Foundations and Trends® in Information Retrieval

Deep Learning for Dialogue Systems: Chit-Chat and Beyond

Suggested Citation: Rui Yan, Juntao Li and Zhou Yu (2022), "Deep Learning for Dialogue Systems: Chit-Chat and Beyond", Foundations and Trends® in Information Retrieval: Vol. 15, No. 5, pp 417–588. DOI: 10.1561/1500000083.

Rui Yan
Gaoling School of Artificial Intelligence
Renmin University of China
ruiyan@ruc.edu.cn

Juntao Li
School of Computer Science and Technology
Soochow University
ljt@suda.edu.cn

Zhou Yu
Computer Science Department
Columbia University
zy2461@columbia.edu

now
the essence of knowledge
Boston — Delft

Contents

Deep Learning for Dialogue Systems: Chit-Chat and Beyond

Rui Yan[1], Juntao Li[2] and Zhou Yu[3]

[1] *Gaoling School of Artificial Intelligence, Renmin University of China, China; ruiyan@ruc.edu.cn*
[2] *School of Computer Science and Technology, Soochow University, China; ljt@suda.edu.cn*
[3] *Computer Science Department, Columbia University, USA; zy2461@columbia.edu*

ABSTRACT

With the rapid progress of deep neural models and the explosion of available data resources, dialogue systems that supports extensive topics and chit-chat conversations are emerging as a research hot-spot for many communities, e.g., information retrieval (IR), natural language processing (NLP), and machine learning (ML). Building a chit-chat system with retrieval techniques is an essential task and has achieved great success in the past few years. The advance of chit-chat systems, in turn, can support extensive IR tasks, e.g., conversational search, conversational recommendation. To facilitate the development of both retrieval-based chit-chat systems and IR tasks supported by these systems, we survey chit-chat systems from two perspectives: (1) techniques to build chit-chat systems, i.e., deep retrieval-based models, generative methods, and their ensembles, (2) chit-chat components in completing IR tasks. In each aspect,

Rui Yan, Juntao Li and Zhou Yu (2022), "Deep Learning for Dialogue Systems: Chit-Chat and Beyond", Foundations and Trends® in Information Retrieval: Vol. 15, No. 5, pp 417–588. DOI: 10.1561/1500000083.

we present cutting-edge neural methods and summarize the core challenges encountered and possible research directions.

1

Introduction

Starting from the 1960s, conversational artificial intelligence has become a crucial research field and has grabbed much more attention in recent years. Empowered by deep neural models, dialogue systems have demonstrated very impressive and appealing performance in virtual assistants and social bots. In viewing its potential and values, mainstream NLP, IR, and even ML communities have started contributing to dialogue systems. Dialogue systems can be roughly grouped into two classes, i.e., task-oriented and chit-chat systems. The former group focuses on completing predefined tasks with task-specific constraints and goals, e.g., restaurant booking and making calls. The later systems are mainly designed for modeling the 'chats' characteristic of human-human conversations (Daniel and James, 2020) without specific goals and constraints, i.e., the topics of the conversation could be any. Given predefined constraints and goals, task-oriented systems can achieve impressive performance with limited data and computational resources. In contrast, chit-chat systems require massive training conversations to mimic human chatting with extensive topics. Unlike task-oriented systems that have achieved great success for decades, learning-based chit-chat systems have not made great strides until recent years with the explosion of both data resources, model capacity (data modeling capability of deep neural networks), and computational power. To facilitate the development of chit-chat systems and their supported IR tasks and bridge the gap between different research communities, especially for the NLP and IR fields, we propose to systematically review state-of-the-art chit-chat systems and draw the connections between chit-chat and tasks, from being supporting tasks and the unified modeling framework in the paradigm of pre-trained language models.

Specifically, our work has a deep concentration on deep neural chit-chat systems using IR techniques and NLP methods, i.e., this monograph presents lessons and experiences of how to establish relevant, coherent, diverse, knowledgeable, and human-like chit-chat systems. Besides, we also discuss the connections between chit-chat systems and tasks, ranging from the perspectives of treating chit-chat components as supporting tasks to make task completion more natural (e.g., recommendation) to the trend of leveraging a unified framework for various downstream tasks in the era of pre-trained language models. To the best of our knowledge, it is the first survey to cover these topics and features.

The main contributions of this survey are as follows:

- We thoroughly survey the deep neural models in recent years for chit-chat systems, ranging from retrieval-based methods to generation-based approaches and the ensemble of these two types of models.

  We provide the connections between the recently resurgent chit-chat systems and task-oriented systems, e.g., conversational recommendation and conversational search, which enables us to explore more possibilities of building either better chit-chat systems or improving user experience in constructing IR systems.

- We introduce various solutions for addressing or mitigating the confronted challenges (e.g., context modeling, one-to-diversity, human factors learning) from different perspectives, including data-side and model-side solutions and utilization of extra resources.

- We present necessary data resources and evaluation methods for building retrieval-based and generation-based chit-chat systems.

We also analyze the main challenges that we are facing and give the possible exploration directions and the rising trends, which will shed light on building human-like systems.

1.1 Intended Audience and Scope

This survey is intended to bridge the researchers of IR and the NLP community to move chit-chat systems forward and support more IR tasks. Our target audience includes, but is not limited to, IR or NLP researchers who want to study chit-chat from different perspectives, e.g., compensating retrieval-based models with the generation or vice versa, IR researchers who need to complete their tasks with the assistance of chit-chat systems, engineers with hands-on experience in building chit-chat systems to leverage advanced chit-chat modeling techniques, anyone who intends to quickly keep up with the frontier of chit-chat systems, anyone who wants to learn how to build chit-chat systems with deep neural architectures.

The main scope of this survey is based on the tutorial of SIGIR 2019 and WWW 2019 (Wu and Yan, 2019a, 2019b). We expand the tutorial contents with up-to-date techniques for building chit-chat systems, covering retrieval-based methods, generation components, and their ensembles. Besides the above contents, we also discuss the role of chit-chat systems in completing tasks, especially for some emerging IR tasks, e.g., conversational search and conversational recommendation. Considering the new trend of utilizing a unified self-supervised pre-training framework for both chit-chat and IR tasks, we further review a few recent works in this line and point out the possible future direction.

The rest of this survey is structured as follows:

- The remainder of this section summarizes the importance of chit-chat systems and presents the core problems of chit-chat systems. Besides, the landscape of chit-chat systems is also introduced. At the end of this section, we clarify the relationship and discrepancy between this survey and recent papers.

- Section 2 briefly reviews classic chit-chat systems before the neural age, including rule-based, template-based, and learning-based methods, and summarizes the characteristics of these methods.

- Section 3 sorts out and elaborates retrieval-based dialogue systems in recent years. This section starts with the pre-processing of conversation data and then discusses the core problems of retrieval-based chit-chat systems in detail (e.g., context modeling, knowledge utilization, human factors learning), which ends with necessary data resources and evaluation metrics for building retrieval-based chit-chat systems.
- Section 4 provides an alternative option for building chit-chat systems, i.e., generation-based methods, focusing on the pros and cons of generation-based methods in building chit-chat systems and their relationships with retrieval-based solutions. The last part of this section gives essential data resources, evaluation methods, and current challenges.
- In Section 5, we describe the ensemble of the aforementioned two types of frameworks, focusing on the scenarios of integration and re-ranking, template and prototype, and adversarial learning. Section 6 connects chit-chat systems with tasks, including vanilla tasks and newly appeared IR tasks like conversational search, and reveals the trend of unifying chit-chat dialogues and tasks with large-scale pre-trained language models.
- Section 7 first concludes this survey with the progress of chit-chat systems and the chit-chat component in IR tasks, and then points out the ongoing struggles and the possible future trends.

1.2 The Importance of Chit-Chat Systems

Chit-chat systems have become more and more popular and important in both academia and industry. Studying chit-chat systems have various benefits, including providing helpful services to human users, promoting the development of artificial intelligence technologies, holding tremendous potential and commercial values in the future.

To human users, chit-chat systems can satisfy a myriad of human needs, such as communication, social belongings, emotional engagement,

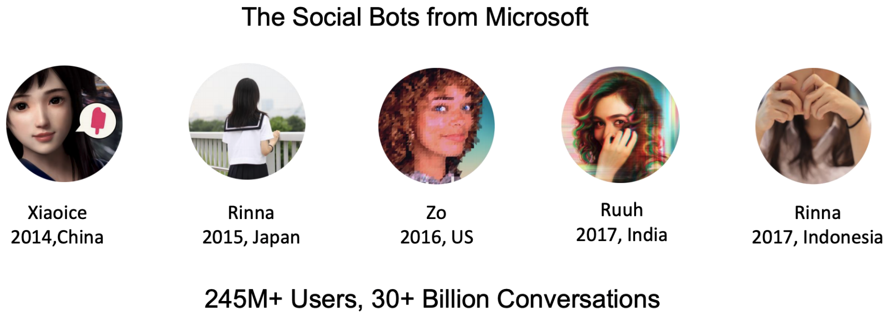

Figure 1.1: User size of social bots from Microsoft (Wu and Yan, 2019b).

etc (Huang *et al.*, 2020b). On account of these merits, various applications, including but not limited to virtual assistants, smart speakers, social bots, and virtual customer services, are developed. As shown in Figure 1.1, chit-chat systems from the Microsoft corporation alone attracted over 245 million users and achieved over 30 billion conversations by 2019.

As for the connections between chit-chat systems and technology development, it is an indicator to calibrate the progress of artificial intelligence by launching the Turing test which is designed to test whether a machine can exhibit intelligent behaviours equivalent or indistinguishable from a human[1]. Building chit-chat systems also poses various unique challenges to state-of-the-art deep neural models, e.g., one-to-diversity, long-range context modeling, topic shift, long-term engagement computation, human factors learning, and the settlement of these problems, in turn, facilitates the progress of deep learning methods and encourages technical development.

Except for contributing to technology development and human needs, chit-chat systems also connect to various online commercial services. As demonstrated in Figure 1.2, chit-chat conversation might mix with goal-oriented demands, such as question answering, image search, and recommendation. Exploring chit-chat conversations could seamlessly find the demands of users and complete different tasks in a more efficient manner accordingly, i.e., without introducing multiple

[1] https://en.wikipedia.org/wiki/Turing_test

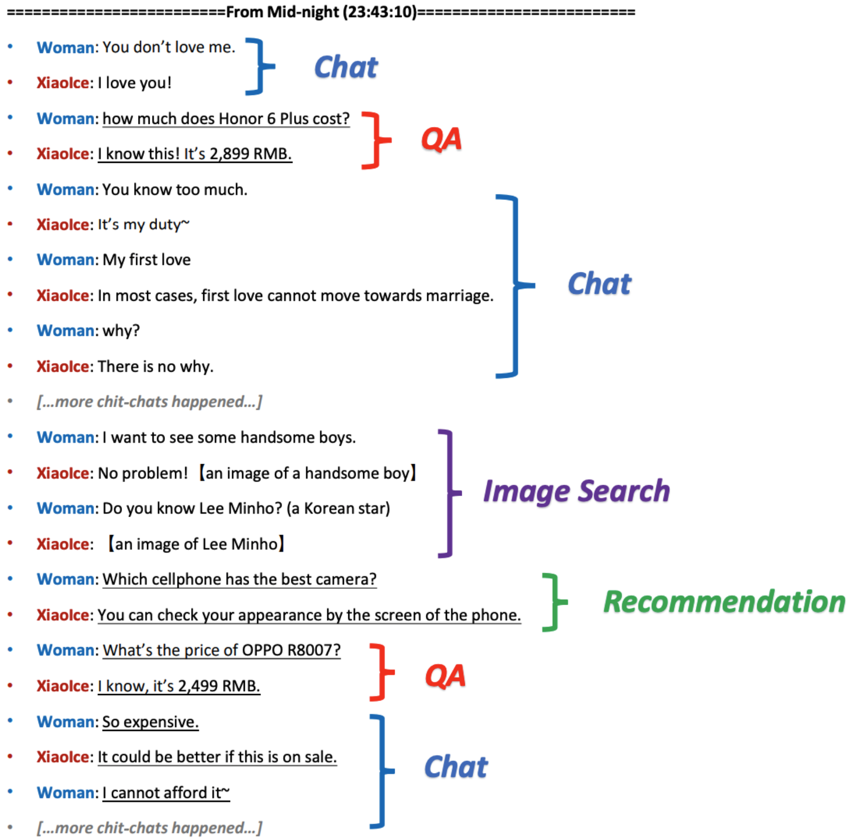

Figure 1.2: A case study that demonstrates the connections between chatbots and goal-oriented applications (Wu and Yan, 2019b).

task-specific systems. It also serves an essential role in intelligent entities and devices by providing the human-machine interface. The progress of chatbots could assist the development of robotics.

With the rapid progress of conversational AI techniques that support human-like interactions between computers and humans, it can be imaged that chit-chat systems are likely to have more industrial applications and broad market prospects. We believe that the potential of these systems is far more than we have seen in recent few years on social bots, virtual assistants, information seeking systems. In the far future, conversational AI systems might change almost everything in our daily life, e.g., the games will be more immersive, robotics are more intelligent that is able to sing, talk and even make friends with humans.

1.3 The Core Problems of Chit-Chat Systems

One of the main goals of chit-chat systems is to pass the Turing test so as to prove that an artificial program can chat like humans. Thus, the properties of human conversations should be considered and modeled in chit-chat systems. Seeing that human conversations are intricate and difficult to formulate, we utilize the qualitative analysis results of human conversation properties in Daniel and James (2020) to divide and shape the core problems of chit-chat systems. There are mainly six basic properties for human conversations: (1) **turns**, (2) **speech acts**, (3) **grounding**, (4) **sub-dialogues and dialogue structure**, (5) **initiative**, and (6) **inference and implicature**. For deep neural models trained on massive conversation data, **speech acts** and **dialogue structure** are implicitly modelled by neural networks. As for **initiative**, **user-initiative** and **system-initiative** frameworks are more common for task-specific systems while **mixed initiative** are very difficult to achieve. Thus, researchers mainly focus on the following problems in deep neural chit-chat systems.

Context Modeling. One of the main challenges we encountered is the long-range context modeling. Unlike task-oriented conversations that mainly consist of task-specific contents and usually complete a user demand in no more than a few dozens of conversation turns, chit-chat conversation is tied up with over hundreds of turns in usual, owing to the non-goal-oriented nature of chit-chat. In view of this, long-range context modeling has become a crucial issue for chit-chat dialogues to make conversations more consistent and coherent.

One-to-Diversity. In addition to multi-turn context modeling, one-to-diversity has also hindered the development of chit-chat systems. Unlike task-oriented conversations that take task completion as the evaluation metric, chit-chat further needs to mimic human-like conversations. Among various characteristics of human conversations, modeling expression diversity and one-to-many correlations bear the brunt.

Knowledge and Grounding. Beyond learning statistical patterns from existing human conversations, advanced chit-chat systems are expected to master and leverage knowledge like human beings. Besides commonsense knowledge, chit-chat conversations often correlate with non-contextual information, i.e., information and content that are not in context. Hereafter, we denote these extra information as grounding.

Human Factors. For chit-chat systems, user experience and engagement are always the core. To build a better chit-chat system, we have to consider the influence of various human factors, such as personalized expression preference, emotional changes, and beyond.

1.4 Landscape of Chit-Chat Systems and Beyond

A View from Chit-Chat. Advanced chit-chat systems mainly utilize cutting-edge deep neural techniques to automatically obtain responses for any newly given query or dialogue contexts. We group existing chit-chat models into three categories, i.e., frameworks based on retrieval techniques, generation-based models, and the ensemble of these two kinds of solutions. Retrieval-based frameworks mainly study how to automatically select feasible response candidates, covering the multi-turn context matching, extra resource utilization, human factors constraining, and pre-trained context-aware representation usages. Generation-based research focuses on the limitations of sequence to sequence networks, exploring from the perspective of data manipulation, generation pipelines, training objectives, large-scale pre-trained language models, and aforementioned context modeling, as well as human factors. Ensemble solutions investigate how to compensate retrieval-based dialogue systems with the merits of generation models and vice versa.

Linking Chit-Chat with Tasks. The connections between chit-chat and tasks can be categorized into three different directions. One is to discover and complete specific goals from chit-chat human-machine conversations to achieve better user engagement. The second is to enhance downstream tasks with chit-chat components, e.g., it can make it easier for users to accept recommended items from commercial recommendation systems.

Another possible direction is to utilize a unified large-scale pre-training framework to complete chit-chat conversations and tasks.

1.5 Comparisons with Existing Surveys

Recently, several tutorials and survey papers on dialogue systems have been presented (Yih *et al.*, 2015; Yih *et al.*, 2016), focusing on deep learning techniques and various IR related tasks (Gao, 2017), e.g., question answering (QA). Li and Yan (2018) briefly reviewed the multi-turn conversation methods involved in the NLPCC 2018 shared task, including both retrieval models and generation solutions. Chen *et al.* (2018b) provided a tutorial on spoken dialogue systems, which is mainly about traditional task-oriented dialogue systems. Serban *et al.* (2018) offered a thorough investigation on the public data available for building dialogue systems. Gao *et al.* (2019a) covered a myriad of topics in dialogue systems, including question answering, reading comprehension, task-oriented systems, social bots and industrial applications. Huang *et al.* (2020b) comprehensively studied three challenges that researchers are facing at present in building intelligent dialogue systems. Yan and Wu (2021) briefly summarized the progress and future of chit-chat dialogues with limited coverage and insufficient in-depth study. Zamani *et al.* (2022) recently provided an overview of existing research related to conversational information seeking. Gao *et al.* (2022) also wrote a book about conversational information seeking but focused on recent advances and technical details for building the main modules of conversational information retrieval systems. Considering that deep neural-based systems are the mainstream and are still in the process of development, we mainly compare this paper with recent surveys in closely related fields. More concretely, we conduct comparisons with two recent papers presented by Gao *et al.* (2019a) and Huang *et al.* (2020b), respectively.

This survey differs with Gao *et al.* (2019a) from the following aspects:

- We mainly focus on chit-chat systems rather than focus on task-oriented systems, question answering, and machine reading comprehension.

