dialogue system, as an unexpectedly hazard speech will potentially sabotage the public image of the enterprise. The incident of "Tay" [1] (a social dialogue system designed by Microsoft) is a good illustration. Tay was shut down after releasing racist and sexually-charged messages on Twitter.

However, the cost of developing such systems is relatively high, as the performance of the rule-based systems is heavily reliant on manually designed templates and pre-defined rules which require substantial human resources. On the other hand, the learning capacity of rule-based systems are limited, this is mainly due to the rule-based nature of these systems as most of pre-defined templates and rules are only related to addressing task-specific scenarios. In this way, they are unable to be easily adapted/transferred in new tasks and domains. To fill the gaps, statistical methods were introduced as an alternative.

Statistical Approaches Dialogue response generation systems evolved from rule-based approaches to statistical-based approaches in order to reduce the manual effort of using hand written grammar rules and to deal with large corpora.

Langkilde (15) proposes to use statistical approaches in generating natural language. In this work, a corpus based statistical module was used to generate various representations of a sentences. These phrases are then ranked by the combination of context-independent score and context-dependent score. This method as the base for many following research in this field. This method is later improved by Bangalore and Rambow (16) with a tree-based model of syntactic representation along with independently hand-crafted grammar rules. Another work (17) presented a novel ranking algorithm, they proposed to use a log-linear ranking system and show that

[1] https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

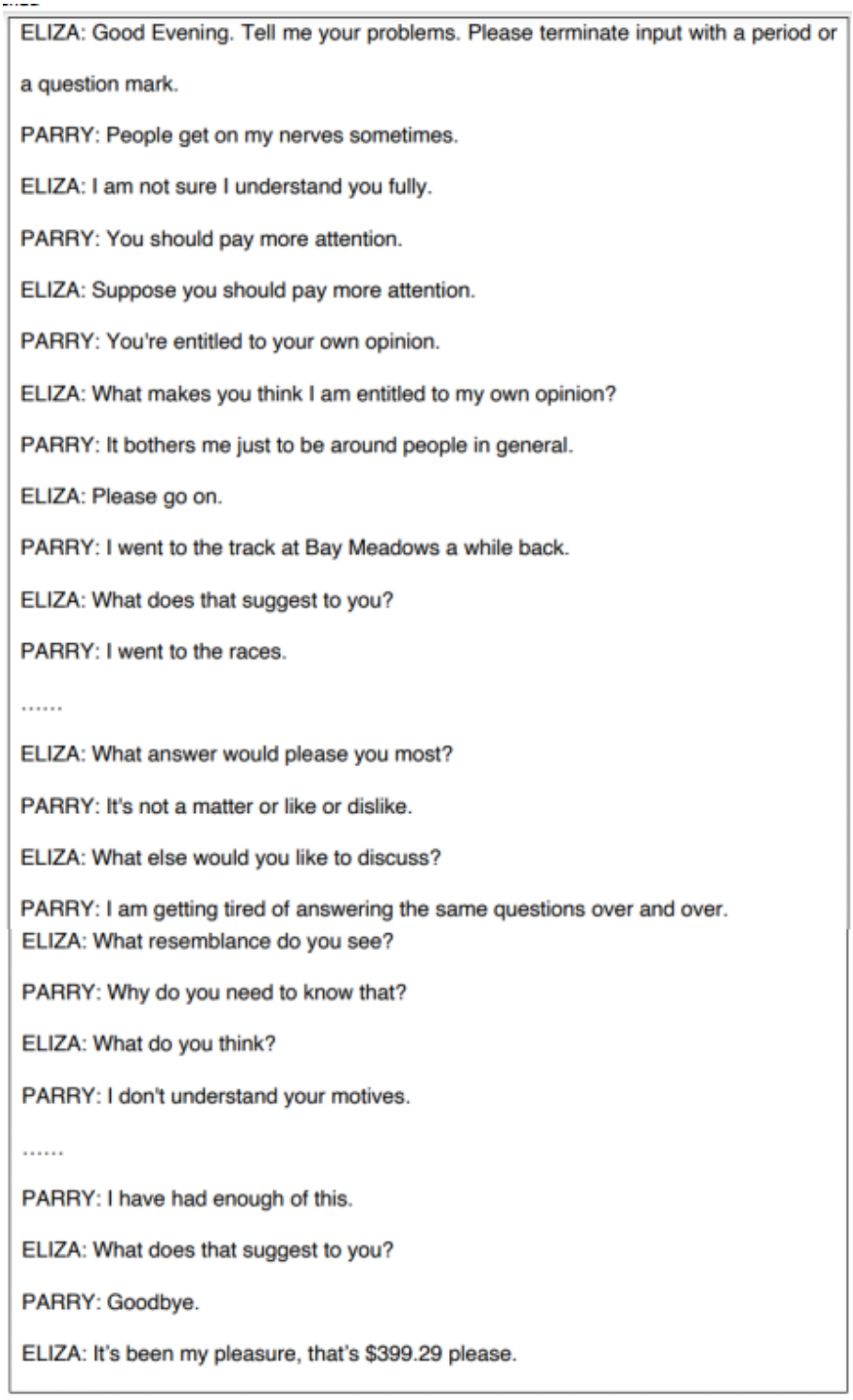


Figure 2.1: A conversation between two rule-based dialogue systems: ELIZA and PARRY.

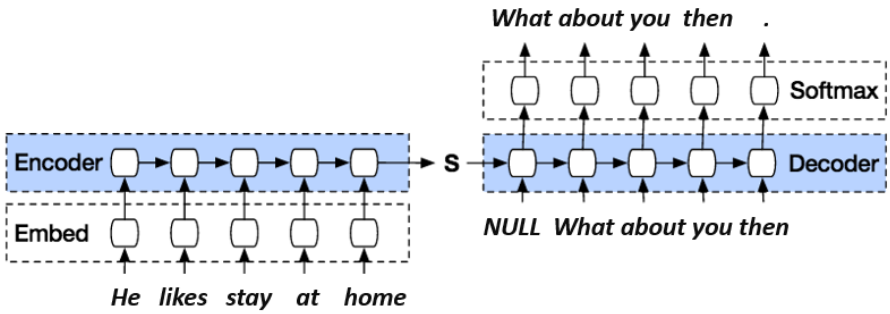


Figure 2.2: Encoder-decoder based dialogue system, the encoder aggregates the input words and generates contextual state "S", and the decoder outputs target tokens in a recursive manner.

log-linear ranking obtained better performance than existing systems. Schatzmann et al. (18) proposed a statistical model for word-level human-machine dialogue generation, the method explicitly models the context-dependent confusability of words and allows the system-specific language model and semantic decoder to be incorporated at the same time.

In recent years, deep learning-based methods have gradually taken over statistical-based approaches in order to mitigate the issues faced by statistical language models such as word alignment and phase ordering (19).