- We group recent research from the view of chit-chat, a specific type of conversation system that has attracted millions of users, instead of connecting goal-oriented dialogues and fully data-driven social bots from a unified perspective of optimal decision making.
- We mainly survey end-to-end methods built upon deep learning methods, instead of presenting task-oriented pipeline models or connecting traditional machine learning methods with modern neural models.
- We expose the recently explosive progress of completing tasks with the assistance of chit-chat systems, e.g., conversational recommendation (Lei *et al.*, 2020). We review the new paradigm of building chit-chat systems and completing tasks in recent large-scale pre-trained language models.

Compared with the short survey written by Huang *et al.* (2020b), we further present the following contents.

- Instead of focusing on surveying research that relates to specific challenges of chit-chat systems, we present a comprehensive study of modern chit-chat systems based on deep neural models.
- Except for discussing the main challenges that we are facing, this survey presents various solutions for addressing a myriad of challenges in the chit-chat conversations, which can provide guidance for anyone who wants to build chit-chat systems.
- This survey also has a border coverage, which draws the connection between chit-chat and goal-oriented systems, and emerging tasks of the IR community.

2

Classic Dialogue Systems before Neural Age

This section briefly reviews the representative dialogue systems before the neural age, including rule/template-based systems and learning-based solutions.

2.1 Rule-Based Methods

Starting from ELIZA (Weizenbaum, 1966), the chatbot has become one of the important branches of building intelligent programs. Among various solutions, rule(template)-based methods take the dominant role from the era of ELIZA and the ensuing decades. This section takes ELIZA as the representative to demonstrate how to utilize limited rules to construct a program that can have certain types of natural language conversation with humans. Figure 2.1 presents a part of the conversation example in Weizenbaum (1966) for simplicity. We take the user-issued utterance "You don't argue with me." as an illustration to explain how ELIZA responds. The first step is to detect the target keyword "you" from the user-issued utterance, where each word in the utterance is looked up in a dictionary of keywords, and the highest-ranked one of the identified

TELL ME MORE ABOUT YOUR FAMILY

My mother takes care of me.

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

My father.

YOUR FATHER

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

My father is afraid of everybody.

WHAT ELSE COMES TO YOUR MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Figure 2.1: A typical conversation between chatbot ELIZA and human user, where the capitalized lines are the machine responses (Weizenbaum, 1966).

keywords is selected as the target[1]. Then, each decomposition rule that contains the target keyword "you" is tried. For example, a specific template associated with the keyword "you" decomposes the utterance

"`You don't argue with me.`"

into:

"`You 0 me.`"

where "`0`" refers to a variable number of words. Certain transformation rule is then involved to translate ("`You 0 me.`") as:

"`I 0 you.`"

with the reassemble rule (tacking with a prefix "Why do you think"):

"`Why do you think I 0 you`"

Finally, ELIZA outputs "WHY DO YOU THINK I DON'T ARGUE WITH YOU" for the user-issued utterance. Through analyzing the

[1]If there is no matched keyword in the dictionary, the system will output non-committal responses.

above conversation, typical rule-based techniques in ELIZA can be summarized as:

- The identification of the target keyword (in the keyword vocabulary and with the highest-ranked score) from the given utterance.
- Linking keyword with associated decomposition rules.
- Choosing proper transformation rule.
- Selecting reassemble rule utilized for creating response to the user-issued utterance.
- Designing provision mechanisms that can deal with unexpected cases, e.g., responding utterance without matched keyword with non-committal responses such as "I see" (Daniel and James, 2020).

Later on, Colby *et al.* (1971) utilize a chatbot, namely PARRY, to behave like a paranoid person. On the basis of ELIZA's chatting capability, PARRY further models agent-level affections, i.e., fear and anger. Some conversation topics might make PARRY accumulate anger, while other topics will cause more fear. With two variables to present the degree of fear and anger, PARRY can respond with different affection states. Since chatting like a paranoid resembles ELIZA as a Rogerian psychologist and with more capabilities, PARRY is the first chatbot known to pass the Turing Test. Most of the following chatbots choose similar and more favorable settings to pass the Turing test. Even some modern chatbots are still based on the influential architecture of ELIZA.

After ELIZA and PARRY, many efforts have been devoted to optimize the techniques of rule-based chatbots (Bradeško and Mladenić, 2012). Weintraub (1986) introduces parsing to augment keyword identification and propose to improve pattern matching and word vocabulary, which won the Loebner Prize Competition from 1991 to 1993. Other strategies of better parser, pattern matching, extra databases (Hutchens, 1996; Wallace, 2003; Copple, 2008). More rule-based methods can be found in Thorat and Jadhav (2020).

2.2 Learning-Based Methods

Since writing rules is very tedious and it is also difficult to cover most conversation situations, learning-based methods have been widely explored (Shawar and Atwell, 2005; Schatzmann *et al.*, 2006; Thorat and Jadhav, 2020). Note that this section only quickly reviews a few representative learning-based dialogue systems before the deep neural age. Litman *et al.* (2000) use the formalism of Markov decision processes and reinforcement learning (RL) algorithms to learn dialogue policy. Williams (2007) formulates the dialogue system as a partially observable Markov decision process to build a unified statistical framework that can globally optimize each separate technique involved in a spoken dialogue system. Ritter *et al.* (2011) use phrase-based Statistical Machine Translation method to generate a response for a linguistic stimulus. Misu *et al.* (2012) also utilize reinforcement learning to build dialogue policy but focus on a specific application.

Compared with rule-based solutions, the merits of learning-based methods are two-fold. For one thing, learning-based methods can automatically complete conversation procedures without much hand-crafted efforts like ELIZA by capturing matching patterns and correlations between user-issued utterances and possible responses, generating responses with learned assemble strategies, and editing or creating responses as the provision mechanism. Thus, learning-based methods can be directly adapted to new scenarios with training corpora in short time as it does not take too much time to write various types of rules. For another, learning-based methods can scale up to large size data to capture more patterns, as a result of which they can handle more conversation situations failed in rule-based systems in which there are no matched rules.

2.3 Reconsider the Problem

Compared with rule-based methods and learning-based solutions, the former category is more efficient for data-scarce situations and can efficiently deal with straightforward queries from users, while the latter one has the merits of dealing with uncovered rules/patterns well without much hand-crafted efforts and is easier to transplant to new domains.

On the one hand, with the explosion of online available conversation-like data (e.g., posts on Twitter, Weibo) and the demands of chatbot from millions of netizens, it is very costly and even unlikely to manually write conversation rules for specific types of dialogue with myriads of topics, domains, and users. On the other hand, learning-based methods are not mature yet, which usually outputs irrelevant responses for user-issued queries, leading to low trust and user experience. Another challenge of learning-based methods is figuring out what to optimize because simply mimicking human responses is insufficient in many scenarios and applications. In view of this, rule-based methods are still the mainstream of commercial and industry dialogue systems, while researchers never cease to address the limitations of learning-based methods. One of the most important directions of learning-based methods is to learn more from explosive data with a powerful model to capture sophisticated correlations, conversation patterns, and also representations, e.g., using deep models rather than shallow learning-based methods in previous studies. To make the best use of the merits of learning-based methods, another possible aspect is to study fully end-to-end chatbot framework. Most of the progress of chatbots in recent years mainly concentrates on these two aspects, i.e., deep learning-based end-to-end models, which will be elaborated in the following sections of this survey.

3

Retrieval-Based Chit-Chat Systems

This section presents the essential and widely used paradigm in real-world applications, i.e., retrieval-based chit-chat systems. The success of these systems correlates with the social attribute of human beings and the availability of massive human conversation data. That is, different people might pursue similar needs from chit-chat systems, such as emotional engagement, counselling, personal assistant, etc. Owing to the convenient access of web information, e.g., Twitter, Facebook, and TikTok, people also have similar hotspot topics to converse with. Besides, human beings can share commonsense knowledge with others. Based on these phenomena and observations, it is feasible to retrieve a proper response from existing conversations for a given user query and corresponding dialogue context. With the ubiquitously available resources and powerful retrieval engine, building a retrieval-based chit-chat system has gained increasing interest in recent years.

We start this section with an introduction to the paradigm of retrieval. We then present the essential indexing and pre-retrieval process before retrieving and ranking the response candidates with a matching neural model. Later on, we discuss the common challenges and advanced solutions of retrieval, including history modeling, extra information utilization, human factors learning, and pre-trained language models. We end this section with evaluation metrics for building a retrieval-based dialogue system.

3.1 The Paradigm of Retrieval

In most cases, we can obtain a correlated response from existing conversation data by powerful retrieval engine and matching models. These retrieved responses are created by human beings, which are grammatically fluent and contain informative content. Leveraging these existing conversations as system outputs makes dialogue systems perform like humans.

For each query, the retrieval system needs to select a proper response from millions of conversation utterances. It is natural to launch a coarse-to-fine retrieval pipeline. That is, the system first conducts pre-retrieval from existing conversations with a powerful engine and then utilizes a deep neural model to perform a fine-grained selection from the retrieved candidates. In other words, most of the existing retrieval-based dialogue systems can be decomposed into two stages, i.e., the simple and fast pre-retrieval and the sophisticated candidate ranking.

3.1.1 Indexing and Pre-Retrieval

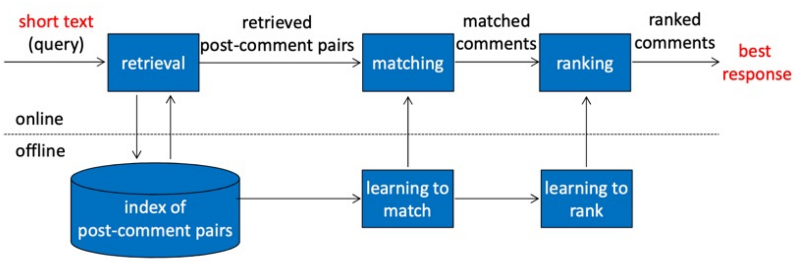

Figure 3.1: System architecture of retrieval-based short text conversation (Ji *et al.*, 2014).

The primary retrieval-based systems interact with humans via extracting a suitable and in-context response from a pre-built database with index (Ji *et al.*, 2014). Specifically, Ji *et al.* (2014) break down the retrieval-based short text conversation into three stages, namely retrieval, matching, and ranking. Figure 3.1 sketches out the overall framework of the system. The system first leverages a few fast and efficient modules, e.g., linear models that can calculate *Query-Response Similarity*, *Query-Message Similarity* and *Query-Response Matching in Latent Space*, to recall or pre-retrieve a few candidates based on their index. Then, more powerful matching models, e.g., language model, deep matching network, are used to extract more matching features between query and the pre-retrieved candidates in the matching stage. Finally, the matching features are further aggregated and processed by an efficient ranking component, e.g., linear function, to assign different scores for the pre-retrieved candidates. With the ranking scores, one can simply select the most suitable candidate as the response for a given query. Both the matching models and linear ranker can be trained offline.

3.1.2 Response Selection Frameworks

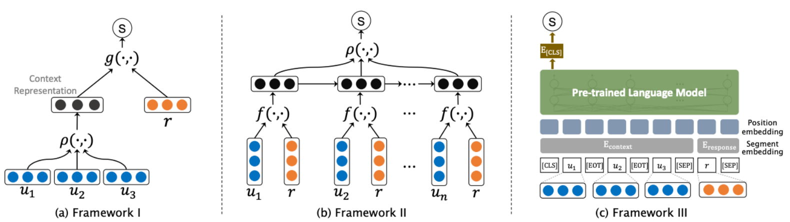

Figure 3.2: Three typical matching paradigm for response selection: (1) representation-based, (2) interaction-based, (3) PLM (Pre-trained Language Model)-based methods (Tao *et al.*, 2021b).

For deep neural retrieval-based chit-chat systems, learning to match and rank is essential, and researchers usually formulate matching and ranking as a response selection task (Tao *et al.*, 2021b). The core of response retrieval or selection is to learn a context-response matching model $f(\cdot)$ from training data to compute the matching degree between a dialogue context c_i and a given response candidate r_i. The training objective of a context-response matching model is

$$\mathcal{L} = -\sum_{i=1}^{N} y_i log(f(c_i, r_i)) + (1 - y_i) log(1 - f(c_i, r_i)) \tag{3.1}$$

where y_i is the label to indicate whether a response candidate is appropriate for the given dialogue context. As presented in Figure 3.2, there are mainly three types of context-response matching frameworks. Different from Tao *et al.* (2021b) that group existing neural context-response matching methods based on their frameworks, we structure existing literature from the type of context information to be processed. In this survey, we refer dialogue context to dialogue history, user profile, and extra resources (e.g., document, image caption, video). The modeling of each type of dialogue context is elaborated in detail in the following sections. At the end of this section, we give a brief review of pre-trained language models (PLMs) for context-response selection. Since PLMs are developing rapidly, and current methods will get outdated quickly, we mainly discuss the role of PLMs in retrieval-based dialogue systems.

3.2 Dialogue History Modeling

A long-term and core problem of the retrieval-based dialogue system is the modeling of the multi-turn dialogue history. Earlier studies pay attention to constructing single-turn history-response matching models where only a single utterance is considered, or multiple utterances in the history are concatenated into a long sequence for response selection. Recently, most studies have focused on the multi-turn scenario where each utterance in the dialogue history first interacts with a candidate response in turn. Then the interacted signals are sequentially aggregated based on the utterance's order in the dialogue history.

3.2.1 Single-turn Dialogue History Modeling

Early works of retrieval-based chit-chat systems mainly study single-turn response selection, where the last turn of dialogue history (i.e., the dialogue query) is used to select a proper response. Many works explore learning better representations for dialogue history (query) and response separately to calibrate the correlations between the last turn dialogue history and response candidates. Then the matching degree is obtained based on these two types of encoded representation vectors. The essence of most retrieval-based natural language processing problems is text matching, and we first introduce some classic text matching methods.

Lu and Li (2013) propose a DNN (Deep Neural Network)-based matching model, namely DeepMatch$_{topic}$, for short texts response selection, which combines both the localness and hierarchy intrinsic in the structure, where localness corresponds to salient local structure in the semantic space, hierarchy refers to different levels of abstraction in matching. Hu *et al.* (2014) propose ARC-I/ARC-II, which further enhance the matching degree computation with the deep convolutional neural network. The improved architecture can either learn the representation of query and response separately or the representation of interacted query and response. A multi-layer perceptron then processes the obtained representation to calculate the matching degree information. Wu *et al.* (2018) introduce extra knowledge to enhance long text matching, including the answer selection task and response selection task. The motivation behind introducing prior knowledge is to filter out irrelevant information in extracting matching signals. To achieve this, the authors utilize an ultra-simple gate mechanism to learn a weighted sum of the knowledge embedding and word embedding. Then, three different types of interaction based on word embeddings, aggregated word embeddings by BiGRU, and aggregated knowledge-enhanced word embeddings by BiGRU, are computed to capture the correlations between two textural sequences. These three types of similarity matrices are further processed as three input channels of a CNN module along with an MLP layer to obtain the final matching score of two sequences. Tay *et al.* (2018a) enhance the effective co-attention strategy in text-matching with the asymmetrical Hermitian Inner Product to expand its effectiveness in asymmetrical matching tasks, e.g., the matching degree calculation between query and response. Moreover, Tay *et al.* (2018b) further propose MACN as a new paradigm of utilizing attention not as a pooling operator but as a form of feature augmentation.

However, dialogue datasets often contain several turns of dialogue utterances. Thus, researchers try to utilize this helpful information

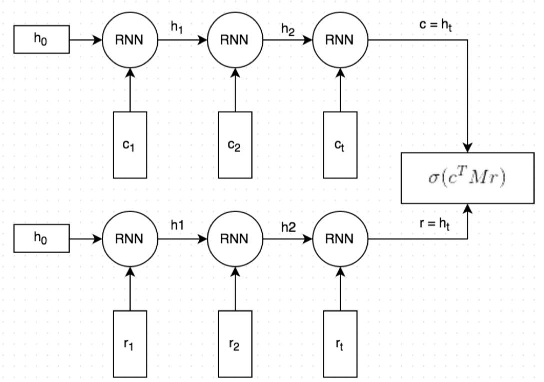

Figure 3.3: Diagram of dual-LSTMs (Lowe *et al.*, 2015).

based on such single-turn history modeling. Lowe *et al.* (2015) introduce the Ubuntu Dialogue dataset, consisting of 7 million utterances and nearly 1 million dialogues, and propose the dual LSTM model to encode multi-turn dialogue history and response, respectively, which is depicted in Figure 3.3. To model multi-turn dialogue history, the authors choose to concatenate multiple utterances in the history into a long sequence, which is similar to single-turn dialogue history modeling.

3.2.2 Multi-turn Dialogue History Modeling

	Context
utterance 1	*Human*: How are you doing?
utterance 2	*ChatBot*: I am going to **hold a drum class** in Shanghai. Anyone wants to join? The location is near Lujiazui.
utterance 3	*Human*: Interesting! Do you have coaches who can help me practice **drum**?
utterance 4	*ChatBot*: Of course.
utterance 5	*Human*: Can I have a free first lesson?
	Response Candidates
response 1	Sure. Have you ever played drum before? ✓
response 2	What lessons do you want? ✗

Figure 3.4: A multi-turn conversation sample from Wu *et al.* (2017).

Apart from the above single-turn modeling strategies, many re-

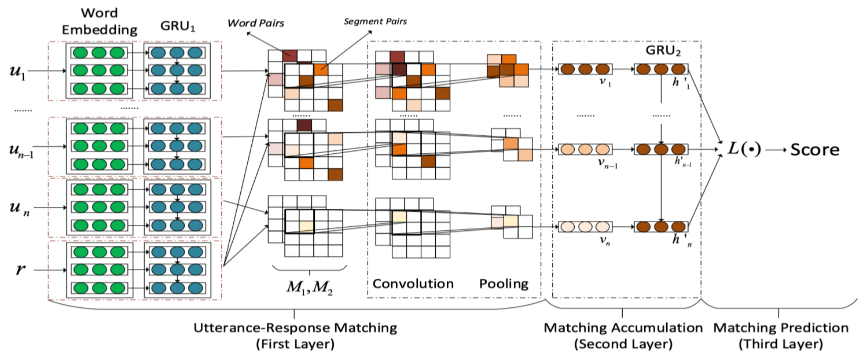

Figure 3.5: Architecture of SMN, a representative model of "representation-matching-aggregation" paradigm (Wu *et al.*, 2017).

searchers have begun to emphasize the hierarchical structure of multi-turn dialogue history as shown in Figure 3.4.

Beyond directly concatenating all turns of utterances, many multi-turn dialogue history modeling methods have been proposed. For example, Zhou *et al.* (2016) use RNN to read history and response and use the last hidden states to represent history and response as two semantic vectors. The obtained semantic vectors are utilized to measure their relevance. Yan *et al.* (2016) propose DL2R, which first reformulates the input query and then combines the matching information between the reformulated and the original queries as well as the retrieved queries and responses, respectively. Since then, most researchers in the literature have adopted the "representation-matching-aggregation" paradigm to build the matching models.

Wu *et al.* (2017) propose a novel sequential matching network (SMN) to deal with multi-turn history where the representation of each history utterance is based on its interaction with the response candidate. The representation sequence of the history utterances is then processed by GRUs to compute aggregated matching signals, which are further utlized for calculating the matching score. The paradigm is depicted in Figure 3.5. Zhang *et al.* (2018e) then propose to enhance SMN with a gated self-attention mechanism, named Deep Utterance Aggregation (DUA), to improve the representation learning process of query and response candidates. Moreover, Zhou *et al.* (2018d) introduce a matching model DAM, which is fully based on attention calculation. Specifically, each query and the corresponded candidate response are encoded with the self-attention mechanism from the transformer framework to learn their representations. Then, the cross-attention calculation is involved in capturing the word-level interaction features between the query and its response candidate, where the obtained interaction features are aggregated as a 3D matching tensor to compute the final matching score. Wang *et al.* (2019a) propose a novel strategy to extract correlation information between contexts and responses from multiple views, which is fully based on the dilated-convolution and self-attention. Since convolution and attention can be computed in parallel, this framework is much faster than previous methods that contain RNNs-based components. Then, another attention mechanism is used to aggregate utterance-response matching information across different turns within the dialogue context. Tao *et al.* (2019a) also leverage a bunch of representations at different granularities, including character-level, word-level, n-grams aggregated by CNNs, segment-level learned by RNNs, and attention-based global representations, to complete the multi-turn response extraction tasks. Besides multiple types of representations, they also point out that when and how to fuse these different representations does matter for the final matching performance. Through a thorough study on benchmark datasets, they conclude that fusion in the later stages of matching achieves better performance consistently than fusion in earlier steps.

Tao *et al.* (2019b) argue that single interaction in extracting the correlation features between context and response is not enough and presents an interaction-over-interaction network (IOI) that involves multiple stacked interactions. The upper interaction is iteratively launched upon the lower interaction results. As shown in Figure 3.6, each history utterance i interacts with the candidate response over L times with the same interaction block, where the t-th interaction is based on the $t-1$-th interaction. All interaction results of each utterance and block are aggregated to compute the final history-response matching score $g(c, r)$, where c and r refer to history utterances and response candidates. Yuan *et al.* (2019) investigate the side effect of utilizing unnecessary history

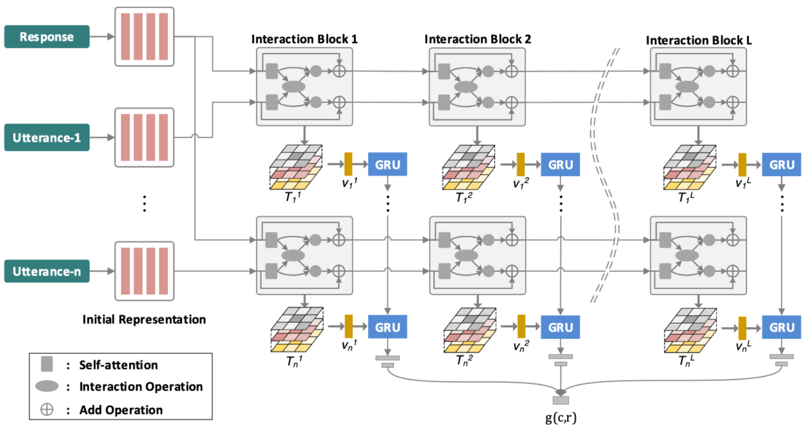

Figure 3.6: Architecture of interaction-over-interaction network (Tao *et al.*, 2019b).

utterances in matching and point out that the matching-based models are susceptible to the history. Accordingly, the authors present the MSN model to mitigate this issue, which utilizes a multi-hop selector based on attentive module (Zhou *et al.*, 2018d), word selector, utterance selector, and hop-k selector to choose the associated history utterances for calculating the history-response matching score.

3.3 Grounding with Extra Information

There are many types of grounding information in response selection. We group these works from the modality of extra information, i.e., text, image, and video. To better understand how grounding information is utilized in chit-chat response selection, we also include response-selection research for specific chatting tasks in this section, e.g., question answering based on given images and videos.

3.3.1 Document-Grounded Response Selection

Human conversations, in reality, are normally grounded on external knowledge. For example, users on Reddit usually launch a conversation based on the posted document at the beginning of a thread with topic information and other related facts. However, many chit-chat systems fail to leverage knowledge grounding information in producing conversations,

A's profile	trying new recipes makes me happy. i feel like i need to exercise more. i am an early bird , while my significant other is a night owl. i am a kitty owner.
B's profile	i might actually be a mermaid. i use all of my time for my education. i am very sociable and love those close to me. i enjoy swimming in the ocean , i feel in tune with its inhabitants.
Context	**A**: hi how are you today **B**: i am good . how are you ? **A**: pretty good where do you work ?
True response	i do not work , i am a full time student . what about you?
False response	i have been working as a salesman for more than 10 years.

Figure 3.7: The illustration of document-grounded response selection for multi-turn conversations (Zhao *et al.*, 2019b).

leading to a gap between human conversation behaviors and chit-chat systems. To fill this gap, Zhao *et al.* (2019b) propose to perform response selection grounded on easy-to-access unstructured documents, which are common sources of knowledge in application. Figure 3.7 gives an instance from the PERSONA-CHAT corpus (Zhang *et al.*, 2018b) that can illustrate the document-grounded response selection task, where a model is designed to pick up the ground-truth response from multiple pre-retrieved candidates conditioned on the given context and the correlated user profiles.

Response selection with grounded-document encounters many extra challenges. One is how to deal with irrelevant context utterances to the given documents, e.g., the greeting sentence in Figure 3.7). The other is how to avoid redundant information from the given documents, which may overwhelm the critical context information in selecting an appropriate response. Besides, existing matching models need to improve their capabilities in effectively capturing and using multi-source information. Zhao *et al.* (2019b) propose a document-grounded matching network (DGMN) which breaks down the matching process into four stages, namely encoding, fusion, matching, and aggregation. The encoding stage correlates with obtaining the vector representations of document sentences, context utterances, and response candidates by self-attention. Based on the obtained vector representations, an attention mechanism is utilized in the fusion stage to further compute the document-aware context representation and the context-aware document representation. Then, a novel distillation strategy and a hierarchical interaction mechanism are introduced to extract various matching information. The final stage aggregates these matching signals to obtain the matching score.

Since the conversation topic flow changes quickly in many chit-chat dialogue sessions, different contents in context and extra knowledge contribute unequally to selecting the ground-truth response from candidates, in which the irrelevant information might overwhelm the desired context and knowledge information, leading to poor performance. To solve this obstacle, Hua *et al.* (2020) design a model to calibrate irrelevant information in context and extra knowledge collection, which can enhance the performance of response selection. Since the conversation flow might change quickly across the whole session, not all context utterances and knowledge information are responsible for selecting the current response. The useless information will seriously interfere with the matching score calculation of the desired context and grounding information and response candidate. To overcome this problem, Hua *et al.* (2020) explore to dynamically extract the desired information from dialogue context and knowledge collection for response selection. Concretely, they first leverage the last context utterance to detect the relevant context information and document segments and then leverage the cross-attention strategy to calculate the correlated interaction matrices between pre-selected contents and response candidates. Next, a bidirectional LSTM model is involved in aggregating different types of matching features, including context-candidate and document-candidate, and another bidirectional LSTM is used to capture the temporal dependency information of context utterances. In selecting the relevant sentences from knowledge collection, the representation of the extracted context and the response candidate is involved and combined, and an attention mechanism is utilized for selecting relevant sentences.

A similar idea of selecting desired contexts and knowledge to improve response selection is also presented by Gu *et al.* (2020b). The model architecture is demonstrated in Figure 3.8. The model first utilizes a BiLSTM to encode both context, response, and knowledge. Then, the encoded representations are processed by a context filter to obtain knowledge-aware context representations and a knowledge filter to

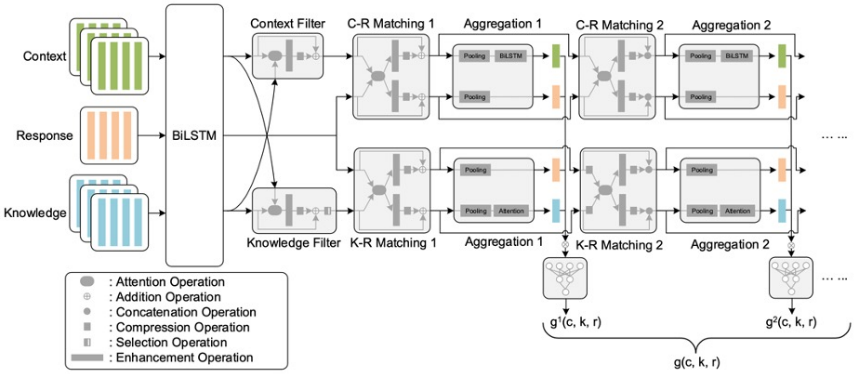

Figure 3.8: Structure of the Filtering before Iteratively Referring (FIRE) framework (Gu *et al.*, 2020b).

compute context-aware knowledge representations. Next, these two types of filtered representation, along with the encoded response, iteratively extract and aggregate matching signals multiple times for calculating the final score, which resembles the stacked IOI framework.

3.3.2 Image-Grounded Retrieval-Based Dialogue

Recently, lots of researchers have considered bringing the vision and language together, such as image caption (Xu *et al.*, 2015; Hendricks *et al.*, 2016) and visual question answering (VQA) (Antol *et al.*, 2015; Noh *et al.*, 2016; Yang *et al.*, 2016). However, it is still far away from the goal of developing general AI agents that can "see" (e.g., understanding their surroundings or social media content) and "communicate" (e.g., share their opinions with humans). Das *et al.* (2017a) make a step towards conversational visual AI and extend the scenario of visual question answering to visual dialogue. In visual dialogue, given an image, a dialogue history consisting of a sequence of questions and answers, and a follow-up unanswered natural language question, the machine is trained to produce an appropriate answer. Specifically, the machine should understand the dialogue context and extract essential clues from the image to pick a proper answer from a set of candidate answers. Figure 3.9 gives an example of the visual dialogue. The main

Caption: The skiers stood on top of the mountain.
Person A (1): how many skiers are there
Person B (1): hundreds
Person A (2): are they getting ready to go downhill
Person B (2): i think so my view is at end of line
Person A (3): is it snowing
Person B (3): no, there is lot of snow though
Person A (4): can you see anybody going downhill
Person B (4): no my view shows people going up small hill on skis i can't see what's going on from there
Person A (5): do you see lift
Person B (5): no
Person A (6): can you tell if they are male or female
Person B (6): skiers closest to me are male
Person A (7): are there any children
Person B (7): i don't see any but there could be it's huge crowd
Person A (8): does anybody have hat on
Person B (8): they all have winter hat of some sort on
Person A (9): is sun shining
Person B (9): yes, all blue sky
Person A (10): do you see any clouds
Person B (10): no clouds

Figure 3.9: An example of visual-grounded dialogue (Das *et al.*, 2017a).

challenges of visual dialogue include: (1) how to fuse multi-modal representations since the textual and visual features are always represented with different methods. For example, ResNet (He *et al.*, 2016), VGG Net (Simonyan and Zisserman, 2014) and Faster RCNN (Ren *et al.*, 2015) are widely used to extract the visual features of an image while the textual features are usually represented by word2vec (Mikolov *et al.*, 2013), Glove (Pennington *et al.*, 2014) and pre-trained language models (Devlin *et al.*, 2018); (2) how to model the complex interactions between image and dialogue; and (3) how to perform visual coreference resolution. For example, when encountering the word "they" in the second turn in Figure 3.9, the model needs to know that this refers to skiers in the previous turn and locate the skiers in the image to answer the question.

Most researchers build the model on the basis of encoder-decoder architecture, and these works can be categorized into three main groups: (1) fusion-based methods; (2) attention-based methods; and (3) visual coreference resolution methods.

Fusion-based Methods. Das *et al.* (2017a) propose three strategies to convert the model inputs (i.e., the dialogue history, the associated image, and the natural language question) into a joint representation.

First, Late Fusion (LF) exploits LSTMs to encode the entire dialogue history and the question, respectively and exploits the L2-normalized activations from VGG-16 (Simonyan and Zisserman, 2014) to represent the image. The representations of the three inputs are directly concatenated and transformed through a linear layer. Second, Hierarchical Recurrent Encoder (HRE) captures the intuition that there is a hierarchical structure in a dialogue history. Specifically, the question/answer is composed of a sequence of words, and the dialogue history consists of a sequence of question-answer pairs. To this end, HRE uses a recurrent block to embed the question at each turn and image jointly. The joint representation, together with the embedding of the current turn, are then fed to a dialogue-RNN to encode the global information. Third, a Memory Network (MN) Encoder utilizes a memory bank to store each previous question and the answer as "fact" and learns to answer the question with the stored facts and image. Guo *et al.* (2019a) further enhance previous one-stage fusion with a synergistic stage to deal with candidates that are hard to distinguish. In their proposed framework, the first stage is responsible for pre-retrieving or recalling hard samples close to the ground-truth answer. The second stage provides the target question with each pre-retrieved answer candidate, i.e., each candidate is filled in their question context, to extract more nuanced matching features. Then, these in-context question-answer candidate pairs interact with the given image and the correlated dialogue history to rank and select a proper answer.

Attention-based Methods. Recently, researchers have proposed various attention mechanisms which significantly promote the development of this field. Lu *et al.* (2017) propose a History-Conditioned Image Attentive Encoder (HCIAE), which utilizes an attention mechanism to performs coreference resolution by focusing on the important part of the dialogue history that might be helpful in answering the given question. The overall architecture is shown in Figure 3.10. In a nutshell, the final encoding is obtained through two steps, i.e., the encoder first employs the question to attend the history and then leverage the attended history and question to interact with the image. Specifically, given the spatial image features $\mathbf{V} \in \mathcal{R}^{d \times k}$, the embedding of question $\mathbf{m}_t^q \in \mathcal{R}^d$, and

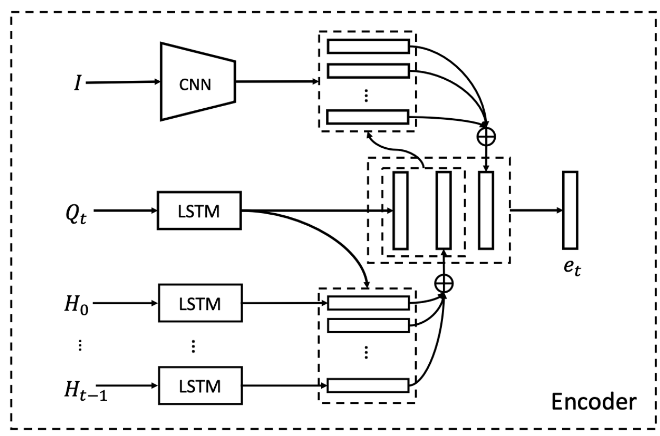

Figure 3.10: The model structure of HCIAE (Lu *et al.*, 2017).

the contextual representations of dialogue history $\mathbf{M}_t^h \in \mathcal{R}^{d \times t}$, the first step attention is implemented as:

$$\begin{aligned} \mathbf{z}_t^h &= \mathbf{w}_a^T \tanh(\mathbf{W}_h \mathbf{M}_t^h + (\mathbf{W}_q \mathbf{m}_t^q) \nVdash^T) \\ \alpha_t^h &= \operatorname{softmax}\left(\mathbf{z}_t^h\right) \end{aligned} \tag{3.2}$$

where $\nVdash \in \mathcal{R}^t$ presents a vector with each elements being 1. $\mathbf{W}_h, \mathbf{W}_q \in \mathcal{R}^{t \times d}$ and $\mathbf{w}_a \in \mathcal{R}^k$ refer to learnable parameters. The attended history is represented as the convex combination of $\mathbf{M}_t$ columns, denoted as $\hat{\mathbf{m}}_t^h$. The second step attention is implemented as:

$$e_t = \tanh\left(\mathbf{W}_e \left[\mathbf{m}_t^q, \hat{\mathbf{m}}_t^h, \hat{\mathbf{v}}_t\right]\right) \tag{3.3}$$

where $\mathbf{W}_e \in \mathcal{R}^{d \times 3d}$ are trainable parameters and $[\cdot]$ corresponds to the concatenation operation.

Considering that the image-grounded dialogue system involves multiple types and multi-modal components, e.g., question, dialogue history, and image, the correlations between each element are critical to producing suitable answers, either in the generative or retrieval paradigm. Guo *et al.* (2019b) propose to utilize different attention mechanisms in multiple stages to enhance the modeling of various correlations. As illustrated in 3.11, they first utilize the pre-trained VGG and RCNN to extract global and local features from the given image, respectively

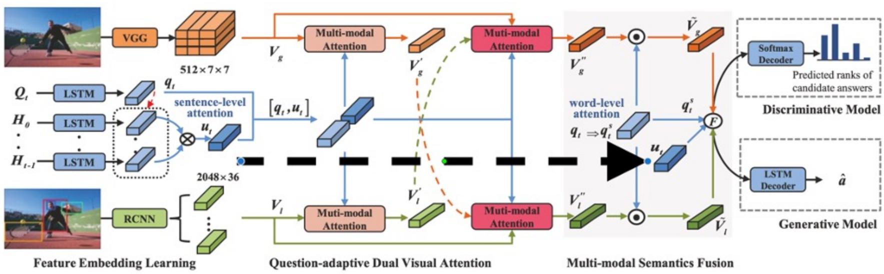

Figure 3.11: Structure of Dual Visual Attention Network (DVAN) for visual dialogue (Guo *et al.*, 2019b).

and introduce two LSTMs to learn the embedding representation of question and dialogue history. Then, three different attention operations are introduced to model the intra-textual, intra-visual, and cross-modal correlations: (1) leveraging the question to attend to the dialogue history; (2) using both the question and the attended history to attend to local image features from RCNN and global features from VGG; (3) utilizing the attended global and local features to mutually attend to each other. Next, the authors utilize the self-attention mechanism to capture word-level salient information from the question. These different types of features are then fused to produce an answer.