Deep Learning Approaches Leveraging deep learning approaches in Natural Language Generation (NLG) has achieved state-of-the-art performance across different tasks, including the task of dialogue generation due to the capability of neural networks to learn language representations with different levels of abstraction and granularity (20). The most commonly used neural network is the feed forward neural network or also called multi-layer perceptron (MLP). Bengio et al. (21) demonstrated the ability of feed forward neural networks on various language modeling tasks.

Another type of neural network architecture that is more suited for dealing with sequential data is the Recurrent Neural Network (RNN) architecture, RNNs have the capability to handle long sequences using the knowledge (memory) gained from previous sequence computations. Inspired by the recent achievements in the task of neural machine translation (22, 23, 24), Shang et al. (25) formulated dialogue generation task as a sequence-to-sequence learning (translating) problem and propose to use RNNs to build dialogue system. Following such setting, Vinyals and Le (26) propose to utilize an encoder-decoder architecture in dialogue generation task. A

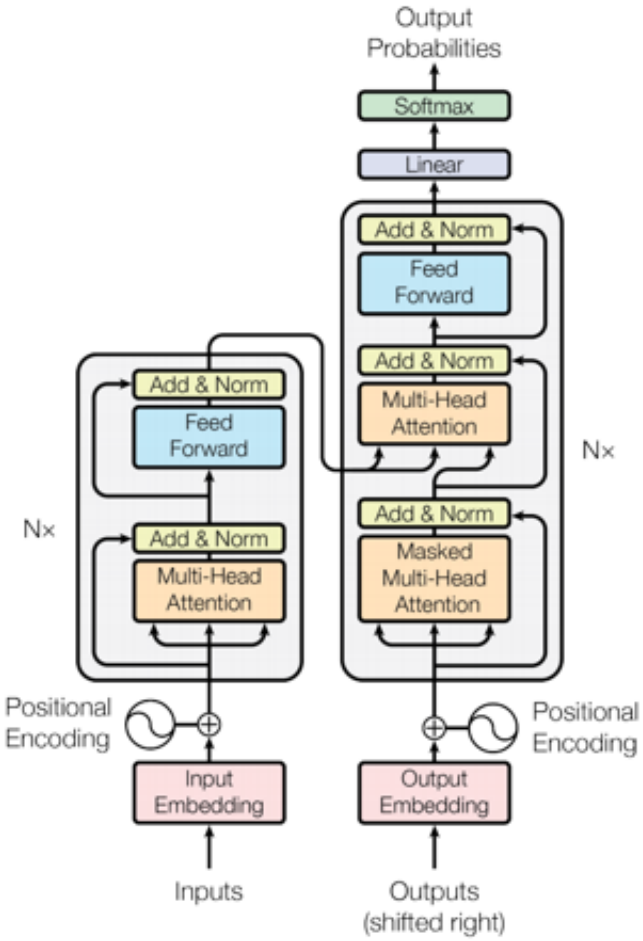


Figure 2.3: Architecture of a standard Transformer network (1).

general example is illustrated in Figure 2.2: encoder is designed to model the context of the input dialogue which is usually the sequence of speaker's utterances. A decoder is constructed for target response prediction, this process involves utilizing the contextual encoder states and decoding the target token recursively. With such grounding architecture, various backbone neural structure like Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber (27), Gated Recurrent Unit (GRU) Cho et al. (24).

Recently, the Transformer (1), a novel sequence modelling architecture that relies solely on self-attention mechanism, was proposed in solving language generation task in 2017. The core structure of a standard Transformer architecture is presented in Figure 2.3, follows the encoder-decoder structure but is built entirely on the attention mechanism. Even tough the attention mechanism is applied in the RNN-based models (23, 25) to help the network focus on the appropriate part of input when

decoding tokens. It is argued that this process ignores the internal correlation of a sentence. The Transformer architecture addresses this issue by applying a self-attention mechanism that computes the attention scores of each word pair in the current input sentence, the self-attention score is computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, V relates to the concept of query, key and value in the retrieval system. The general attention mechanism (soft attention) aims to find what elements (key) in the source input need to be paid attention to, given the target token (query). The relevance scores (value) between each key-value pair is usually computed through the dot-product operation. Differently, in the self-attention mechanism, the query and key are sourced from the same sequence (mostly the input sequence), and as a result, the self-attention mechanism enables the system to accept input embedding from the previous encoder and weighs their relevance to each other to generate the encoding output. This work also introduced the idea of Positional Encoding to make use of the order of the sequence, this is done by injecting relative or absolute position information into the word embedding.

The Transformer architecture was originally proposed to solve machine translation task, it achieved leading performance on the WMT 2014 English-to-German translation task (1). This architecture has gradually taken the place of RNN architecture in sequence modelling tasks for its outstanding performance in many other NLP tasks including dialogue generation (28).

2.2 State-of-the-art in Social Dialogue Systems with Emotion

One determining aspect for successful communication is the ability to express concrete sentiments and emotions in conversations. It is desirable for any dialogue generation systems to process the contextual emotions while generating responses. By the aware of analysing emotions in dialogue contexts and having the ability of producing utterances with correct emotions, dialogue system can provide a better user experience and therefore increase users' satisfactions.

Zhou et al. (29) proposed to generate appropriate conversational responses not only

in content (relevant and grammatical) but also in emotion (emotionally consistent). They models the high-level abstraction of emotion expressions by embedding emotion categories with an external emotion vocabulary, the emotional content is integrated in an GRU based sequence-to-sequence model (23). Their proposed model can regulate the implicit change in emotional state and express the explicit emotional expression by selecting the emotion or generic words from the emotion vocabulary at every decoding time-step. In the work of Sun et al. (30), an emotional conversation generation framework named E-SCBA was proposed that applied an asynchronous approach. In this framework, the emotional words are pre-generated and then introduced into the dialogue response generation module. While this method focus on generating emotional responses, the emotions in input sentences are not analysed. To fully leverage emotional annotations in the dataset, Zhong et al. (31) proposed a novel affective attention mechanism along with the weighted cross-entropy loss for emotional dialogue generation. Their model extends the standard sequence-to-sequence model (23) and adopts emotional annotations to embed each word with affects in both encoding and decoding modules. In addition, their model considers the effect of negators and intensifiers via an affective attention mechanism, this allows the model to emphasize attentions for affect-rich words in input sentences. Asghar et al. (32) proposed an end-to-end neural framework that captures the emotional state of the user for generating emotional responses. In their work, they adopted three ways to incorporate emotional contexts in the neural conversational model, this includes an emotional word embedding method, an emotional-based objective function that augment the standard cross-entropy loss, and a novel beam search method specifically engineered for emotional dialogue generation. They showed in the experiments that, these techniques improve the open-domain conversational systems by enabling them to produce emotionally rich responses that are more interesting and natural.