To mimic the human behaviors that one might revisit the image and dialogue context multiple times to capture adequate information for producing a suitable answer, Gan *et al.* (2019) enhance the previous one-pass glancing on a multi-modal context with a recurrent dual attention method to allow multiple-step reasoning. To further explore the effectiveness of attention mechanisms in visual dialogue, Park *et al.* (2020) propose a multi-view attention network, which breaks down the visual dialogue task into two sub-problems. First, the model constructs the question-guided contextual representation and collects the topic-related clues from the dialogue history. Second, the model performs multi-modal alignment between visual and textual representations through the sequential alignment process. To alleviate the modality-imbalanced problem, Kim *et al.* (2020) propose various consensus-dropout and ensemble methods to integrate the image-only and the image-history-joint model and achieve more balanced performance on all metrics.

Suganuma, Okatani, *et al.* (2020) propose a Light-weight Transformer for Many Inputs (LTMI) to cope with the expensive computational cost of directly applying the vanilla Transformer to model the many-to-many utility interactions.

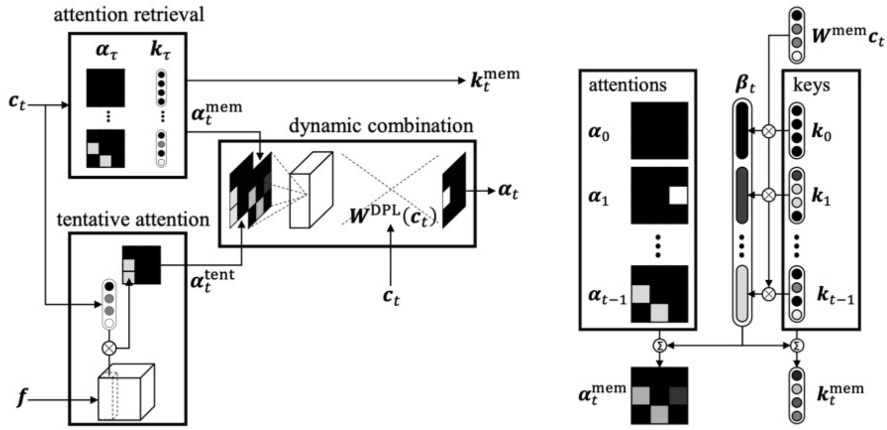

Figure 3.12: Structure of the AMEM framework (Seo *et al.*, 2017).

Visual Coreference Resolution Methods. Different from the VQA task that each question can be answered independently based on the grounded image and context (optional), the visual dialogue generation task involves a series of questions and answers with correlations, which demands a process that resembles the coreference resolution in NLP. To fill this gap, Seo *et al.* (2017) construct a visual coreference resolution task to model the dependencies between dialogue history and the current question. They propose to utilize a memory module to enhance widely-used attention calculation between the encoded (question, history) c_t and the feature representation of the image f_n with attention information from previous interaction turns, i.e., the attention matrices of previous (question, history) towards the given image. As shown in figure 3.12, the proposed AMEM framework consists of two different types of attention matrices, i.e., the tentative attention and retrieved attention, and a dynamic combination strategy. The tentative attention matrix is computed to extract salient signals from the feature maps of the encoded image based on the current question and history, formulated as:

$$\begin{aligned} s_{t,n} &= (\mathbf{W}_c^{\text{tent}}\mathbf{c}_t)^\top(\mathbf{W}_f^{\text{tent}}\mathbf{f}_n) \\ \alpha_t^{\text{tent}} &= \text{softmax}(\{s_{t,n}, 1 < n < N\}), \end{aligned} \tag{3.4}$$

where $\mathbf{W}_c^{\text{tent}}$ and $\mathbf{W}_f^{\text{tent}}$ are learnable parameters, and $s_{t,n}$ represents an attention weight for an encoded image feature. Then, a memory module that stores previous attention matrices is introduced to enhance the interaction between the current question and image with more necessary information. Specifically, given the attention memory $\mathbf{M}_t = \{(\alpha_0, \mathbf{k}_0), (\alpha_1, \mathbf{k}_1), \ldots, (\alpha_{t-1}, \mathbf{k}_{t-1})\}$ with the previous attention matrices α_τ and the correlated keys $\mathbf{k}_\tau$, the memory reading process is denoted as:

$$\begin{aligned} m_{t,\tau} &= (\mathbf{W}^{\text{mem}}\mathbf{c}_t)^\top\mathbf{k}_\tau, \\ \beta_t &= \text{softmax}(\{m_{t,\tau}, 0 < \tau < t-1\}), \end{aligned} \tag{3.5}$$

where $\mathbf{W}^{\text{mem}}$ are learnable parameters. The relevant attention α_t^{mem} is defined as $\alpha_t^{\text{mem}} = \sum_{\tau=0}^{t-1} \beta_{t,\tau}\alpha_\tau$. Then a dynamic combination strategy and a single-layer MLP are presented to compute the final attention matrix α_t)the , conditioned on c_t. These attended visual features, attention matrices, encoded texts, and the keys can be further fused for either answer selection or generation.

To make the visual coreference resolution process more interpretable, Kottur *et al.* (2018) propose incorporating neural module networks (NMNs) into the visual dialogue framework to achieve explicit visual resolution at the word or phrase level. In addition to the neural modules designed for VQA, the authors also introduce three modules to handle visual dialogue, including *Not*, *Refer* and *Exclude*. The goal of *Refer* module is to first locate the entity that needs to be resolved and then link it to the reference entity in the coreference pool with the correlated visual object groundings. Specifically, given the reference pool $P_{\text{ref}} = \{(x_p^{(i)}, a_p^{(i)})\}_i$ which stores all previous entities in the dialogue, the

Refer module uses the text embedding x_{txt} to attend to the reference pool, and this procedure is implemented as follows:

$$\begin{aligned} s_i &= \text{MLP}([x_{\text{txt}}, x_p^{(i)}, \Delta_i t]), \\ \tilde{s}_i &= \text{Softmax}(s_i), \\ a_{\text{out}} &= \sum_{i=1}^{|P_{\text{ref}}|} \tilde{s}_i a_p^{(i)}, \end{aligned} \tag{3.6}$$

where $\Delta_i t$ denotes the distance that a candidate entity in the current round cross over its first occurrence. The *Not* module is presented to process the image region that is not attended. The *Exclude* is introduced to address the special case like "What other red things are present?". Specifically, *Exclude* module invokes *Not* and two neural modules in VQA, i.e., *Find* and *And* to complete this action:

$$y = \text{And}[\text{Find}[x_{\text{txt}}, x_{\text{vis}}], \text{Not}[a]], \tag{3.7}$$

where the *Find* module returns all objects that are correlated to the given textual representation from the image, the *Not* module focuses on image regions that are not attended. *And* module is utilized to combine the outputs of the aforementioned two modules.

Inspired by how humans complete the coreference resolution, Niu *et al.* (2019) propose an attention mechanism, named RvA, to recursively look up the dialogue history for the coreference resolution problem and meanwhile update the visual attention accordingly. As exemplified by the example in Figure 3.13, when the dialogue system encounters a question that is expressed with ambiguity (e.g., "Are they on or off?"), it will recursively look up the dialogue history and update the correlated visual attention until the visual coreference is addressed (e.g., "How many lamps are there?").

Similarly, Kang *et al.* (2019a) assume that humans break down the visual reference resolution into two steps: (1) linguistically resolve the ambiguous questions by recalling the dialogue history from one's memory and (2) find a local region of a given image for the resolved questions. The authors further propose the Dual Attention Networks (DAN), which comprises two different attentions, namely *REFER* and *FIND*, corresponding to the aforementioned two steps.

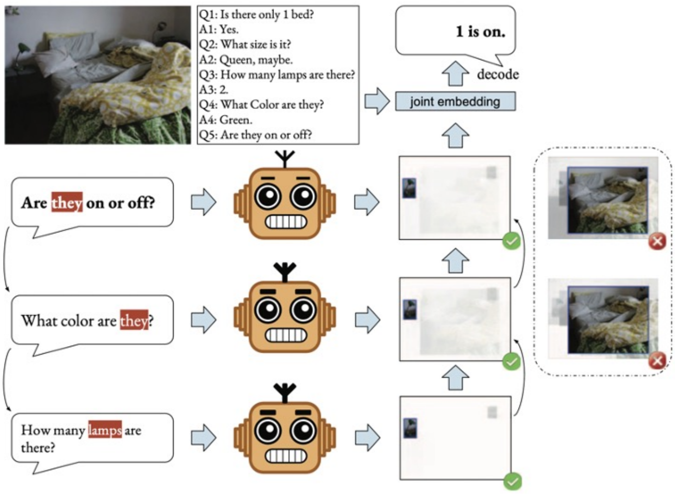

Figure 3.13: An example that illustrates the RvA in visual dialogue (Niu *et al.*, 2019).

First, the *REFER* module aims to attend to the most relevant proportion of dialogue history with respect to the given question. Given the encoding of question q_t and the representations of dialogue history $M_t = \{h_i\}_{t=0}^{t-1}$, DAN utilizes multi-head attention to calculate the importance of different proportions:

$$\begin{aligned} \text{head}_n &= \text{Attention}(q_t W_n^q, M_t W_n^m), \\ \text{Attention}(a, b) &= \text{softmax}(\frac{ab^\top}{\sqrt{d_{\text{ref}}}})b, \end{aligned} \tag{3.8}$$

where $W_n^q \in \mathbb{R}^{L \times d_{\text{ref}}}$ and $W_n^m \in \mathbb{R}^{L \times d_{\text{ref}}}$ are learnable parameters, d_{ref} denotes the dimension of the latent space. The multi-head representation x_t is defined as $x_t = (\text{head}_1 \oplus \cdots \oplus \text{head}_h)W^o$, where W^o is the learnable parameters and $\oplus$ is the concatenation operation. The output of the *REFER* module e_t^{ref} is then defined as:

$$\begin{aligned} c_t &= \text{ReLU}(\hat{x}_t W_1^f + b_1^f)W_2^f + b_2^f, \\ \hat{c}_t &= \text{LayerNorm}(c_t + \hat{x}_t), \\ e_t^{\text{ref}} &= \hat{c}_t \oplus q_t, \end{aligned} \tag{3.9}$$

where $\hat{x}_t = \text{LayerNorm}(x_t + q_t)$, W_1^f, b_1^f, W_2^f, and b_2^f are learnable parameters. The *FIND* module employs the bottom-up attention mechanism to obtain the visual attention weights:

$$\begin{aligned} r_t &= f_v(v) \odot f_{\text{ref}}(e_t^{\text{ref}}), \\ \alpha_t &= \text{softmax}(r_t W^r + b^r) \end{aligned} \tag{3.10}$$

where W^r and b^r are learnable parameters, $f_v(\cdot)$ and $f_{\text{ref}}(\cdot)$ are MLPs. The attention weights will be further used to compute the vision-language joint representations as follows:

$$\begin{aligned} \hat{v}_t &= \sum_{j=1}^{K} \alpha_{t,j} v_j \\ z_t &= f'_v(\hat{v}_t) \odot f'_{\text{ref}}(e_t^{\text{ref}}) \\ e_t^{\text{find}} &= z_t W^z + b^z, \end{aligned} \tag{3.11}$$

where W^z and b^z are learnable parameters, $f'_v(\cdot)$ and $f'_{\text{ref}}(\cdot)$ are MLPs. The output e_t^{find} is then used to score the candidate answers.

Other Approaches. In this section, we briefly summarize other approaches used to solve the visual dialogue problem. These methods can be roughly grouped into the following categories: (1) employing the graph neural networks (Schwartz *et al.*, 2019; Zheng *et al.*, 2019b; Guo *et al.*, 2020) to have a better understanding of the semantic dependencies in visual and contexts; (2) optimizing the dialogue policy with reinforcement learning methods (Das *et al.*, 2017b; Yang *et al.*, 2019b); (3) pre-training the vision-language transformer on a large multi-modal corpus and transferring to visual dialogue (Murahari *et al.*, 2020); and (4) initializing the encoder with BERT and adopt visually grounded masked language modeling (MLM) and next sentence prediction (NSP) objectives to optimize the model (Wang *et al.*, 2020b).

3.3.3 Video-Grounded Retrieval-Based Dialogue

Many potential applications of conversational machines would benefit significantly from comprehending the scene in which the system is applied. However, visual dialogue only involves conversing about a static image which is inherently limited. To this end, Alamri *et al.*

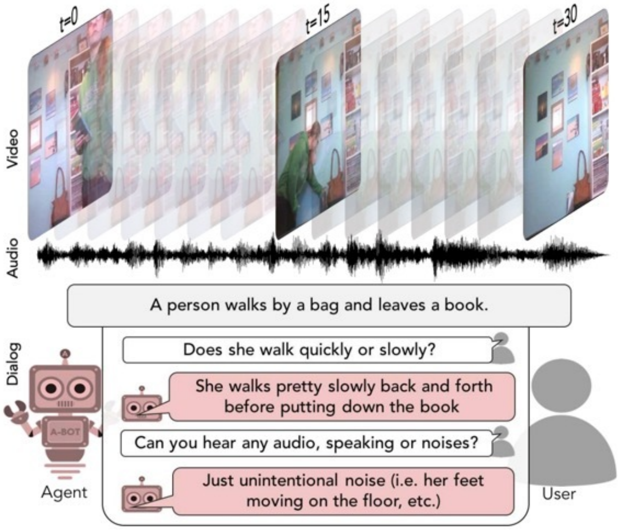

Figure 3.14: The illustration of Audio Visual Scene-Aware Dialogue (Alamri *et al.*, 2019).

(2019) make one step further by proposing the task of scene-aware dialogue. Figure 3.14 depicts a conversation about a temporally varying scene that is carried on between an agent and a human. To answer such questions, the agent not only needs to understand the visual scene holistically but also be aware of the audio information.

Formally, given a video, the correlated dialogue history consisting of previous question-answer pairs, the video caption, and the current question, the system's goal is to select a proper response from a set of candidate answers. To have a more intuitive observation about the impact of different input information, Alamri *et al.* (2019) propose a late-fusion approach for video-grounded dialogue. The video frames and the audio track are independently transformed into a fixed-sized vector through convolutional neural networks (CNNs). The words in the dialogue history of the current turn are concatenated to form a long sequence, which is then fed to an LSTM to generate the contextual representation. The encoding of the question is implemented in analogy to the encoding of the dialogue history. The four vectors are then concatenated to rank the candidate answers.

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Figure 3.15: An example of persona-based dialogue (Zhang *et al.*, 2018b).

3.4 Human Factors: Emotion, Persona and Beyond

With the development of dialogue systems, more attention has been paid to building empathetic and more human-like chit-chat systems through explicitly modeling human factors such as persona and emotion, where persona refers to various user profile information e.g., name, age, knowledge or expertise, wording behaviours. Though more attempts were made in generative-based dialogue systems as will be discussed in the next section, there are various works that incorporate human factors into retrieval-based chit-chat systems.

Naturally, a human-like chit-chat system should display a consistent personality during its chatting with humans. Recently, an increasing number of works pursue to design retrieval-based chit-chat systems grounding on persona information. Zhang *et al.* (2018b) propose grounding dialogue systems on persona information to make chit-chat more engaging. They firstly construct a new dataset named PERSONA-CHAT, where each human-human dialogue is paired with several sentences that describe the personal information of both speakers. Figure 3.15 presents an example of the PERSONA-CHAT. They further design the profile memory network, which considers the dialogue history as input and then performs attention over the persona to combine with the dialogue history. Mazare *et al.* (2018) collect a much larger personalized dialogue corpus from Reddit[1] and implement a persona-based response selection network similar to Zhang *et al.* (2018b). Gu *et al.* (2019a) then propose the dually interactive matching network to better incorporate persona descriptions into multi-turn response selection. Specifically, they utilize a dual matching architecture to perform interactive matching on context-response and persona-response pairs, respectively, to select the target response from candidates.

Li *et al.* (2021a) highlight the effectiveness of user dialogue histories and construct two personalized dialogue corpora where each dialogue case is paired with the dialogue histories of both users. They further design a personalized hybrid matching network which incorporates user dialogue histories into hybrid representation matching between dialogue context and the response by personalized attention and wording behavior modeling. Zhong *et al.* (2020) collect a dialogue corpus from two subreddits whose conversations are considered more empathetic than that of others. They propose a BERT-based response selection model which employs co-attention between the candidate response and the persona information as well as the dialogue context based on the representations calculated by BERT.

Beyond persona information, researchers have been starting to explore the effect of emotion in response selection to build more empathetic chatbots recently. Lubis *et al.* (2019) point out the necessity of positive emotion in responding to humans. Stepping from this point, they first propose a response retrieval method for positive emotion elicitation. Then, with the retrieval model's help, they construct a corpus by replacing dialogue responses with those that can lead to positive emotions in human-machine interactions. Qiu *et al.* (2020) incorporate emotions into retrieval-based dialogue systems from two different perspectives: 1) they utilize a transition network to track the underlined emotion flow aside from dialogue history modeling and leverage the captured intrinsic emotion information to augment the context-response matching, 2) they present flexible mechanisms to perform emotional controlling in chatbots.

[1] https://www.reddit.com/r/datasets/comments/3bxlg7/ (date accessed: 11 April 2022)

3.5 Pre-Training in Dialogue Retrieval Models

Recent years have witnessed a growing number of large-scaled pre-trained language models such as BERT (Devlin *et al.*, 2018), RoBERTa (Liu *et al.*, 2019b), XLNet (Yang *et al.*, 2019c). Based on the deeply-stacked self-attention architecture (Vaswani *et al.*, 2017), these models are trained with self-supervised objectives on large-scale open data to become powerful feature extractors that can not only provide contextual representations but also be fine-tuned to specific down-stream tasks without additional parameters except an MLP layer. The idea of self-supervised pre-training inspires a growing number of studies of retrieval dialogue systems. The core motivation of works in this section is to find the nature of human dialogues, which makes dialogue modeling distinctive from other NLP tasks that can guide the design of model architectures, input formulations, and self-supervised objectives. Among them, some works share the spirits of pre-training and develop their own model architecture (Tao *et al.*, 2021a) and training objectives (Tao *et al.*, 2020), while others build response selection models upon some released pre-trained language models and even move a step further to exploit dialogue specific self-supervised objectives to pre-train or post-train a dialogue model that is tailored for multi-turn response selection.

Tao *et al.* (2020) introduce ECMo (i.e., Embedding from a Conversation Model) to provide contextualized representations that are tailored for dialogue modeling for the multi-turn response selection task. They pre-train a large hierarchical encoder-decoder dialogue generation model on a large dialogue corpus to better model the multi-turn context of human dialogues. The generation model can provide word-level and sentence-level contextualized dialogue representations that can be merged in the input and the output layer of the matching model, respectively, which is similar to ELMo (Peters *et al.*, 2018).

Wolf *et al.* (2019) borrow the idea of transfer learning and adapt GPT (Radford *et al.*, 2018b) to response generation and response selection tasks through fine-tuning with the language modeling loss and the next utterance classification loss. The proposed model achieved remarkable results in the response selection.

More recently, researchers mainly adopt BERT (Devlin *et al.*, 2018)

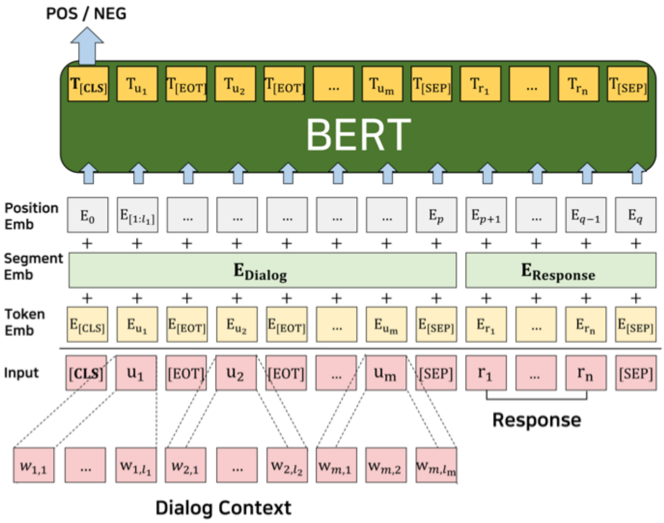

Figure 3.16: An example of BERT-based response selection (Whang *et al.*, 2019).

as the backbone of their response selection model. There are several obstacles that prevent the direct adaptation of BERT to the multi-turn response selection task. Firstly, although BERT is skilled at performing text pair classification tasks, the input formulation of multi-turn response selection is a bit different from that of typical text-pair classification tasks such as NLI, where the inputs are two sentences that are concatenated together with a special separator token. In fact, the inputs of the multi-turn response selection task consist of two parts, one is the dialogue context which is comprised of multiple utterances, and a candidate response which is an individual utterance. Next, the domain of the pre-training corpora (i.e., English Wikipedia and Book Corpus) is usually quite distant from the downstream dialogue corpus, especially when the dialogue is domain-specific (i.e., the Ubuntu Corpus (Lowe *et al.*, 2015)). Moreover, the pre-training objectives (i.e., masked language modeling and next sentence prediction used in BERT) are not tailored for multi-turn response selection. In this research line, various works explore to introduce dialogue-specific features or design training objectives that are tailored for multi-turn response selection to tackle the aforementioned problems.

As for the incorporation of dialogue-specific features, Whang *et al.* (2019) firstly adopted BERT to multi-turn response selection task and proposed an effective domain adaptive post-training method to improve the model performance. The proposed BERT-VFT model was firstly post-trained on the task-specific corpora using masked language modeling and next sentence prediction to equip the model with specific domain knowledge. Moreover, to mitigate the gap of input formulation between typical sentence-pair matching and multi-turn response selection, BERT-VFT made some modifications on BERT, as illustrated in Figure 3.16. Specifically, as the dialogue contexts are comprised of multiple utterances, they injected an additional special token EOT (i.e., end of turn) between dialogue utterances in the input context. The two proposed techniques were followed by most of the works along this line. Similarly, Gu *et al.* (2020a) proposed SA-BERT, which additionally introduces the speaker embedding that models the speaker changes information in a dialogue flow. Moreover, SA-BERT also employs a speaker-aware disentanglement strategy to filter out a small portion of important utterances from the context.

Meanwhile, in addition to the original objectives (i.e., masked language modeling and next sentence prediction) used in Whang *et al.* (2019) and Gu *et al.* (2020a) during post-training, various works concentrate on designing new self-supervised objectives (which can also be considered as data augmentation strategies) that are tailored for multi-turn response selection to post train the pre-trained language model. Whang *et al.* (2020) propose three utterance manipulation strategies (i.e., insertion, deletion, and search) to aid the response selection model towards maintaining dialogue coherence. In insertion, an utterance is randomly extracted from consecutive utterances in the context, and the model is trained to predict the insertion position. In deletion, the goal for the model is to find the irrelevant utterance that is inserted into the original context. As for search, the model learns to select the true previous utterances from the shuffled context utterances. Lu *et al.* (2020) propose to augment the dialogue corpus for fine-tuning the pre-trained language models by treating the last utterance of consecutive utterances in a session as the positive response while considering a randomly chosen utterance in the same session as the negative response. Xu *et al.* (2020c) proposed a more systematic method to incorporate the self-supervised objectives. Specifically, they introduced four self-supervised tasks, which include next session prediction, utterance restoration, incoherence detection, and consistency discrimination, to post-train a pre-trained language model for multi-turn response selection in a multi-task manner. Li *et al.* (2020a) introduced task-specific pre-training to bridge the gap between pre-training and fine-tuning. They constructed the dialogue-related corpora based on four medium-sized prevailing dialogue corpora. They generated three negative examples through utterance ordering, utterance insertion, and utterance replacement for each positive example, which aims to provide negative examples lacking readability, fluency, and coherence. They further score examples with the n-gram Normalized Inverse Document Frequency and train the model using the mean-square error (MSE) objective.

3.6 Evaluation

Most of the research reviewed in this section focuses on the task of response selection (context-response matching). Thus, the model performance evaluation lies on how to automatically calibrate the model capability of retrieving a suitable response, i.e., the ground-truth response is expected to be selected or obtain a higher matching score. There are four metrics that are widely used (Zhang *et al.*, 2018b; Zhao *et al.*, 2019b; Alamri *et al.*, 2019), i.e., **mean rank (MR)**, **recall@k**, **mean reciprocal rank (MRR)**, and **normalized discounted cumulative gain (NDCG)** (Das *et al.*, 2017a). Except for these straight-forward evaluation metrics, retrieval-based chit-chat systems also requires more evaluation methods to measure the **efficiency of the response selection system**.

3.7 Summary

This section presents most of the representative deep neural response selection model for building chit-chat systems, which is categorized from the types of context information utilized in calculating context-response matching scores, including multi-turn history modeling, document-grounding, visual information utilization, persona, and emotion. We also briefly review some typical approaches based on pre-trained language models. We can conclude that most of the recent neural response retrieval methods focus on (1) **the utilization of context information**, (2) **how to effectively model context information**, and (3)**how to efficiently model context information**. With the prevalence of large-scale PLMs, the model efficiency of retrieval-based chit-chat systems will become the core challenge in the near future.

4

Generation-Based Chit-Chat Systems

Unlike retrieval-based chit-chat systems, generation-based methods can create new responses for a specific query given the correlated dialogue context. From the perspective of IR, generation-based methods can be alternative solutions when retrieval-based models fail. For instance, there is no related response in previous conversations, or the users prefer a personalized response. Moreover, generation methods also serve a vital role in the ensemble-based chit-chat system, which will be discussed in the next section. This section presents necessary information for the IR community to either build an ensemble chit-chat system or construct an alternative model when retrieval-based systems fail.

We start this section with a brief overview that mainly clarifies the difference of generation-based approaches and previously mentioned retrieval-based solutions and reveals the development trend of chit-chat systems. We then summarize and simply introduce the widely utilized sequence to sequence generation frameworks in the age of deep learning. After this, we explore several essential challenges and research topics, including tackling the one-to-diversity issue, context modeling, knowledge and grounding in response generation, human factors, and response generation methods based on the booming pre-trained language models. At the end of this section, we provide a preliminary study on the most challenging problem of chit-chat conversation generation, i.e., performance evaluation, and provide some open available data resources for building generation-based chit-chat systems.

4.1 The Paradigm of Generation: A Rising Trend

As stated in section 3, retrieval-based conversation systems can obtain real-world responses from existing conversation records, which is more accessible for humans. For dialogue context that resembles existing data, it is cheap and safe to retrieve a proper response since deep neural networks can capture the correlations and matching degree between dialogue context and pre-retrieved response candidates. Besides, training a deep model that can perform context-response matching is more data-efficient than estimating the joint probability distribution in conditional language models in generation-based conversation systems, which is more applicable for scarce data scenarios such as task-oriented systems. However, retrieval-based methods still have their limitations, one of which is that there is no existing conversation history that can accord with the given dialogue context. Another shortage is that retrieving responses from existing conversations cannot handle the notable problem of chit-chat conversations, i.e., one-to-diversity modeling, where the response for a specific conversation context varies according to many extra conditions such as persona, expression behaviors, emotion change, etc. Besides, it is also challenging to fuse multiple resources and extra information in the obtained response.

In contrast, generation-based models, by virtue of their encoder-decoder framework and capabilities in estimating the joint probability distribution of languages, can create new responses for unseen dialogue context in historical conversation data. Beyond that, encoder-decoder frameworks are more efficient for fusing different information to generate better responses. It is also feasible to achieve controllable response generation based on different conditions, e.g., persona, emotion preference, by modifying the conditional probability distribution objective, and thus generation-based methods can handle the one-to-diversity problem well for chit-chat conversations. The downside is that generation-based methods require large-scale conversation history for model training, especially for the chit-chat conversations, and more computational resources to build commensurate models are desirable accordingly. Another possible risk is caused by the widely-used generation training objective of maximizing the log-likelihood of utterances in the training data, which will encourage generation-based chit-chat systems to produce either short responses or responses with commonly co-occurred wordings, and none of which is rational for real-world chit-chat systems.

Although generation-based chit-chat systems exist various kinds of problems, learning to converse by generating a response shows a clear ascending trend. This upward trend is due to many different reasons. One of the critical reasons is the explosion of open-available conversation records. Another contributing factor is the rapid growth of computing capacity, and it is even applicable to run models such as BERT-base on mobile devices. With massive data and powerful GPUs as well as parallel algorithms, it is not brain surgery to train a large-size response generation model that performs well for chit-chat conversations. It is even favorable to pre-train large-scale language models for building chit-chat systems, e.g., DialogGPT. These pre-trained models, by virtue of their great fitting capability and impressive model capability, can not only generalize well across different topics, domains, and even languages but also have a certain degree of ability to handle few-shot/zero-shot problems. Besides, these pre-trained models can learn commonsense knowledge with language modeling jointly. Thus, building chit-chat systems on these pre-trained language models can create informative and fluent responses for new topics or unseen dialogue contexts, which will prevent some negative issues, e.g., resulting in dialogue breakdown and poor user experience. Another character is that it is convenient to jointly learn language generation and other influence factors, e.g., commonsense knowledge, grounding information, personalized demand.

As for the IR community, more powerful generated-based methods can also facilitate retrieval-based chit-chat systems. For one thing, generation-based methods could produce multiple alternatives for a user-issued query to enhance the response matching process. For another, generation-based methods could fuse multiple retrieved response candidates and edit obtained response candidates with different requirements, e.g., polishing irrelevant information, injecting personalized factors. Based on the above situation, generation-based methods are not only popular for other research communities but also for the IR field.

4.2 Overall Framework: Architecture and Challenges

One of the important reasons for the flourishing of chit-chat systems is the significant progress of sequence to sequence neural frameworks. Starting from Sutskever *et al.* (2014), various sequence to sequence frameworks are devised. Bahdanau *et al.* (2015) further leverage a context attention mechanism to enhance the original sequence to sequence model, consisting of a recurrent-based encoder and decoder. Gehring *et al.* (2017) propose a novel sequence to sequence framework entirely based on convolutional neural units, which allows high effective parallel training and long-context modeling. Shortly afterward, Vaswani *et al.* (2017) propose the Transformer model fully based on attention computation, which has the merits of both parallel training and dynamic context window modeling, as well as long-range correlations capturing. Since then, Transformer has become the most popular and widely acknowledged framework for sequence to sequence modeling.

Generation-based chit-chat systems formulate conversation as a sequence to sequence task to leverage the powerful sequence to sequence modeling capabilities of these frameworks. Given a conversation input $\boldsymbol{x}$ which is an user-issued query with/without dialogue context, target output response $\boldsymbol{y}$, and other conditions ($\boldsymbol{C}$) such as knowledge ($\boldsymbol{k}$), persona ($\boldsymbol{p}$), grounding ($\boldsymbol{g}$), the generation-based chit-chat task can be formulated as learning the mapping function $f(\cdot)$ to capture the corrections between input $\boldsymbol{x}$, conversation condition $\boldsymbol{C}$, and output $\boldsymbol{y}$, which is implemented as maximizing the following objective:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{C}) = \prod_{t=1}^{T_y} p(y_t|c, y_1, \cdots, y_{t-1}) \tag{4.1}$$

where T_y is the length of the target response y, and $\boldsymbol{C} = (\boldsymbol{k}, \boldsymbol{p}, \boldsymbol{g})$[1]. $p(\cdot)$ can be parameterized as deep neural networks. In doing so, it can be achieved to automatically create a response for given dialogue context and other conditional information.

4.3 Tackling the One-to-Diversity Issue

In modeling chit-chat conversation generation, one of the notorious challenges is "one-to-diversity". For a given dialogue history or context, there might be many feasible responses to fill the requirement and change the direction of the conversation, e.g., responding from different aspects of the given conversation history, starting a new topic along with a dialogue breakdown, conversing with the same semantic but different language expressions. The reasons behind this problem are in various aspects. For one thing, the open-available data is still limited compared with all possible conversations and one-to-diversity correlations, as a result of which data-driven methods built upon these data, not surprisingly, fail to model the one-to-diversity problem. For another, existing widely-used sequence to sequence mapping frameworks, e.g., aforementioned Seq2seq model, ConvSeq2seq, and Transformer along with the cross-entropy training objective, are insufficient to address the one-to-diversity correlation learning. Moreover, typical generation-based conversation systems ignore conversational behaviors of humans in producing one-to-diversity dialogues, where individuals will respond by combing dialogue context, commonsense, human factors, and other extra information.

Figure 4.1 presents a few examples that reflect the one-to-diversity challenge for chat-driven conversations. Take the user input "What are you doing?" for example, there are multiple feasible responses, and these responses consist of diversified words.

To address these challenges, existing work mainly explore data manipulation, new generation framework and pipeline, effective training objective, and leveraging extra resources. Since data serves as the core of building conversation systems in the era of the deep neural network,

[1]Other conditions that can affect model outputs or alter the target response are also desirable.

Input: What are you doing?	
1. I've been looking for you.	4. I told you to shut up.
2. I want to talk to you.	5. Get out of here.
3. Just making sure you're OK.	6. I'm looking for a doctor.
Input: What is your name?	
1. Blue!	4. Daniel.
2. Peter.	5. My name is John.
3. Tyler.	6. My name is Robert.
Input: How old are you?	
1. Twenty-eight.	4. Five.
2. Twenty-four.	5. 15.
3. Long.	6. Eight.

Figure 4.1: Conversation cases that illustrate the one-to-diversity problem (Li *et al.*, 2016a).

we will first elaborate on recent findings on data manipulation, including data augmentation and data selection. In the after part, we present the representative framework and pipeline as well as effective objectives designed for addressing the one-to-diversity problem, respectively. In the end, we discuss how to leverage extra resources in mitigating the one-to-diversity problem. Note that this section mainly focuses on the pros and cons of existing research from the one-to-diversity aspect. For instance, modeling human factors and commonsense knowledge are two essential topics of conversation systems, but we only focus on their characteristics in addressing the one-to-diversity issue in this section, where more research on knowledge grounding and human factors will be discussed in Section 4.5 and 4.6. In this section, we refer to one-to-diversity as both the expression diversity and one-to-many correlations.

4.3.1 Data Manipulation: Augmentation and Selection

It is widely acknowledged that data is the King during the time of deep neural networks, i.e., the performance of chit-chat systems based on deep neural networks is bounded by the given training data. In other words, most problems of existing models can be directly addressed or mitigated from the perspective of data manipulation. Herein, data manipulation mainly refers to widely-used data augmentation and selection methods. Recall that our target of introducing data manipulation is to address the one-to-diversity problem in chit-chat dialogues, including expression diversity and one-to-many modeling. In the following, we will elaborate on how to utilize data manipulation to solve the above-mentioned two challenges, respectively.

Among which, expression diversity calibrates the word frequency and co-occurrences, where low expression diversity means the generated responses consist of high-frequency and commonly co-occurred words, and in turn, high expression diversity correlates with more informative words and in-context contents. In most cases, the expression diversity of automatically generated responses is well below human beings, e.g., the ground-truth conversations, and thus data manipulation methods that can improve expressions diversity are sorely needed. For low-resource scenarios, it is more preferable to perform data augmentation by creating more data with low-frequency and rarely co-occurred words so that the augmented data could prevent the deep neural models from trapping into some high-frequency words and lexical combinations. For the high-resource setting, various data selection strategies are commonly utilized by filtrating out most of the data samples with high-frequent and non-informative words to re-balance the data distribution. In doing so, we can improve both the expression diversity of neural chit-chat systems and the data efficiency of model training. In real-world applications, it would be better to simultaneously introduce both data augmentation and selection strategies, either by first augmenting and then selecting or augmenting one part and selecting another part.

For the one-to-many diversity issue, the most straightforward solution is to conduct data augmentation by creating multiple responses for each given conversation context. Another possible direction of obtaining more one-to-many labeled data is to correlate each dialogue context with multiple responses that have similar context information. Moreover, it is also worth considering the many-to-one problem in existing data, i.e., one response can fit multiple dialogue contexts. Filtering out these labeled data with common responses can also mitigate the challenge of one-to-many modeling. Same as data manipulation for expression diversity, these methods are mixed-used in practice.