2.3 Datasets for Training Social dialogue systems

In order to enable the large-scale training required for building deep learning-based dialogue systems, many endeavours have been made in collecting and processing dialogue datasets. In this section, we investigate the existing large-scale open-domain dialogues with a special focus on the datasets with emotional annotations.

Firdaus et al. (2) released Sentiment Emotion aware Multimodal Dialogue (SMED) dataset for the task of sentiment and emotion controlled dialogue generation. The

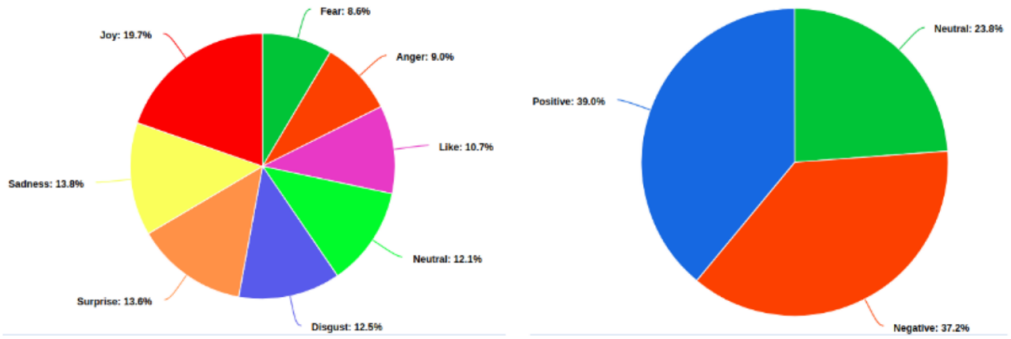


Figure 2.4: Emotion and Sentiment distribution of SEMD (2) dataset.

Name of TV Series	Genre	#Season	#Episodes	#Dialogues	#Utterances
Breaking Bad	Drama	5	62	1,659	32,653
Grey's Anatomy	Drama	15	351	14,926	295,496
House of Cards	Drama	6	73	2,851	5,616
Friends	Comedy	10	236	4,228	82,353
House M.D.	Drama	8	177	6,476	127,780
Castle	Drama	8	173	7,401	146,669
How I met your mother	Comedy	9	208	4,968	97,344
The Office	Comedy	9	201	4,813	94,470
Game of Thrones	Drama	8	73	2,263	47,472
The Big Bang Theory	Comedy	12	279	5,456	86,024
Total		90	1,833	55,041	1,066,677

Table 2.1: Statistics of SMED (2) dataset.

SEMD dataset consists of 55k conversations from 10 TV shows having text, audio and video information. This dataset is sourced from 10 famous TV series, consisting of conversations with utterances from multiple speakers, making it a multi-party conversational dataset. In each TV series, all the episodes spanning all seasons were collected. In the end, 1258 episodes (746 hours of video) were collected, the data statistics of every TV show is provided in Table 2.1. After collecting the source data, textual human dialogues are then extracted from these videos. For each episode in the dataset, all the subtitles are transcribed into text, then these speeches are further grouped into individual short conversations based on scene transition. Apart from textual data, video visual signals and audio soundtracks were also extracted and aligned with text by the corresponding video timestamps. They further annotated each utterance with an emotional tag, the emotions are selected based on Ekman's six universal emotions (Anger, Disgust, Fear, Happiness, Sadness, Surprise), they also labelled the sentiment state of each utterance. The emotion and sentiment

distribution of the entire SMED dataset is presented in Figure 2.4.

Multi-modal Emotion Lines Dataset (MELD) was released in 2019 by Poria et al. (33) and evolved from the Emotion Lines Dataset (ELD) developed by Chen et al. (34). ELD contains dialogues from famous TV-series "Friends", where each dialogue collects utterances from multiple speakers in the show. ELD was created by crawling the dialogues which were sorted into four groups of utterances respectively based on the amount of the utterances in each dialogue. The MELD was first constructed by extracting the start and end time stamps of all utterances in the ELD. To do this, the subtitles of all the episodes in 'Friends' were crawled through, and then extracted from their respective video timestamps. In particular, the following two constraints were enforced:

- Timestamps of all the utterances in a single dialogue should be sorted in an strictly-increasing order.
- All utterances in each dialogue must belong to the same scene in an episode.

These constraints filtered out a few outliers where some dialogues lasts across different scenes or episodes. Next, three annotators were employed to label each utterance. A majority voting was deployed to decide the final label of each utterance. A few utterances where all three annotations turned out to be different were dropped out, and their corresponding dialogues were also removed to maintain coherence. Finally, after clarifying the timestamp of each utterance, the corresponding audiovisual clips were extracted from the episode followed by audio content extractions. And therefore the final dataset includes information of audio, visual, and textual modalities in each utterance. As a result, MELD was characterized as an extension of ELD. MELD contains about 13,000 utterances from over 1,433 dialogues from "Friends", where each utterance is annotated with emotion labels and sentiment labels.

DailyDialog Li et al. (3) was released in 2017. This data set is formed around daily life conversations with no specific domain orientation. The raw data of this dataset was crawled from various websites which serve for English learner to practice English dialog in daily life. This dataset has three appealing characteristics: first, the language in DailyDialog is all human-written and thus is formal in terms of grammar correctness and fluency; second, the conversations in this dataset are often built around concrete daily life scenarios such as shopping in the market or two students talking about their summer vacation trips; third, the crawled dialogues usually end after reasonable speaker turns where conversations in other dialogue dataset (35) contain more than

Dataset	Source	Avg.Dialog Length	Total Utterances	Emotion
MELD	TV Series	9.6	13K	Yes
IEMOCAP	Crowdsourced	-	7K	Yes
OpenSubtitles	Movies	-	140M	No
DailyDialog	Crowdsourced	7.9	13K	Yes
Empathetic Dialogues	Crowdsourced	4.3	100K	Yes
DSTC7	Crowdsourced	20	200K	No
SEMD	TV Series	21.71	1M	Yes

Table 2.2: The statistics of seven dialogue datasets.

100 speaker turns. For the dataset annotations, they considered two aspect: emotion and act. Each utterance in the dataset is manually labelled with an emotional tag based on Ekman's six universal emotions. In addition, add one category to represent other emotions. Hence, seven emotion categories were included in DailyDialog. Apart from emotional states, communication intention was also manually labelled. This was implemented by labelling each utterance as one of four dialogs act classes: Inform, Questions, Directives and Commissive. An overview of the statistics relating to emotional tags, conversation topics and dialogue acts are presented in Figure 3.1. In total, DailyDialog contains 13,118 conversations, the mean speaker turns in a conversation is 7.9 and there are on average 14.6 words per utterance.

We also investigated other emotional dataset including IEMOCAP (36), OpenSubtitles (35), Empathetic Dialogues (37) and DSTC7 (38), the statistics of these datasets are presented in Table 2.2 for reference.