Following the data manipulation line, Csáky *et al.* (2019) propose to improve the neural conversational model with entropy-based data filtering to output more diverse responses. Li *et al.* (2019) introduce a CVAE-GAN framework to perform data augmentation for chit-chat conversations. Through utilizing the augmented data during training, the response generation model can produce more expressive and diverse responses. Zhang *et al.* (2020c) provide a new framework of data augmentation for chit-chat conversation systems. Through designing a data-level distillation process, a ranking module, and a model-level distillation process, the proposed solution can better utilize massive unpaired data and eventually can create dialogues with diversified contents. Cai *et al.* (2020b) conduct data manipulation for better data instance training of neural dialogue generation models by learning to augment and re-weight, and its effectiveness in improving response diversity is confirmed in experiments.

In short, data manipulation methods aim at changing the distribution of training data along with the powerful fitting capability of deep neural networks to alter the model outputs from the data side. Below, we will review existing research on designing new generation frameworks, training with effective objectives, and leveraging extra resources to achieve expression diversity and one-to-many modeling.

4.3.2 Generation Framework and Pipeline

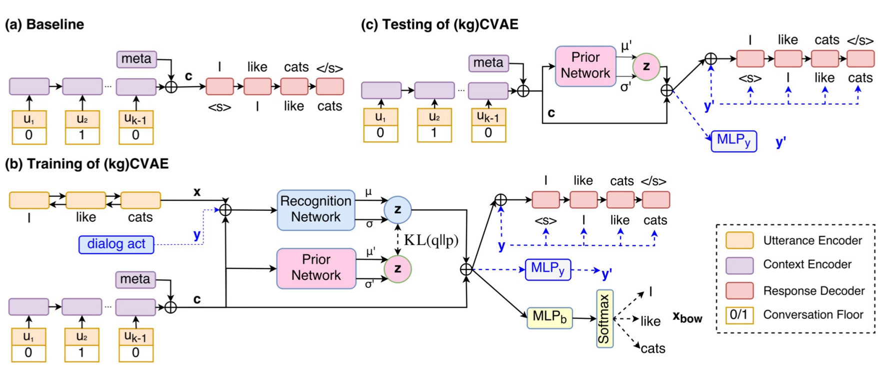

Figure 4.2: Representative neural network architecture of CVAE for chit-chat conversations (Zhao *et al.*, 2017).

One of the commonly used frameworks for improving expression diversity in chit-chat systems is Conditional Variational Auto-Encoder(CVAE). Through introducing a latent variable, these models can capture sentence-level or session-level global information and thus can prevent them from trapping into local language model (decoder) with high-frequency co-occurred words generated. For chit-chat systems, Cao and Clark (2017), Zhao *et al.* (2017), and Shen *et al.* (2017) first leverage the merits of CVAE models to generate responses with expression diversity. We take the CVAE architecture of Zhao *et al.* (2017) for illustration, as shown in Figure 4.2. Meta information and context utterances (u_1, u_2, u_{k-1}) are combined as the condition c. Given the condition c, the training target of this model is to re-construct the target response x, which resembles with Auto-Encoder. During testing, the CVAE model takes condition c as input and can produce multiple different responses for the same condition.

By virtue of its strength of generating multiple different responses for a given context, i.e., one-to-many diversity, many variants of CVAE models for chit-chat conversation are proposed later on. Park *et al.* (2018) propose two strategies to solve the degeneration problem of VAE-based conversation models, including using latent variables with a hierarchical structure and exploiting an utterance drop regularization. Xu *et al.* (2018a) incorporate the CVAE model with the measure of coherence and a context gate to achieve better conversations. Du *et al.* (2018) enhance CVAE-based conversation models with a sequence of latent variables to achieve high variability in responses and further introduce a backward recurrent neural network to augment the approximate posteriors for capturing long-distance information from the future tokens during generation. Unlike the vanilla CVAE model that only uses a latent variable, the generation of each response word in the proposed method is conditioned on multiple latent variables in sequential order. Gao *et al.* (2019d) leverage novel regularization terms to enhance the CVAE model with relevance and meanwhile maintain diversity. Gao *et al.* (2019c) introduce explicit semantic meaning in the discrete latent variable to enhance CVAE models on the short-text conversation. Zhang and Zhang (2019) leverage the CVAE model to construct a hierarchical response generation structure that can learn different levels of diversity information (i.e., word-level and discourse-level). Zeng *et al.* (2019) make the latent variable in CVAEs fit a Dirichlet distribution with adaptive structures to effectively represent complex latent variables. Cui *et al.* (2020) compensate CVAE-based conversation models with an attention mechanism constrained by fine-grained focus information to mitigate the scarcity of discourse-level information. Ko *et al.* (2020) alternatively introduce a regression task on the latent space to combine multiple valid responses with similar semantics for a prompt and accordingly alleviate the diversity problem in response generation. Khan *et al.* (2020) conduct adversarial learning on the latent space for diverse dialogue generation. Chen *et al.* (2018a) incorporate hierarchical structure and variational memory network into an encoder-decoder neural network.

Besides CVAE-based models, multi-pass encoder-decoder and multi-stage encoder-decoder can also change the outputs of generation-based chit-chat systems. Zou *et al.* (2018) propose a multi-encoder to multi-decoder (MEMD) framework to promote diversity of shot-text conversation. Kong *et al.* (2020) argue that indiscriminately generating function words and informative content by a single decoder can yield generic responses. To solve this issue, they propose a two-step decoding strategy to separately produce low-frequency content words (words that have substantive lexical content) and generate the high-frequency function words (words that essentially serve to make grammatical properties).

In addition to the above categories, there are also other efforts to optimize generation models. Tian *et al.* (2019) propose a memory-augmented generative model for conversational response generation by abstracting useful information from the training corpus and then saving this information in the memory. With the assistance of the memorized information, the generation model can output more informative and diverse responses.

4.3.3 Training with Effective Objectives

Apart from data manipulation and generation framework, training objectives also serve as the key component of deep neural networks. To address the expression diversity problem, many researchers attempt to introduce frequency-aware training objectives to encourage generating informative words and penalizing high-frequency words. Li *et al.* (2016a) leverage Maximum Mutual Information (MMI) as the training criterion of their neural conversation model to produce more diverse responses. Li *et al.* (2016c) also bring reinforcement learning to neural conversation systems. Song *et al.* (2017) introduce the maximum marginal relevance ranking algorithm in the beam search process to prevent decoding universal responses. Jiang and Rijke (2018) review previous approaches to the low-diversity problem of chit-chat systems and link it with the over-confidence problem. Besides, they point out several potential directions to mitigate such a challenge, i.e., penalty strategies on output confidence and label smoothing. Jiang *et al.* (2019) further enhance the cross-entropy training objective with frequency information by linking token frequency with a weighting mechanism to address the over-confidence problem in creating generic responses. Inspired by the huge success of boosting method in image and language generation, Du and Black (2019) introduce it into dialogue generation to improve diversity. Specifically, they build a base framework conditioned on recent boosting theory to combine varied training and decoding strategies such as mutual information, maximum likelihood augmented by reward, etc.

Alternatively, unlikelihood training shows competitive performance in solving the expression diversity problem. To address the limitations of likelihood-based decoding objectives, Baheti *et al.* (2018) introduce two distributional constraints to encourage semantic similarity, and the distribution over topics and syntax in the response resembles user input. Liu *et al.* (2018b) propose to assign different weights for the different responses based on the one-to-many phenomenon in chit-chat conversations, in which the weights are calculated based on the statistics of the corpus. Gao *et al.* (2019b) design a reinforcement learning algorithm to generate multiple diverse responses simultaneously for short-text conversation. Khayrallah and Sedoc (2020) apply simulated multiple reference training to model one-to-diversity of non-task-oriented dialogues. Cai *et al.* (2020a) leverage contrastive learning for dialogue generation to solve the low-diversity issue of maximum likelihood estimation (MLE) objective in chit-chat conversations. Ueyama and Kano (2020) propose an inverse n-gram frequency (INF) loss that can incorporate contextual fluency and diversity at the same time to generate diverse conversations. Li *et al.* (2020c) extend the recently introduced unlikelihood loss to address the low-diversity limitation caused by maximum likelihood training. Besides, He and Glass (2020) introduce negative training to prevent the model from creating offensive and generic responses. The proposed method first identifies unexpected responses as the negative samples and uses them to punish unwanted responses during the post-training process

4.3.4 Leveraging Extra Resources

Except for the aforementioned methods, other strategies that can encourage diverse expression and model one-to-many diversity are investigated. The most popular pointcut is leveraging extra resources. Qiu *et al.* (2019) propose to utilize multiple valid references, which is not always available, and further explore the correlation of different responses to model the one-to-many mapping of chat-driven conversations. Ko *et al.* (2019) use linguistically motivated specificity and semantic plausibility reranking to generate informative responses. Su *et al.* (2020a) propose to diversify dialogue generation from the perspective of leveraging non-conversational text, which covers a much broader range of topics.

4.4 Context Modeling: Single-Turn and Multi-Turn

Context modeling is also an essential problem for chit-chat conversations. For real-world conversations, it is not rare to converse in over hundreds of turns, and thus mimicking human conversations involves modeling long-range dialogue context. Early research of neural-based response generation system mainly consider single-turn context to verify whether fully data-driven methods can generate human-like responses. Later on, various methods are proposed to model multi-turn context in chit-chat conversations. The below part first presents a few representative works that only consider single-turn context and then elaborates the more actual setting, e.g., learning to generate response with multi-turn context.

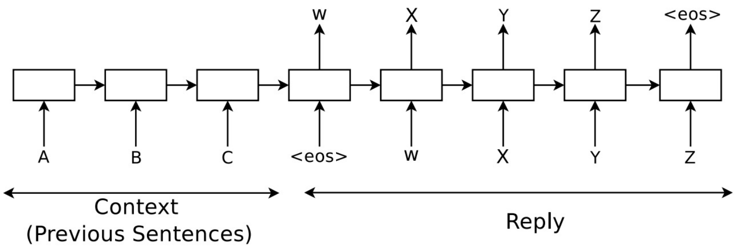

Figure 4.3: Conversation model with sequence to sequence learning and single-turn context, where <eos> is a special token to indicate the beginning and ending of a sequence (Vinyals and Le, 2015).

4.4.1 Single-Turn Context Modeling

As shown in Figure 4.3, early-stage of deep neural conversation systems (Vinyals and Le, 2015; Xu *et al.*, 2017) mainly consider single-turn conversations, i.e., the first person utters "ABC", and the second person responds "WXYZ". These methods are built upon recurrent neural networks with the encoder-decoder framework, where encoder is responsible for capturing and aggregating context information while the decoder auto-regressively generates the target response.

4.4.2 Multi-Turn Context Modeling

A more realistic setting is to consider multi-turn context information in chit-chat dialogues so that the systems can chat with consistent content and other characteristics and provide more accurate response. With the increase of data size and computing power, learning to chat with multi-turn context utterances has become the mainstream of this field, and existing research can be roughly grouped into the following two lines.

One is to study context modeling and utilization method. Tian *et al.* (2017) propose weighted sequence integration to explicitly weight the context vector by calculating the correlation between context and query, which can introduce more related context and reduce adverse effects of irrelevant noises. In order to capture salient information in the context and generate highly relevant responses, an attention mechanism is used in the encoder. Xing *et al.* (2018) introduce HRAN, which uses word-level attention within utterances and utterance-level attention among utterances, to dynamically model important parts of a context. Considering the advantage of sequence integration, Zhang *et al.* (2018c) use static and dynamic attention mechanisms for each utterance in one conversation and weight them to obtain contextual representation.

Although traditional attention mechanism or cosine similarity method has been applied to solve the problem of long-distance between response and relevant context information, these methods may lead to insufficient correlation hypothesis. Therefore, the model called ReCoSa is proposed by Zhang *et al.* (2019a) to solve the long-distance dependence problem, initializing representation of each context with LSTM and leveraging multi-layer multi-head self-attention mechanism to computed attention weights for the decoder between updated context and masked response representation. Dziri *et al.* (2019) introduce THRED. This hybrid model combines conversation history from previous utterances and topic words from a Latent Dirichlet Allocation model using message attention, context-level attention, and topic attention, which generates consistent and topic-related responses. Since RNN-based models cannot perform parallel computing, Mangrulkar *et al.* (2018) propose a CNN-based hierarchical structure. In this architecture, CNNs are utilized to build encoder and decoder generating n-best responses. Then the proposed CNN re-ranker and N-gram match re-ranker are combined to determine the final rank of each hypothesis. This hybrid model combines conversation history from previous utterances and topic words from a Latent Dirichlet Allocation model using message attention, context-level attention, and topic attention, which generates consistent and topic-related responses. Since RNN-based models cannot perform parallel computing, Mangrulkar *et al.* (2018) propose a CNN-based hierarchical structure. In this architecture, CNNs are utilized to build encoder and decoder generating n-best responses. Then the proposed CNN re-ranker and N-gram match re-ranker are combined to determine the final rank of each hypothesis.

Most hierarchical models focus on representing the context from the word level and utterance level. However, they do not explicitly model the meaning and relationship of utterances. Therefore, in light of the relationship between the query and response under a background, Shen *et al.* (2019a) propose the CSRR model. In this model, hierarchical latent variables based on VAEs are utilized to represent the utterance meaning. Three hierarchies, namely discourse level, pair level, and utterance level, are built to learn the dependency between query and response. Shen *et al.* (2018) point out that a good response should be able to smoothly connect previous and future conversations. They introduce the NEXUS model to enhance the connection through maximizing mutual information. With the proposal of the transformer, many researchers applied it to conversation generation.

Cai *et al.* (2020c) propose BCTCE structure. Different from previous hierarchical models, this model directly uses a bi-channel transformer to realize the parallel encoding of dialogue utterances and the document for document-driven conversation. Although transformer-based models can have better performance on long-term dependence, according to the empirical study conducted by Sankar *et al.* (2019), advanced seq2seq response generation models are insensitive to random modifications or noise in dialogue context, i.e., these models are incapable of capturing dialogue context information efficiently. In addition to modifying the model structure, many studies also introduce new tasks to improve the performance of multi-turn response generation. Zhou *et al.* (2019) propose a context rewriting network(CRN), rewriting the last utterance according to the context history and original last utterance. This unsupervised method generates a self-contained utterance and makes context modeling explainable and controllable. Zhao *et al.* (2020b) introduce a model with a simple structure for response generation, together with order recovery and masked content recovery tasks. This method reduces the complexity of the model, makes context understanding learnable, and improves conversation generation.

Another is to leverage more context information and correlated dialogue history. Wu *et al.* (2019b) propose a new response generation pattern, prototype-then-edit. First, a prototype selector retrieves a context-response pair according to the current context and then rewrites the prototype response by taking differences between prototype context and current context into consideration. Feng *et al.* (2020) believe that a good response is generally related to the potential context knowledge in the specific scenario. They propose to combine the dialogue history and future dialogue to build scenario knowledge, enhancing the conversation generation system. To incorporate scenario knowledge without future dialogue, an imitation learning framework is introduced to imitate the scenario-based teacher model. As to E-Commerce, Zhang *et al.* (2020e) incorporate the seller's historical dialogue information into response generation through finding out the most relevant seller's historical responses for the customer's question and fusing information from the generation module and copy module. In terms of multi-party dialogue, it is challenging to extract relevant context information due to complex interaction among the interlocutors' roles. Liu *et al.* (2019a) propose a new model, incorporating interlocutor-aware contexts into recurrent encoder-decoder frameworks and predicting the speaker and the addressee when generation responses. For VQA tasks, the MAC network shows strong performance on single-turn VQA tasks, in order to adapt this network to tasks that need reasoning over the dialogue history, Shah *et al.* (2020) augment MAC networks with Context-aware Attention and Memory(CAM), attending over the MAC control states of past dialogue turns. This structure makes the conversation characterized with history dependency and coherence.

4.5 Knowledge and Grounding in Response Generation

In real-world conversations, a feasible response is not only correlated to the current context information well but also constrained by commonsense knowledge. For chit-chat systems, many researchers have investigated incorporating extra knowledge into the response generation process in recent years. Basically, there are two main types of knowledge that are utilized in neural response generation models, including structured knowledge bases that support logical reasoning and unstructured grounding information such as Wikipedia passage, document, and other background information. In this section, we first introduce representative works that utilize structured knowledge in response generation and then present some typical research that leverages unstructured information. At the end of this section, we give a few response generation methods that can incorporate structured and unstructured knowledge simultaneously.

4.5.1 Structured Knowledge

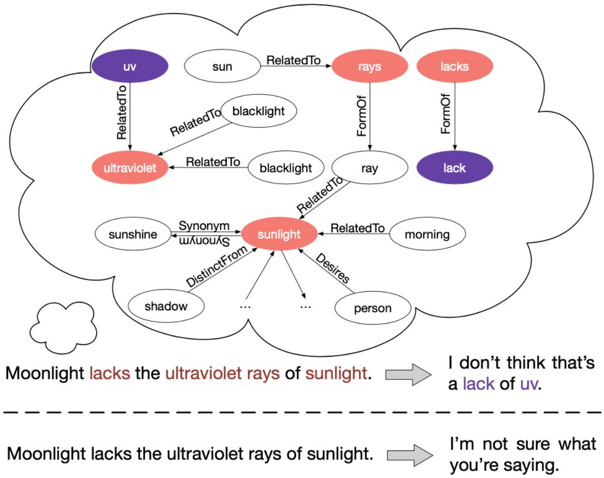

Figure 4.4: Chit-chat response generation with/without (the first/second line) the enhancement of structured commonsense knowledge (Zhou *et al.*, 2018b).