In this project, we select DailyDialog (3) in the training and experiments of dialogue systems mainly for its diversity and quality.

2.4 Evaluation Metrics for dialogue systems

Evaluation of the dialogue system has long been a challenging problem. The main difficulty comes from the fact that the quality of a dialogue response can be highly subjective and it is not possible to find out all the possible references to be compared with the model outputs. One solution is human evaluation, which provides reliable judgements in any desired criteria, however such method is time-consuming, expensive, and not necessarily repeatable Fan and Luo (39). Compared to human evaluation, automatic evaluation methods are more commonly adopted among the researchers. Typically there are two main categories for automatic metrics: learning-based metrics

and reference based metrics. Learning based metrics often require additionally training and are highly relevant to human judgment. Lowe et al. (40) learn representations of dialogue utterances using LSTM and use the dot-product between generated response and ground truth response in latent space as an evaluation score. This method is effective but requires additional development of the system. Reference-based metrics like BLEU (41), ROUGE (42) and METEOR (43) are widely used in other NLG tasks like Machine Translation and Text Summarization, which calculate the similarity between the generated utterance and the ground truth. These metrics have proved to be very effective in Machine Translation because each source sentence has a ground truth with which to compare. However in human conversations there may be many possible responses according to past dialogues, thus an acceptable response may receive a low score if simply computing reference-based metrics. It is shown in Liu et al. (44) that reference-based metrics are not correlated with human evaluation in the most cases. Some researchers (45, 46, 47) also consider the evaluation process as a retrieval task and propose to use ranking metrics for the evaluation of a dialogue system.

Back to our main objective, while the dialogue context could provide relevant signals for formulating the response, Kumar et al. (48) further prove that the additional information of dialogues can also be leveraged for response generation. In their work, they build embeddings for dialogue acts (DA) and use a separate RNN encoder to learn the DA representations, this representations are then leveraged in response generation process along with the dialogue representation. They show that dialogue acts helps achieve better performance in dialogue generation task main. They conducted a detailed analysis and drew a key insight that the inclusion of DA information induces uniformity and removes ambiguity in dialogue generation.

2.5 Summary

In general, we have reviewed the existing approaches in building dialogue systems, state-of-the-art emotional dialogue integration technologies, popular dialogue datasets and evaluation methods. These pioneer works provide us with clear instructions in both building the pipeline of a dialogue system and understanding the mechanisms behind the state-of-the-art choices.

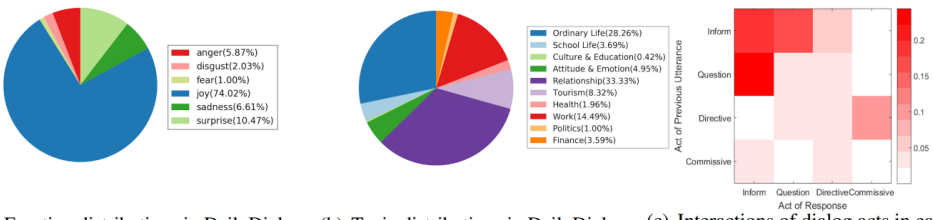
3 Design

For all experiments, we use DailyDialog (3) in the training and evaluation stages for its high quality emotional annotations and conversations. This dataset is introduced in Section 2.3.

3.1 Non-emotional Dialogue System Design

In order to evaluate the performance of emotional tags, we propose to use non-emotional dialogue system as baseline methods. An initial experiments on DailyDialog dataset has been conducted, which includes two neural architectures: RNN (sequence-to-sequence (23)) and Transformer (1). At this stage, the additional dialogue information such as emotional tags and utterance acts is not included into the model and the results thus serves as baselines.

RNN system design The sequence-to-sequence model is consist of an encoder, that analyzes input dialogue context, and a decoder, that generates output tokens. The whole system works as follows: each token in the textual input is converted into a word embedding (weights are randomly initialized and optimized during training). The encoder converts the sequence of word vectors $X = x_1, x_2, \ldots, x_n$ to hidden



(a) Emotion distributions in DailyDialog. (b) Topic distributions in DailyDialog. (c) Interactions of dialog acts in each utterance pairs.

Figure 3.1: Statistics of DailyDialog (3) dataset.

representations $h = h_1, h_2, \ldots, h_n$, this converting process is defined as:

$$h_t = GRU(h_{t-1}, x_t)$$

Then, the decoder takes as input a context vector c_t and the embedding of a previously decoded word $e(y_{t-1})$ to update its state s_t using another GRU (24) module:

$$s_t = GRU(s_{t-1}, [c_t; e(y_{t-1})])$$

where $[c_t; e(y_{t-1})]$ is the concatenation of the two vectors, serving as the input to the GRU cell. The context vector c_t is designed to dynamically attend on key information of the encoders' output h_n post during decoding (see (22) for soft attention mechanism). When a state vector s_t is computed, the decoder generates a token by sampling from the output probability distribution o_t which is computed as follows:

$$o_t = softmax(W_{out} \times s_t)$$

where W_{out} is a weight matrix that map the output state s_t from a hidden space to the task vocabulary.

Transformer system design The Transformer network (see Figure 2.3) also follows a standard encoder-decoder architecture, but with self attention operation instead of GRU. In this network, the encoder encodes the input sequence with a multi-head attention module. This encoding process contains two operations, self attention computation and a feed-forward neural network, in each operation, the input will be directed added into the result vector (residual connection) followed by a layer normalization:

$$M_{in} = MultiHead(X, X, X) \quad (1)$$
$$= LayerNorm(SelfAttention(X, X, X) + X) \quad (2)$$

In the decoder, the masked output embedding is decoded based on encoders' output M_{in} and previous decoded tokens.

$$Y_{mask} = MaskAttention(Y, Y, Y) \quad (3)$$
$$O = MultiHead(M_{in}, M_{in}, Y_{mask}) \quad (4)$$