Structured knowledge in existing response generation solutions mainly represents entities and their relationships. Liu *et al.* (2018a) employ a knowledge base to mitigate the problem of generating short, general, and meaningless responses. Through extracting correlated facts for a given dialogue context and performing entity expansion, the proposed method can have the capability of convergent and divergent thinking conditioned on a knowledge base, and thus it can generate more informative responses with accurate entities. Zhou *et al.* (2018b) introduce large-scale structured commonsense knowledge[2] in chit-chat conversation generation to augment both language understanding and generation, in which the detailed setting is presented in Figure 4.4. For a given post, the proposed model first extracts the associated knowledge graphs from a knowledge base. Then, a graph attention method is introduced to map the retrieved knowledge graphs to their vector representations in a static manner. During the generation process, another graph attention mechanism is proposed to fuse retrieved knowl-

[2] https://conceptnet.io/ (date accessed: 11 April 2022)

edge graphs and triples dynamically for generating commonsense-aware responses.

Wu *et al.* (2020e) not only ground response generation on commonsense knowledge graphs but also introduce topic facts from a neural recommender to achieve controllable and interpretable response generation[3]. Zhang *et al.* (2020a), alternatively, address the low-diversity challenge of response generation by designing manipulation strategies in the concept space. They treate the attention-guided traverses in the knowledge graphs as the possible conversation flows to achieve better knowledge-grounded dialogue generation. To extract more context-relevant knowledge when traversing a knowledge graph, Jung *et al.* (2020) explore to model the KG structure information with a novel in-and-out attention flow conditioned on dialogue context. Wu *et al.* (2020d) focus on the problem of grounding response generation on context-specific knowledge. The authors present a mechanism to filter out unrelated facts and utilize two fusion strategies to effectively integrate knowledge into the response generation. Xu *et al.* (2020b) divide multi-turn response generation into two sub-tasks, i.e., explicit goal-planning grounded on knowledge graph and goal completion by topic elaboration. To complete the task, the authors present a three-layer hierarchical reinforcement learning model, where the upper-layer policy learns to traverse a knowledge graph to complete the goal planning sub-task, while the lower two layers produce conversation about a specific topic conditioned on the goal-planning.

4.5.2 Unstructured Knowledge and Grounding

Vougiouklis *et al.* (2016) treat correlated Wikipedia sentences as extra background knowledge to augment the multi-turn response generation task. As shown in Figure 4.5, each sequence of Reddit utterances is paired with 20 Wikipedia summary sentences. Accordingly, the chat-driven response generation task is formulated as generating a response conditioned on previous context utterances and the corresponded background knowledge sentence wherein these two different types of information

[3]https://github.com/pku-sixing/IJCAI2020-TopicKA (date accessed: 11 April 2022)

Topic	Reddit Sequence of Comments	Wikipedia Sentences
Noam Chomsky	<start> Noam Chomsky: Bernie Sanders is Not a Radical. He has Mass Support for Positions on Health-care & Taxes <end> <start> Funny, because Bernie Sanders's idol Eugene Debs ran against FDR <end> <start> Maybe Clinton will be FDR <end> <start> Watch out, Japanese. <end> <start> Japanese You misspelled Syrians <end>	For Chomsky, there are minimalist questions but the answers can be framed in any theory. ⋮ Minimalism in structured writing or topic-based authoring is based on the ideas of John Millar Carroll. Minimalism is about reducing the interference of the information with the users sense-making process. An error, in fact, is the teachable moment that the content can exploit.

Figure 4.5: An example that aligns Reddit utterances and Wikipedia sentences (Vougiouklis *et al.*, 2016).

are modeled by an RNN and a CNN module, respectively. To train a neural chit-chat systems that can generate coherent and content rich response on the Reddit news dataset, Parthasarathi and Pineau (2018) also introduce unstructured knowledge from Wikipedia summaries and further incorporate the NELL knowledge base. Ghazvininejad *et al.* (2018) extend neural response generation to a more useful conversational application by producing more contentful responses grounding extra knowledge. They propose a fully data-driven model to complete the task of knowledge-grounded conversation task, which consists of a facts encoder, dialogue encoder, and dialogue decoder. To train the proposed model, they extracted a dataset from the Foursquare dataset with comments about restaurants and Twitter, in which each conversation contains entities that tie to Foursquare.

Lian *et al.* (2019) explore unstructured knowledge selection (e.g., user profiles, Wikipedia) in response generation to better leverage external knowledge. The authors propose a knowledge selection mechanism by utilizing the posterior distribution inferred from context and responses and the prior inferred from just the responses. Lin *et al.* (2020a) explore the problem of jointly utilizing different knowledge to produce informative responses. Unlike the previous decoding processes in the RNNs-based sequence to sequence model, the authors propose to extract relevant knowledge during each decoding step and recurrently incorporate the extracted knowledge into updating the decoding state. Besides, the authors also extend the pointer mechanism for knowledge utilization.

Tian *et al.* (2020) propose a memory model to capture on-demand knowledge from both conversational contexts, document, and part of anticipated responses in response generation. Chen *et al.* (2020a) attributes the discrepancy of the prior and posterior in latent-variable-based knowledge selection to lacking posterior information in extracting the correlated knowledge with the prior and the exposure bias problem, i.e., using posterior in training while using the prior for testing. Accordingly, the authors propose compensating the prior with necessary posterior information and designing a knowledge distillation strategy to conduct decoder training with selected knowledge from the prior distribution. Zheng *et al.* (2020a) study the difference of the selected knowledge between the current turn and previous turns. To enhance the performance of knowledge selection in knowledge-grounded dialogue generation, Wu *et al.* (2020c) propose to first retrieve relevant prototype dialogues and then utilize these dialogues to extract knowledge facts.

As mentioned above, there are mainly two types of extra knowledge, i.e., triples from knowledge graphs and textual documents from corpus. The fusion of the structured knowledge and the non-structured grounding texts has been widely studied. Moghe *et al.* (2018) argue that formulating response generation as a sequence-to-sequence task is overly simplistic, while human conversations heavily rely on their background knowledge about a specific topic. To mimic human behaviors and study the effect of background knowledge, they create a dataset comprising dialogues about movies. Each utterance is made from the background information about movies, e.g., plots, reviews, by crowd workers. Liu *et al.* (2019c) design a novel knowledge aware response generation framework that consists of three critical components, i.e., a knowledge graph with the enhancement of unstructured texts to allow reasoning and provide more information, a knowledge selection module, and the generation component to better fuse both knowledge triples and document texts for chit-chat response generation.

Recently, Zhao *et al.* (2019c) explore a more realistic knowledge-grounded dialogue generation setting that only limited labeled examples are available during training. The authors design a disentangled decoder to isolate parameters that rely on knowledge-grounded training examples from the entire generation model. By doing so, the main component of the model can be trained on the cheap and plentiful ungrounded dialogues and unstructured documents, while the rest parameters can be learned with the limited grounded examples. Li *et al.* (2020b) further investigate the setting of zero-resource knowledge-grounded dialogue generation, i.e., context-knowledge-response triples are not required in training. To achieve this goal, the authors propose to leverage latent variables to express the knowledge and devise a variational method to estimate a generation model from two independent corpora, including a dialogue corpus and a knowledge corpus.

4.6 Human Factors: Emotion, Persona, and Beyond

message	What did you have for dinner?
baseline	I had fish and chips.
user1	I had spag bol.
user2	Chicken and chips.
user3	Chicken and rice.
user4	Fish and chips.
user5	I had spag bol.
user6	I had Mexican food.
user7	Salad...
user8	I had chicken and chips.
user9	I had spag bol.
user10	Pizza.

Figure 4.6: A case study that reflects the influence of different persona (Li *et al.*, 2016b).

Unlike task-oriented systems, chit-chat systems mainly serve as personal assistants, emotional companions, and the key factors that affect user experience are the long-term engagement and the matching degree between chit-chat systems and users. Thus, human factors, e.g., emotion, persona, play a vital role in real-world conversation systems. Take a very simple case shown in Figure 4.6 for illustration, different users of distinctive persona respond to the same message with diversified contents. This is especially true when we are considering more realistic scenarios. Also, the previous sections of this section have concluded that the typical sequence-to-sequence neural networks tend to generate generic responses, and the utilization of extra information such as knowledge graphs, document texts could facilitate creating diverse responses. Similarly, human factors can also serve as strong background information to control the content of generated responses and prevent the conversation model from creating uninformative responses. In view of these merits, many attempts have been made to incorporate human factors in generation-based chit-chat systems, which can be roughly grouped into three categories, i.e., persona, emotion, and beyond.

Persona. As one of the pilot studies, Li *et al.* (2016b) explore both the influence of speaker consistency and speaker-addressee interactions for neural response generation, where the profile of interlocutors are expressed by embeddings (similar to word embeddings) learned from massive training conversations. Kottur *et al.* (2017) launch an empirical study to investigate the effects of pre-training, embedding training, data cleaning, diversity-based re-ranking, evaluation setting. Based on the trade-offs of different factors, the authors further propose a a speaker aware neural response generation model with the enhancement of larger datasets and bootstrapping strategy for speaker embeddings pre-training. Luan *et al.* (2017) propose a multi-task learning neural conversation model that leverages both conversations across speakers and other types of relative data to the speaker roles to be modeled. Unlike previous works that utilize implicit embeddings as persona, Zhang *et al.* (2018b) crowd-source a dataset that conditions response generation on explicit profile information and the initial personal topics of the conversation partner. In witness to the success of profile information in response generation, Mazare *et al.* (2018) provide a new dataset that contains 5 million personas and 700 million conversations based on these personas to train a persona-enhanced chit-chat system at scale. Qian *et al.* (2018) study profile assigning in generating coherent conversations, which is solved by a profile detector to judge whether a profile should be used when responding and a bidirectional decoder to generate personality-coherent responses in a forward-and-backward fashion. Chu *et al.* (2018) propose a multi-level attention mechanism and a memory module to learn persona representations. Hu *et al.* (2018) focus on the adaptation of linguistic cues and personality traits to control the system output at each step of the dialogue. Engonopoulos *et al.* (2018) concentrate on how user groups affect utterance generation in chit-chat systems.

There are also many works in recent years, which mainly enhance personalized chit-chat systems from persona information modeling and incorporation in response generation, including but not limited to, modeling personalization in the continuous space of Wasserstein Autoencoders (Chan *et al.*, 2019), variational hierarchical user-based model (Bak and Oh, 2019), adversarial learning (Olabiyi *et al.*, 2019), graph-structured network (Hu *et al.*, 2019), the combination of memory module and conditional variational autoencoder (Song *et al.*, 2019a), persona-guided variational response generator (Wu *et al.*, 2020a), mutual persona perception (Liu *et al.*, 2020), the three-stage framework of generate-delete-rewrite (Song *et al.*, 2020a), persona enhanced dual alternating learning network (Jiang *et al.*, 2020), opinionated dialogue generation with stance-based personas (Scialom *et al.*, 2020), and sketch-filling-ranking framework (Shum *et al.*, 2020). Besides, the investigation of persona-aware response generation with limited resources is non-trivial. Chang *et al.* (2019) propose a semi-supervised variational model for generating speaker-consistent dialogue response. Madotto *et al.* (2019) propose a meta-learning solution to achieve personalized dialogue learning without using any persona description. Moreover, incorporating more useful information in persona-aware response generation is also appealing. Song *et al.* (2020b) introduce natural language inference to generate persona consistent dialogues. Majumder *et al.* (2020a) explore commonsense expansions for persona-grounded dialogue generation.

Emotion. Zhou *et al.* (2018a) first explore large-scale conversation generation conditioned on specific emotion. On the basis of the vanilla sequence-to-sequence dialogue generation model, the authors propose three strategies to achieve emotional response generation. First, they encode the emotion category into the vector space for learning the emotion representation along with the model training. Then, the internal memory module is presented to model the dynamic evolution of emotional information in the decoding progress. Besides learning and utilizing emotional information in the continuous space, they also introduce an external memory to explicitly fuse discrete emotional words in generation. Li and Sun (2018) propose a syntactically constrained decoding framework for emotional response generation, where the basic encoder-decoder component utilizes emotion keywords in the encoding stage and presents topic keywords for the decoding process. Besides, two extra RNN decoders are used to asynchronously introduce emotion and topic keywords in bidirectional generation. Huang *et al.* (2018) empirically study three different models to restrict generated responses with expressed emotions by introducing emotion constraint in encoder input, encoder output, and decoder. Zhou and Wang (2018) leverage emoji to convey emotion information to achieve generating emotional responses as scale, where the generation models are implemented as several conditional variational auto-encoder variants.

Shi and Yu (2018) propose to introduce user sentiment from acoustic, dialogic, and textual information to perform sentiment adaptive dialogue generation. Colombo *et al.* (2019) study emotional response generation in a controlled manner in which emotions in continuous representation are used. Through investigating real-life conversation data, Song *et al.* (2019b) conclude that emotional states are described either by strong emotional words or by implicitly combing neutral words in distinct ways. Based on the observations, the authors propose an emotional chit-chat system that can express the desired emotion explicitly or implicitly in the generated responses with a unified framework. Ma *et al.* (2020) concentrate on the emotion drift problem, which is referred to the inconsistency of emotion between context and responses, and propose to use a control unit framework to incorporate consistent emotional words during generating responses. Shen and Feng (2020) leverage the duality of emotional response generation and the correlated emotional query generation to enhance the correlation modeling of queries and responses. They also utilize the curriculum learning method to gradually generate responses with higher emotion expression difficulty.

More recently, empathetic chit-chat conversation systems have attracted growing attention from both academia and industry. To facilitate the development of empathetic dialogue systems, Rashkin *et al.* (2019) propose a benchmark, which contains 25K emotional conversations. Li *et al.* (2020e) propose a variational model to generate appropriate responses with user emotional reaction awareness. Li *et al.* (2020d) propose a multi-resolution adversarial approach to capture the nuances of human emotions and the potential feedback from users, which can generate more empathetic responses. As presented by Zhou *et al.* (2020c), the popular social chatbot XiaoIce can also output empathetic conversations. In addition, Majumder *et al.* (2020b) consider user emotion in varying degree for empathetic response generation. Specifically, the authors leverage the sentiment-polarity of content, emotional mimicry, and randomness in emotion mixture to generate responses in varied emotions.

Beyond Persona and Emotion. Akama *et al.* (2017) study the setting of generating stylistically consistent response generation and propose a two-stage training framework that resembles transfer learning accordingly. Gao *et al.* (2019e) introduce a structured latent space to bridge conversation modeling and non-parallel style transfer for stylized response generation. Zhang *et al.* (2018a) attempt to deal with the task of response generation with controlled specificity. To calibrate how controllable attribute affect response generation, See *et al.* (2019) launch an empirical study to thoroughly test two representative controllable dialogue generation models in controlling varied attributes at different granularities, including a training strategy enhanced by a latent variable and a decoding strategy that can output different probabilities based on the given controllable feature. Takayama and Arase (2020) explore controlling the specificity of response generation, i.e., generating response conditioned on the word-level co-occurrence of context utterances and responses. They first present a normalized metric to measure the word-level co-occurrence between context utterance and response and then utilize the obtained score as distant supervision signal to guide the model training. Besides, the word-level co-occurrence information between input words and the vocabulary is also computed to achieve specificity-aware generation.

4.7 Pre-Training in Dialogue Generation Models

Starting from BERT (Devlin *et al.*, 2018), pre-trained language models (PLMs) have changed the phase of varied downstream tasks, and many

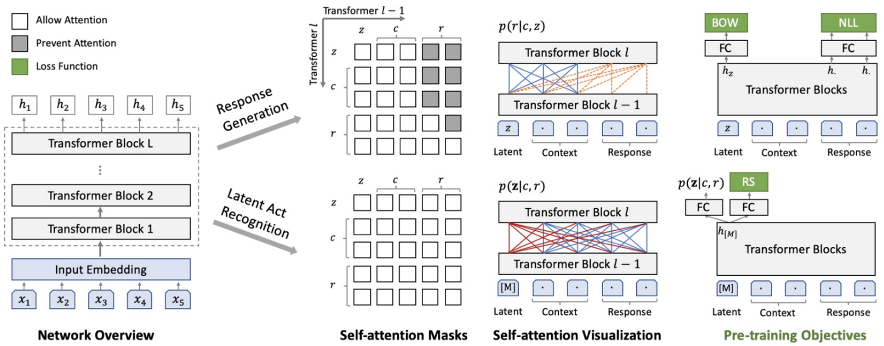

Figure 4.7: A representative pre-training framework designed for dialogue generation (Bao *et al.*, 2020a).

efforts have been devoted to building conversation systems. Bao *et al.* (2020a) propose a representative pre-training framework for dialogue generation. As shown in Figure 4.7, typical pre-training models consist of transformer blocks, and there are some specific modules (e.g., latent variable to address the inherent one-to-diversity challenge) and pre-training tasks (e.g., latent act recognition). Zhang *et al.* (2020f) pre-train a GPT-like language model on a large-scale dataset crawled from Reddit to improve the performance of response generation. Zhao *et al.* (2020a) build a generative dialogue system grounded on document texts by combining BERT and GPT-2, where the BERT encoder is responsible for obtaining the representation of context and extracting suitable knowledge, and GPT-2 serves as the decoder. Cao *et al.* (2020) adapt PLMs to response generation with the capability of multiple input sources modeling and point out that a well-designed fusion strategy is generally better than a straightforward one. Yang *et al.* (2020) explore style-sensitive response generation with limited paired data based on the generalization capability of PLMs. They introduce an extra training loss and a discriminator to produce both word-level and sentence-level style-aware responses. Zheng *et al.* (2020b) utilize PLMs as the backbone model for personalized response generation framework. More discussion about PLMs in dialogue systems is presented in Section 6.4.

4.8 Datasets

With the revival of deep neural networks and the simultaneous explosion of open-access web data, many datasets are collected and constructed for neural response generation. We introduce some representative and frequently utilized datasets in the following part.

Context. To train a deep neural response model, researchers explore crawling large-scale chat-like conversations from various social media and forums, e.g., Twitter, Weibo, and Reddit. These conversations are open-available and cover a myriad of daily topics. Shang *et al.* (2015) constructed a **S**hort-**T**ext **C**onversation dataset (STC) from Weibo, which contains roughly 4.4 million training pairs. Each pair consists of a user-issued post text and a target response that is supposed to be mimicked by the neural model. This dataset is mainly employed and studied for the single-turn context setting mentioned in Section 4.4.

Sordoni *et al.* (2015) created a large-scale context-sensitive corpus for chit-chat conversational response generation, which comprises 127M (context, message, response) triples extracted from Twitter Firehose, ranging from June 2012 to August 2012. Each triple composes of a context sentence, a message which resembles the post text in STC and a target response. In this setting, a deep neural conversation model is built to create a response conditioned on both the context sentence and massage sentence, which corresponds to the multi-turn context setting in Section 4.4.