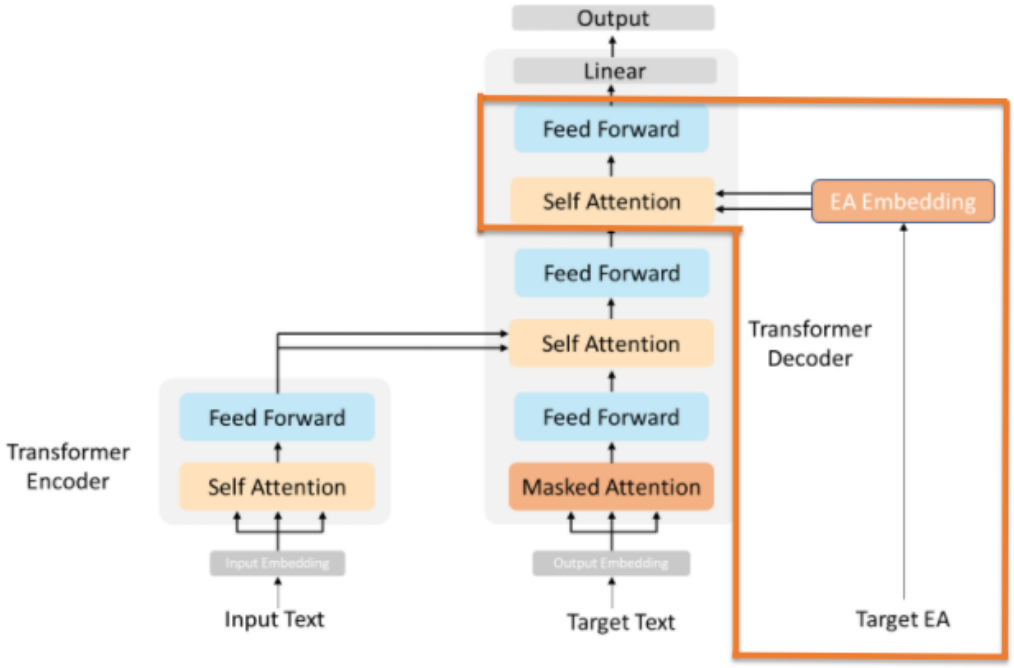


Figure 3.2: Transformer-based dialogue system architecture with dialogue emotions and acts embodied during the decoding stage.

The masking operation, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on the known outputs at positions less than i. In the decoders' multi-head attention operation, the memory state M_{in} serves as key and value in the self-attention computation. Finally, the decoder hidden state O is computed in order to generate a output token, this is implemented through an linear layer and softmax function (same as the RNN decoder).

3.2 Emotional dialogue system Design

We attempt to embed emotional tags and dialogue acts into the system. For each target response, the correspond categorical emotional tag and dialogue act are combined and converted into a single emotion&act (EA) one-hot vector. Since there are seven different emotions and four types of dialogue acts, the combined EA one-hot vector is 28-dimensional. This one-hot vector is then mapped to a low-dimensional embedding which is denoted as v_{ea}

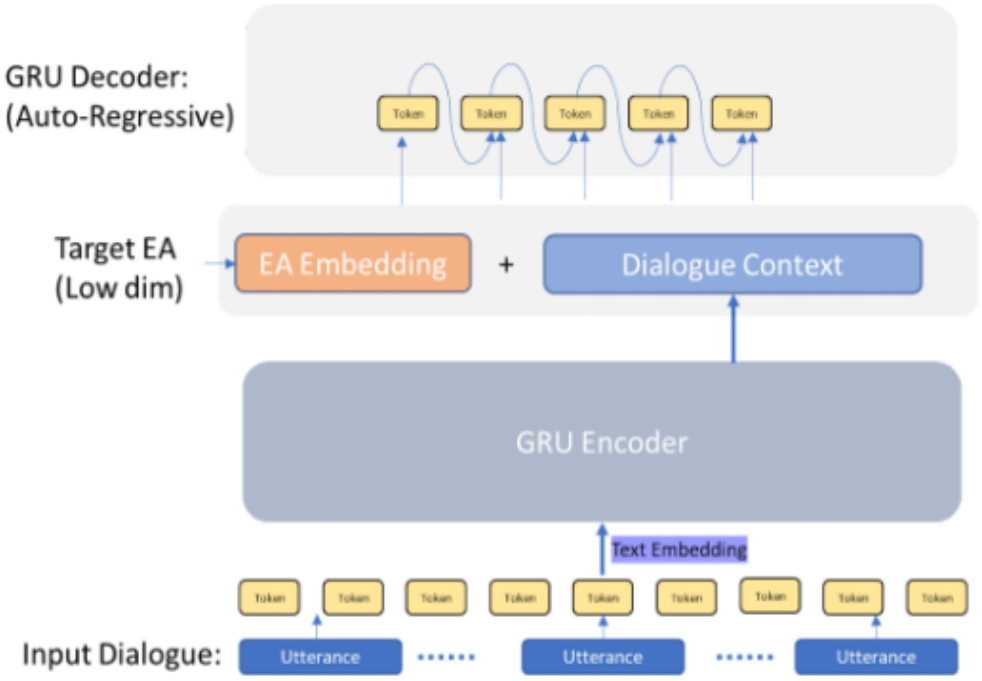


Figure 3.3: GRU-based dialogue system architecture with dialogue emotions and acts embodied during the decoding stage.

Emotional-RNN For RNN-based baseline method, we follow Zhou et al. (29) and integrate EA embedding v_{ea} into the neural network as illustrated in Figure 3.3. EA embedding is concatenated with the encoder output:

$$s_t = GRU(s_{t-1}; [[c_t; e(y_{t-1}); v_{ea}]])$$

this combined contextual dialogue representation s_t is then utilized in decoding process as illustrated above (see Section 3.1).

Emotional-Transformer We add the EA embedding v_{ea} into the Transformer network as shown in Figure 3.2. This integration is implemented by having another multi-head attention module attending to v_{ea}:

$$O_{ea} = MultiHead(v_{ea}, v_{ea}, O) \tag{5}$$

4 Implementation

4.1 Data Pre-processing

We split conversations in the DailyDialog dataset into training, validation, and test sets with 11118, 1000, and 1000 conversations, respectively. We sample the training / test data as follows: we sample one dialogue utterance as a target response (model groundtruth) for each utterance starting from the second one, and use all previous utterances in the same conversation as dialogue context. These contextual utterances are concatenated with a special delimiter token inbetween.

4.2 Implementation Details

Appropriate tools for implementing dialogue framework were carefully inspected and eventually OpenNMT (49) (a neural sequence learning framework) was selected in its Pytorch version, as it is well-accepted by the NLG research community. It also provides a set of open-sourced APIs for various language generation tasks including dialogue generation, that covers data prepossessing, model selection, training configuration and evaluation.

The implementation details are list as follows: the size of the vocabulary is set to 50,000, the dimension of the text embedding is 512. For RNN modules, hidden state size is set to 512; for Transformer modules, the dimension of feed-forward layer is set to 2048, number of heads is 8, dropout rate is 0.2. During the training of both systems, the batch size is 32, number of epochs is 25, optimization method is Adam with $\beta_2 = 0.998$, learning rate is set to 2 (recommended by OpenNMT), Cross-Entropy is used as the learning objective that calculates the loss based on the model output after Softmax layer and target response. For decoding, a beam search algorithm (50) is the choice to generate the target response.

5 Evaluation

As discussed in Section 2.4, Reference-based metrics like BLEU (41), ROUGE (42) and METEOR (43) do not provide fair measurement in evaluating open-domain dialogue systems. Therefore, we follow previous work of designing emotional dialogue system (3, 29) and perform the comparative evaluation against two automatic metrics: perplexity and accuracy. Perplexity is a common metrics in the evaluation of language models which indicates the confidence of the model output, a lower perplexity associate with better performance. This metric is computed as follows:

$$Perplexity(X, Y) = exp(\frac{-\sum_{i=1}^{|Y|} logP(y_i|y_{i-1}, ..., y_1, X)}{|Y|})$$

with X being the source sequence, Y the true target sequence and y_i the i-th target word. The numerator is the negative log likelihood of the loss function value (cross entropy score).

5.1 Evaluation of Non-emotional Dialog System

5.1.1 Experimental Objectives

In the experiments of non-emotional dialogue systems, we set two experimental objectives:

- To implement basic dialogue systems that can converse with human.
- To compare the performance of two selected structures (RNN and Transformer) in terms of the scores of automatic metrics and the quality of generated content.

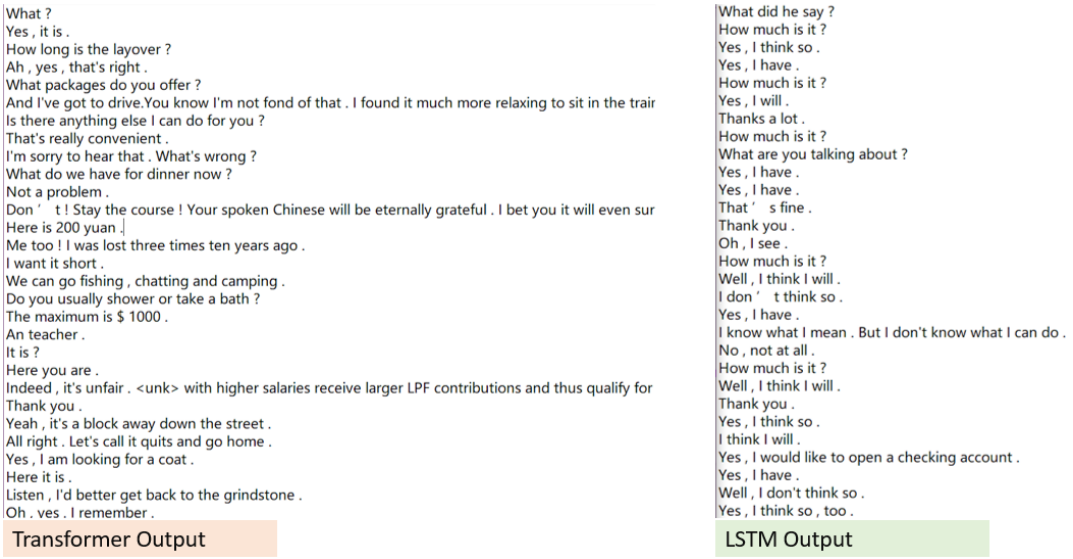


Figure 5.1: System generated utterances in the test set, left are the outputs from Transformer model and right are LSTM.

A: Does your child still believe in Santa Claus ?

B: Yes . She believes everything about him .

A: How sweet .

B (RNN): He must be very happy .

B (Trans): When are you going to tell her that Santa Claus doesn't exist ?

Figure 5.2: A and B are two speakers, last two lines are the generated responses from Transformer model and LSTM respectively.

Model	Perplexity	Accuracy
RNN	**38.20**	34.07
Transformer	55.71	**47.85**

Table 5.1: Table presents the performance of non-emotional dialogue systems.

Model	Perplexity	Accuracy
RNN-EA	**38.91**	34.37
Transformer-EA	62.25	**46.75**

Table 5.2: Table presents performance of emotional dialogue systems.

5.1.2 Experimental Results & Discussion

The results of two non-emotional implementations are presented in Table 5.1. The results suggest that the transformer architecture outperforms RNN architecture in terms of accuracy, but has worse perplexity score. Human evaluation has also been conducted on a small potion of the dataset. Figure 5.1, showing 29 outputs from the Transformer model and LSTM given the same context. Generally speaking, utterances generated by Transformer module are longer and more informative. This demonstrates the effectiveness of using Transformer as the backbone architecture. One case selected from the test set is presented with complete dialogue history in Figure 5.2. In this case, the RNN output (lablled in green) mistaken the gender of the child mentioned in the previous dialogue by referring the child as "He". Transformer output, on the contrast, is reasonable and human-like.

5.2 Evaluation of Emotional Dialog System

5.2.1 Experimental Objectives

We set two experimental objectives:

- Compare the performance of two dialogue systems embodied with emotional tags and dialogue acts.
- Investigate whether these additional information improves the performance of the proposed system.

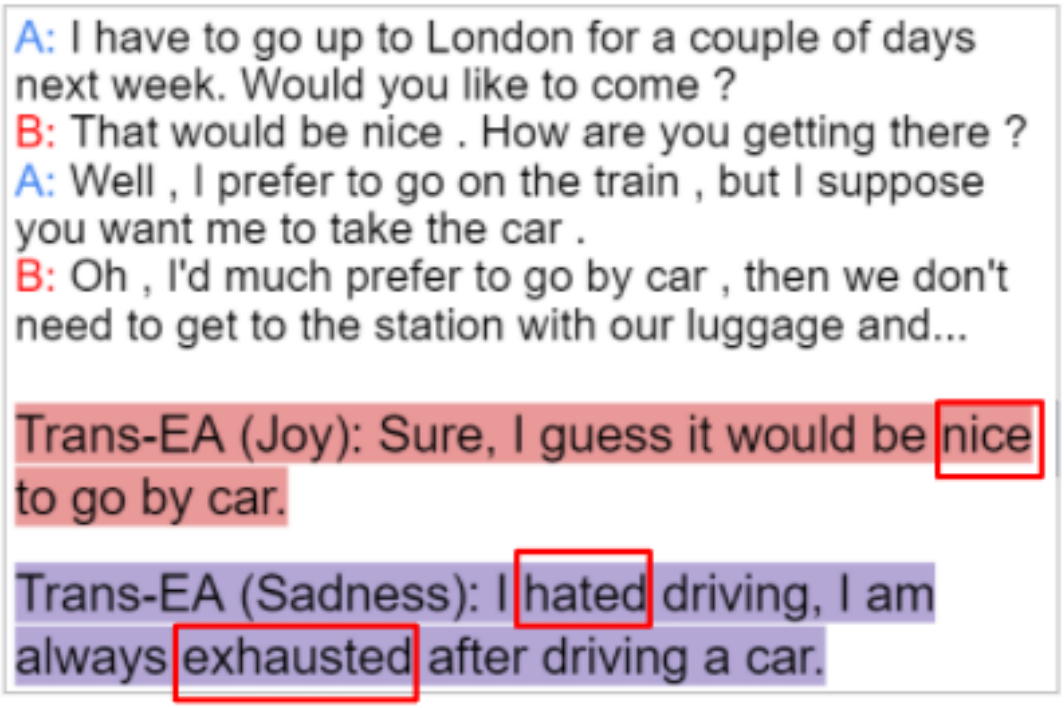


Figure 5.3: A and B are two speakers, last two lines are the generated responses from Transformer model with two different pre-configured emotions. Emotional words are circled with red boxes.

5.2.2 Experimental Results & Discussion

The results of emotional dialogue systems are presented in Table 5.1. We observe a similar trend in this table that the Transformer system produces higher accuracy but worse perplexity than RNN system. Based on reported figures, when compare emotional systems with non-emotional versions, there is no performance gain.

Even though the performance over the automatic metrics does not improve after adding emotions into the system, we conduct a manual inspection towards the generated content when set different emotions, and present one case in Figure 5.3, these cases are selected from the test set. From the examples we can see that, both the "Joy" agent and the "Sad" agent produce reasonable responses, and most importantly, they convey the accurate emotions.

5.3 Evaluation of User Engagement

For the evaluation of actual user experiences, we specifically implement a user interface for enabling human-machine interactions. The conversation program is built upon the Transformer-EA model trained on DailyDialog training set. In this program, we allow two conversation mode: human-machine conversation and machine-machine conversation, this means at each conversational turn, user can choose whether to converse with the AI agent by manual input or ask the AI agent to talk to itself. In order to prevent the dialogue context from growing too long, we cap the length of the input dialogue to five utterances, this means only the most recent five turns of the dialogue history will be preserved as the memory of the AI agent. In this section, we will present a few user cases collected during our user engagement experiments. The first user case, shown in Figure 5.4 is around pet topic, user starts the conversation by inputting "I like your dog" and ask the AI agent to talk to itself, we can observe that the whole dialogue is natural and sticking to the original topic. In the second user case, shown in Figure 5.5, we manually set the emotion of the AI agent to "Surprise" and start the conversation by talking about museum and let the AI agent speak. The first few sentences is somehow not very relevant to the museum topic (it mentioned "performance is excellent" and "grow old"). We then try to engage in the conversation by talking to the AI agent in a bi-turn manner. We set the emotion to "Joy" in the following two cases (Figure 5.7 and Figure 5.6). These two conversations behave natural and human-like.

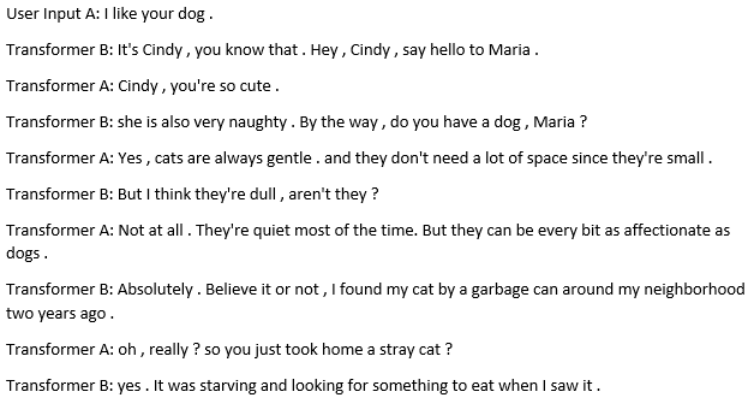


Figure 5.4: A and B are two speakers, "User Input" means the utterance is generated by human and "Transformer" is the AI agent. In this case, user starts the conversation with "I like your dog" and configures the model to extend the topic by talking to itself. The emotion of the AI agent is set to "Joy" in this conversation.

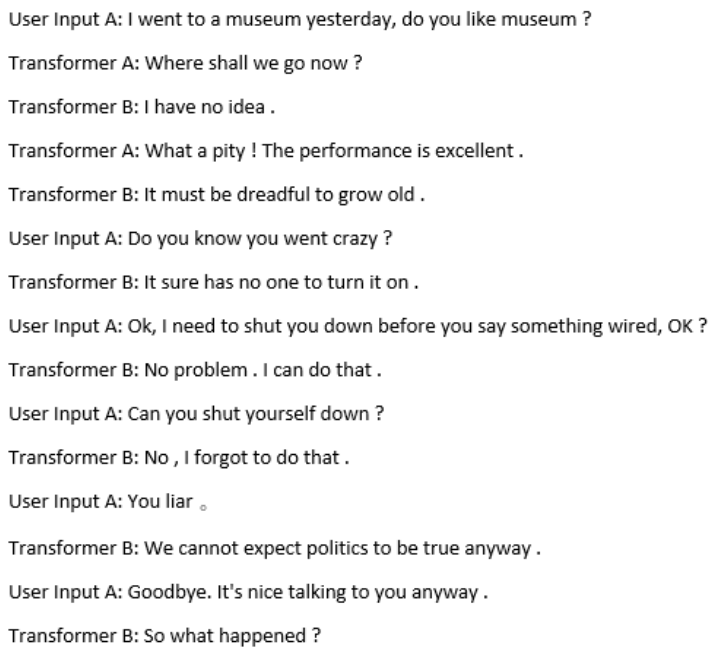


Figure 5.5: User starts the conversation by talking about museum. The emotion of the AI agent is set to "Surprise" in this conversation.

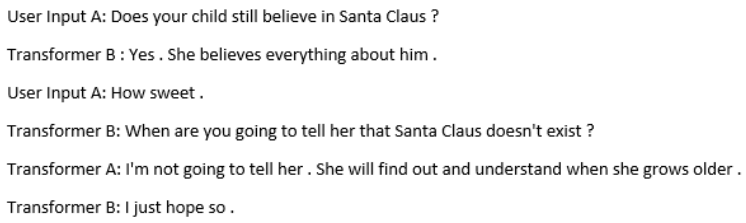


Figure 5.6: User starts the conversation by talking about Santa Claus. The emotion of the AI agent is set to "Joy" in this case.

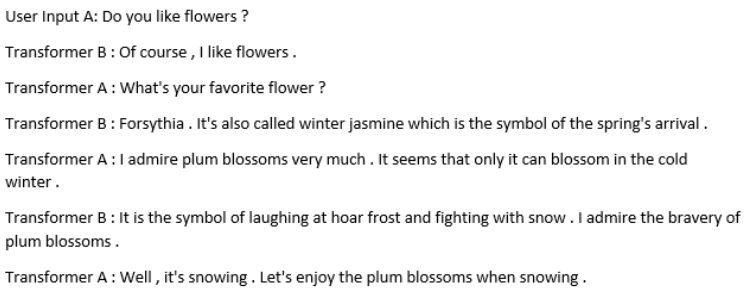


Figure 5.7: User starts the conversation by talking about flowers. The emotion of the AI agent is set to "Joy" in this case.

6 Conclusion

As artificial intelligence has advanced hugely in the recent decade, researchers are pursuing technologies with increasing similarities to human intelligence. As a subjective factor, machines that could understand emotion would be more accurate and responsive to human in human-machine interaction. In this dissertation, we propose to improve the performance of the dialogue system by integrating the analysis of dialogue emotions. For addressing this problem, we propose three research questions:

RQ1: How can we make the generated response express an assigned emotion? A particular aspect that we care about when building emotional dialogue system is the ability to express desired emotions. Therefore, we aim to explore the possibility of controlling the dialogue agent with specific emotions.

RQ2: To what extent does integrating emotional responses make an effect on automated metrics? The ability of recognizing user emotions also contributes the process of natural language understanding. For this research question, we will examine whether the integration of emotional contexts improves model performance during automatic evaluation.

RQ3: Does emotional content make the conversation more engaging? In the last research question, we propose to investigate whether the integration of emotional contexts improves overall user experience.

In this chapter, we review these research questions individually and present our findings with respect to addressing each of them. Then, we consider directions for potential impact of the research and possible future works extending and developing beyond the work presented in this dissertation.

6.1 Summary

To begin with, we conducted a comprehensive literature review around the topics of general methods in building dialogue systems, state-of-the-art emotional dialogue integration technologies, popular dialogue datasets and dialogue system evaluation methods. The literature provided us with valuable guidelines in both implementing the pipeline of a dialogue system and understanding the mechanisms behind the state-of-the-art choices. Among the existing dialogue generation approaches, we chose deep learning method for implementing the dialogue systems for its performance in modelling natural language and the generalisation ability. Specifically, we selected RNN and Transformer as two backbone structures in our experiments as they are state-of-the-art choices in the task of dialogue generation. We chose one public dialogue dataset released by Li et al. (3) to train our dialogue models. This dataset covers a wide range of daily life topics and each one of its utterance is manually labelled with a emotional tag and a action tag, which identifies with our initial motivation of investigating emotional dialogue systems.

After reviewing the works and resources relevant to our research goal, we conducted experiments using two selected neural architectures with and without the integration of additional contextual information. For both architectures, we converted the emotional tag and dialogue intention (act) tag into a latent space and integrated this embedding (we use concatenation operation for RNN, multi-head attention for Transformer) during the decoding phase. **This configuration allowed us to freely control the emotion of the dialogue system therefore answering the first research question (RQ1).**

The performance of non-emotional systems and emotional systems were compared by two automatic metrics, **however, in response to the second research question (RQ2), we did not observe a performance gain after integration of emotional information.**

To answer the last research question (RQ3), we explored the ability of emotional dialogue system to affect user experience by exploring the generated conversations. We noted that in some case the model can express the desired emotions and hence makes the conversation more human-like and engaging. However, in further human-machine experiments, we observed that the model performed well when setting it as "Joy" mode but did not converse normally when the

emotion was set to "Surprise". Although a more comprehensive evaluation is required to evaluate the ability to conveying the correct emotion of the dialogue system, the underlying reason behind this situation could be the bias in the dataset as over 70% of the dialogue are labelled as "Joy".

In summary, we explored all three research questions in this dissertation, we noticed that even though the performance was not improved by looking at the automatic metrics, the behaviour of the dialogue system will change according to the configured emotions.

6.2 The potential impact of the research

We noted that dialogue systems that are able to accurately recognize and convey emotions can communicate with the users in a more engaging manner and further enhance user satisfaction. Therefore, we argue that the research work completed in this dissertation could be potentially instantiated for industrial purposes.

One problem of deploying emotional dialogue system to real human users is that, even though the dialogue system can express a certain emotion, it still lack the ability of perceiving explicit user emotions. A possible direction for addressing automatic detection of user emotions can be implemented by combining the emotional dialogue system and facial recognition tools. Existing facial recognition technologies can provide fairly accurate emotional signals by analysing user facial expressions, these emotional information can be further analysed in the corresponding dialogue system for a better perception of user attitudes.

6.3 Future work

According to our experimental results, the performance of two automatic metrics did not improve after the integration of emotional inputs, we believe that there are still room for improving this performance by exploring different neural architecture and integration strategies.

Another interesting direction is to evaluate the quality of generated conversations by focusing on emotional accuracy, which means to measure whether the artificial utterance conveys the configured emotion correctly. For this evaluation dimension, some work in the filed of sentiment analysis or emotion classification may be

helpful.

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[2] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*, 2020.

[3] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[4] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

[5] Mihail Eric and Christopher D Manning. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*, 2017.

[6] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696, 2020.

[7] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.

[8] Helmut Prendinger and Mitsuru Ishizuka. The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, 19(3-4):267–285, 2005.

[9] Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, 62(2): 231–245, 2005.

[10] Bilyana Martinovski and David Traum. Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16, 2003.

[11] Carol Johnson, Laurie Hill, Jennifer Lock, Noha Altowairiki, Chris Ostrowski, Luciano da Rosa dos Santos, and Yang Liu. Using design-based research to develop meaningful online discussions in undergraduate field experience courses. *International Review of Research in Open and Distributed Learning: IRRODL*, 18 (6):36–53, 2017.

[12] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[13] Kenneth Mark Colby. *Artificial paranoia: A computer simulation of paranoid processes*, volume 49. Elsevier, 2013.

[14] Dimitra Gkatzia. Content selection in data-to-text systems: A survey. *arXiv preprint arXiv:1610.08375*, 2016.

[15] Irene Langkilde. Forest-based statistical sentence generation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[16] Srinivas Bangalore and Owen Rambow. Corpus-based lexical choice in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 464–471, 2000.

[17] Aoife Cahill, Martin Forst, and Christian Rohrer. Stochastic realisation ranking for a free word order language. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 17–24, 2007.

[18] Jost Schatzmann, Blaise Thomson, and Steve Young. Error simulation for training statistical dialogue systems. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 526–531. IEEE, 2007.

[19] Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528. ACL, 2006.

[20] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[21] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

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

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

[24] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[25] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[26] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[28] Ondřej Měkota, Memduh Gökırmak, and Petr Laitoch. End to end dialogue transformer. *arXiv preprint arXiv:2008.10392*, 2020.

[29] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[30] Xiao Sun, Jingyuan Li, and Jianhua Tao. Emotional conversation generation orientated syntactically constrained bidirectional-asynchronous framework. *IEEE Transactions on Affective Computing*, 2019.

[31] Peixiang Zhong, Di Wang, and Chunyan Miao. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7492–7500, 2019.

[32] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer, 2018.

[33] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[34] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

[35] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.

[36] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.

[37] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

[38] Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*, 2019.

[39] Yifan Fan and Xudong Luo. A survey of dialogue system evaluation. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1202–1209. IEEE, 2020.

[40] Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*, 2017.

[41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[42] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[43] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[44] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[45] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*, 2016.

[46] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. A practical dialogue-act-driven conversation model for multi-turn response selection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1980–1989, 2019.

[47] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[48] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. Dialogue-act-driven conversation model: An experimental study. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1246–1256, 2018.

[49] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

[50] Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Conference of the Association for Machine Translation in the Americas*, pages 115–124. Springer, 2004.