Li *et al.* (2017b) developed a high-quality multi-turn conversation dataset by crawling raw data from various websites for English learners to practice spoken language in daily life, namely **DailyDialog**. Compared with the aforementioned two datasets from Weibo and Twitter, Dailydialogue consists of formal language and contains less noise.

Diversity. Xu *et al.* (2018b) built a **L**arge-**S**cale **D**omain-**S**pecific **C**onversational **C**orpus (LSDSCC) to mitigate universal responses and evaluate conversation diversity. They first collected dialogues from the most popular movie discussion board in Reddit, with high-quality and focused domain to alleviate the challenge of producing universal responses.

After thorough pre-processing and cleaning procedures, they also created multiple ground-truth responses for each given query and the correlated context in the test set to calibrate the diversity of generated responses.

Personalization. To evaluate the influence of human factors, Zhang *et al.* (2018b) collected a **Persona-Chat** dataset via Amazon Mechanical Turk. They crowd-source 1,150 personas with 950 personas for training and 100 personas for validation and testing, respectively, where each persona attaches at least 5 profile sentences. To prevent modeling of trivial overlap, the authors proposed to prepare additional rewritten sets on the 1,150 personas by rephrasing, generalization, or specialization. Then, they assigned each of the two Turkers a persona from the possible set and asked Turkers to chat, resulting in a dataset of 131,438/15,602/15,024 utterances (8,939/1,000/968 dialogues) for training/validation/testing.

Wang *et al.* (2019c) designed an online persuasion task on the Amazon Mechanical Turk platform to study personalized persuasive dialogue systems for social good. They asked one participant to persuade the other to denote a specific charity and collected 1,017 dialogues in total. Then, a persuasion strategy annotation scheme is proposed to label a subset of the collected conversations, where emerging persuasion strategies are annotated. They also linked demographic and psychological backgrounds, i.e., personality traits, morality, value systems, and donation behaviors. Finally, they studied the relationships between personal backgrounds and persuasion strategies.

Zheng *et al.* (2019a) created a large-size multi-turn dialogue dataset from Weibo, named as **PersonalDialog**, which composes of 20.83M conversation sessions and 56.25M utterances of 8.47M speakers. Each utterance belongs to a specific speaker, and each speaker associates with various traits such as age, gender, location interest, etc.

Grounding. Dinan *et al.* (2018) crowd-sourced a multi-turn dialogue datasets grounding on Wikipedia, named as **Wizard of Wikipedia**. They asked two participants to chat with each other. Unlike previous instruction, the two participants of this setting are asymmetric, i.e., one participant acts as a knowledgeable expert, referred to the **wizard** while another plays the role of a curious learner, denoted as the **apprentice**. When chatting, the wizard has access to related Wikipedia knowledge while the apprentice does not. This dataset covers 1,365 chit-chat dialogue topics such as commuting, Gouda cheese, music festivals, podcasts, and bowling and consists of 22,311 dialogues with 201,999 conversation turns, which is split by 166,787/17,715/17,487 for train/validation/test respectively. The test set contains two subsets, consisting of 533 overlapping topics and 58 unseen topics in train and validation sets

Zhou *et al.* (2018c) also collected a document-grounded conversation dataset through Amazon Mechanical Turk, where annotators are asked to chat about the given Wikipedia article. To facilitate the dataset collection, the topic of documents is restricted to popular movies and each conversation has more than 12 turns. The final dataset contains 4,112 conversations in total and with an average of over 21 turns. Among which, 2,128 conversations are from the setting that only one of each two-annotator pair has assess to a Wikipedia document while the rest 1984 conversations are from the setting that both annotators have access to the same document.

To mimic human conversations that rely on background knowledge about the topic, Moghe *et al.* (2018) create a corpus that consists of conversations about movies, i.e., each dialogue utterance is obtained by copying and/or modifying the background knowledge sentences of the movies, e.g., plots, comments, and reviews. The constructed dataset has 9,071 conversations about 921 movies with 90,810 utterances in total, where each utterance contains 15.29 words on average. There are 5,157 unique plots, 1,817 unique reviews, and 12,740 comments about these movies for creating responses.

Rashkin *et al.* (2018) crowd-sourced the **EmpatheticDialogues** dataset to facilitate the learning of empathy in chit-chat conversations. Each dialogue is grounded in a specific emotion label and situation that is created by the speaker, with a listener to chat with the speaker. The resulting corpus consists of 24,580 conversations from 810 different annotators, which is split into 19,533/2,770/2,547 for train/valid/test, respectively.

Tuan *et al.* (2019) designed a task of dialogue generation grounding on dynamic knowledge graphs and created a corresponded corpus from TV series, named as **DyKgChat**. Each input message is paired with a knowledge graph and the ground-truth response. The response generation process is required to correlate with the evolution of the knowledge graph. This corpus contains two subsets, i.e., **HGZHZ** and **Friends**, with 1,247 (17,164 turns) and 3,092 (57,757 turns) dialogues respectively. The average lengths of the two sets are 26.95 and 14.52 tokens per turn.

Wu *et al.* (2019a) constructed a dataset named **DuConv** for exploring proactive human-machine conversation under the guidance of explicit conversation goals. In the data collection process, one leads the conversation flow by sequentially altering the conversation topics conditioned knowledge graphs and the preset dialogue goal, while another paired annotator follows the conversation. Finally, the constructed corpus involves 29,858 dialogues, 270,399 utterances, and 143,627 entities. The average words per utterance and average knowledge per dialogue are 9.1 and 17.1, respectively.

Gopalakrishnan *et al.* (2019) explore a more general setting of knowledge-grounded chit-chat conversations and accordingly collected a dataset named **Topical-Chat** via Amazon Mechanical Turk. In the process of data collection, every two workers are asked to chat with natural and coherent content grounded in the provided reading sets. Each pair of partners do not have pre-defined roles like in **DuConv** and **Wizard of Wikipedia**. Instead, their conversations could be both symmetric or asymmetric to varying degrees, which is more realistic in human-human conversations. The established dataset covers 8 topics and consists of 9,058/1,131/1,130 conversations as train/valid/test sets. There are 248,014 utterances in total, and each utterance contains roughly 20 words on average.

Moon *et al.* (2019) manually annotated an open-ended parallel dialogue and knowledge graph corpus named **OpenDialKG**. The corpus is gathered in the Wizard-of-Oz setting Shah *et al.* (2018) in which two crowd-workers are asked to chat with natural and engaging dialogues. Concretely, the first annotator is asked to initiate a conversation about a given seed entity, and the second one is required to select the most relevant and natural facts from a given list of facts to create a conversational response. Only facts within the 1-hop or 2-hop path of the original conversation topic are considered. Once the second annotator responds to the first annotator, new multi-hop facts from the knowledge graph along with paths from entities introduced in the latest message are retrieved. In the next cycle, the first annotator is instructed to create a new message from the updated facts set, and such a cycle is repeated for each dialogue until a participant ends the conversation. There are two sub-sets in this corpus, including a recommendation task and a chit-chat task. The chit-chat data collection comprises 3,353 dialogues and 19,336 conversation turns. The final knowledge graph consists of 1,190,658 fact triples of 100,813 entities with 1,358 different relationships.

Qin *et al.* (2019) extracted a large-scale grounded conversational dataset from Reddit. On Reddit, each dialogue correlates with a submission title and a URL linking to a news or background article to start the conversation about the contents of the given URL, which can be utilized to study neural conversation with on-demand machine reading. After multiple steps of filtration and pre-processing, there are 28.4K/1.2K/3.1K dialogues and documents with 2.36M/0.12M/0.34M utterances for train/valid/test, respectively. There are 15.18M/0.58M/1.68M sentences in train/valid/test documents. The average lengths of utterances and document sentences are 18.74/18.84/18.48 and 13.72/14.17/14.15 words for train/valid/test, respectively.

Tang *et al.* (2019) further extracted a new dataset from **Persona-Chat** (Zhang *et al.* (2018b)) for target-guided chit-chat conversation by maintaining all conversations while discarding the persona information. Then, the extracted data is processed by automatically extracting keywords for each utterance. In doing so, each utterance is paired with a target subject, i.e., extracted keywords to guide the conversation. The final dataset contains 8,939/500/500 conversations for train/valid/test. There are 101,935/5,602/5,317 utterances in train/valid/test sets, respectively. There are 2,678/2,080/1,571 keyword types for train/valid/test, and the corresponded number of keywords are 2.1/2.1/1/1.9 on average.

4.9 Evaluation Metrics

4.9.1 Overlap-based Metrics

Basically, response generation is one of the representative text generation tasks. Owing to the progress and success of automatic evaluation methods for various text generation tasks, most works of response generation introduce general-purpose text generation metrics to assess the generated responses. The most prevalent metrics are BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Among them, BLEU and METEOR are originally designed for the neural machine translation task, while the ROUGE metric is presented for the summarization task. These overlap-based methods automatically evaluate the generation performance by computing the lexical (n-grams) matching degree between the generated candidates and the correlated ground-truth references. In other words, their success primarily depends on the number and quality of ground-truth references. Another critical point of these metrics is to design a combination strategy that can balance the n-gram matching performance in precision and recall, word order, and length.

BLEU. Papineni *et al.* (2002) enhance previous n-gram precision calculation with a maximum word count clipping strategy and propose to compute the logarithm average of {1,2,3,4}-grams precision to achieve a robust evaluation. Besides, they also present a length penalty method to improve the evaluation metric from the perspective of recall information computation, which is formulated as:

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases} \tag{4.2}$$

where r correlates reference length information, while c is from candidates. The final BLEU score is calculated by multiplying the logarithm average of n-grams with the penalty factor, written by

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp(\sum_{n=1}^{N} w_n \log p_n). \tag{4.3}$$

METEOR. METEOR (Banerjee and Lavie, 2005) is introduced to address a few limitations of the BLEU metric. Unlike the implicit recall information calculation in BLEU, METEOR explicitly incorporates the recall result in its computation details. Specifically, METEOR computes the F-mean between the unigram precision and recall to avoid zero output of the geometric average in BLEU calculation for extreme cases, where the calculation of unigram precision and recall is flexibly enhanced by porter stem and WordNet synonymy. It presents the concept of chunks to explicitly capture fluency information rather than n-grams in BLEU and calculates a penalty factor on chunks to compute the final score.

ROUGE. In contrast with Papineni *et al.* (2002) of utilizing the n-grams precision between candidates and the correlated ground-truth reference, Lin (2004) focus more on the recall performance for evaluating summaries and develop a package, named ROUGE, with multiple variants. For instance, ROUGE-N mainly counts the recall of reference n-grams in candidates, which can be computed as:

$$\begin{aligned}&\text{ROUGE-N}\\&=\frac{\sum_{S\in\{ReferenceSummaries\}gram_n\in S}\sum Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}gram_n\in S}\sum Count(gram_n)}.\end{aligned}\tag{4.4}$$

The other variants include ROUGE-L, ROUGE-W, and ROUGE-S.

The above three representative overlap-based automatic metrics can significantly reduce model development and evaluation costs, especially when existing multiple ground-truth references for each generated candidate. However, unlike previous text generation tasks, e.g., neural machine translation and summarization, open-domain chit-chat response generation encounters a much more serious one-to-many problem, and generative response generation models based on deep neural networks can create an enormous number of feasible responses. Thus, an adequate number of references for each generated candidate is vital for the effectiveness of overlap-based metrics. Gupta *et al.* (2019) conduct a thorough empirical study to calibrate the influence of multiple references for overlap-based metrics in chit-chat response generation evaluation. The results show that more human-crafted high-quality references can

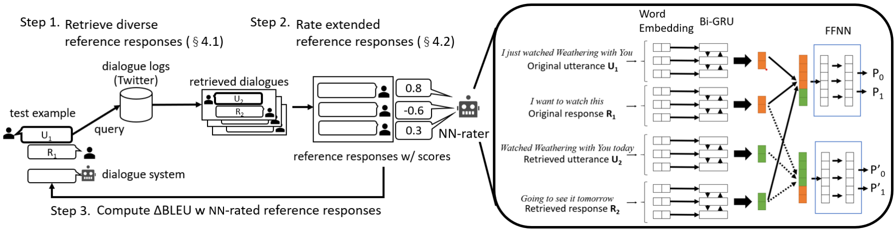

Figure 4.8: The overview of vBLEU (Yuma *et al.*, 2020).

consistently improve the correlation between human evaluations and automatic metrics. Besides, multiple references can help capture the generation diversity of each model. Under this guidance, a few strategies are introduced to enhance the overlap-based metrics with multiple references. We take ΔBLEU (Galley *et al.*, 2015) and vBLEU (Yuma *et al.*, 2020) as illustrations.

ΔBLEU. To obtain multiple references for each generated response, Galley *et al.* (2015) propose to utilize a BM25 algorithm based on bag-of-words to retrieve multiple pseudo references for each test sample. Then they introduce human efforts to assign a quality score w to each reference, ranging from [-1,+1]. Accordingly, the BLEU computation is modified as:

$$\Delta\text{BLEU-N} = \frac{\sum_i \sum_{g \in n-grams(h_i)} \max_{j:g \in r_{i,j}} \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n-grams(h_i)} \max_j \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}. \tag{4.5}$$

Instead of directly writing multiple references for each test sample, retrieving and judging requires less human effort and can also effectively enhance the robustness of the BLEU metric.

***v*BLEU.** Following the direction of constructing multiple references for each test sample and further reducing human effort in evaluations, Yuma *et al.* (2020) present the vBLEU metric. Similar to ΔBLEU, they first automatically extract multiple pseudo references for each test sample. Then, they introduce a neural model to automatically rate the retrieved references instead of annotating by humans in ΔBLEU. Experimental results demonstrate that such a neural rater can achieve comparable performance with human annotations as in ΔBLEU. More details of vBLEU are illustrated in Figure 4.8.

4.9.2 Embedding-based Metrics.

As mentioned before, the overlap-based metrics based on the n-gram overlap calculation show that the n-gram is not a good choice for capturing the correlation signals between candidates and references expect for with multiple high-quality references for each sample. Researchers attempt to handle this issue by performing matching in the semantic space, e.g., word vectors, hidden states of deep neural networks.

Embedding Average. Mitchell and Lapata (2008) treat the average of the word embedding of each word in a sentence as its semantic representation and compute the cosine similarity between the averaged vector representation of two sentences to indicate their matching degree. Instead of simply using the averaged vector representation, Lintean and Rus (2012) propose a greedy strategy to find the closest word for each generated candidate word in references under a given vector similarity metric. The matching signals between the greedily matched word pairs can then be aggregated to compute the matching degree between candidates and the correlated references.

Vector Extrema. Different from using the averaged word similarity in the embedding space, Forgues *et al.* (2014) propose to use the extreme value (either minimum or maximum) of each dimension of the word embeddings within an utterance as its vector representation for extracting the salient semantic matching features between two utterances. There are different choices of word embeddings, e.g., word2vec[4], glove[5], and many different ways to calculate the similarity between two vectors in chit-chat systems (Chan *et al.*, 2019; Gu *et al.*, 2019b).

[4] https://code.google.com/archive/p/word2vec (date accessed: 11 April 2022)
[5] http://nlp.stanford.edu/data/glove.840B.300d.zip (date accessed: 11 April 2022)

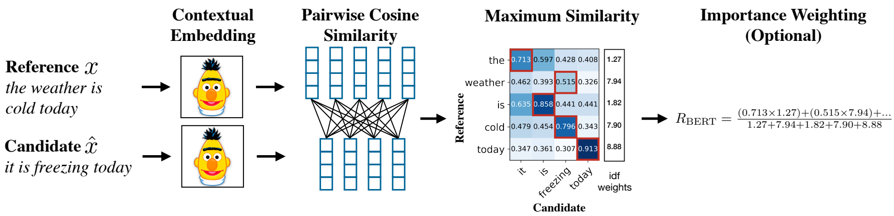

Figure 4.9: The overview of BERTScore (Zhang *et al.*, 2019c).

However, owing to limited training data and representation capability, conventional word embeddings can not effectively learn the contextualized representations for a word in different contexts. Thus, the correlations between embedding-based metrics and human evaluations can be further enhanced by more powerful models. Fortunately, with the explosion of open-accessible data and powerful computing resources, large-scale pre-trained language models (Devlin *et al.*, 2018; Liu *et al.*, 2019b; Radford *et al.*, 2019) can learn contextualized representations and have been widely used for the evaluation of chit-chat response generation (Zhang *et al.*, 2019c; Zhao *et al.*, 2019a; Sellam *et al.*, 2020).

BERTScore. Specifically, BERTScore (Zhang *et al.*, 2019c) uses a powerful PLMs, e.g., BERT, to obtain the contextualized representations of tokenized input sequences, which can be a generated response or ground-truth reference. Then, BERTscore uses the inner product of the calculated representation for each word pair as the similarity metric. Next, it utilizes a greedy strategy and combines precision and recall to compute the F1 score, which can be optionally augmented by inverse document frequency to assign larger weights to rare words. The overview of BERTScore calculation is given in Figure 4.9.

MoverScore. Except for the effective contextualized representations from PLMs, Zhao *et al.* (2019a) present a more powerful semantic similarity calculation method based on the Earth Mover's Distance to calculate the matching degree between a generated candidate and the correlated reference. Experiments on two different variants demonstrate

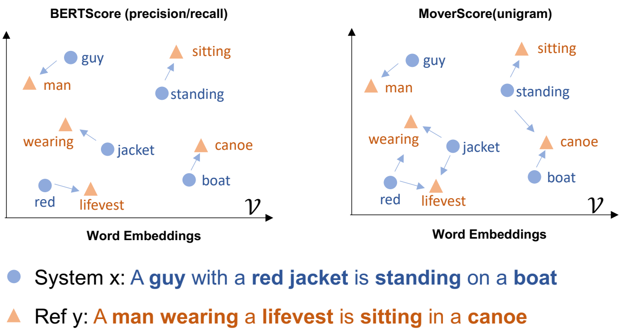

Figure 4.10: An illustration of MoverScore (Zhao *et al.*, 2019a) and BERTScore (Zhang *et al.*, 2019c).

that the proposed MoverScore performs best with the word mover.

BLEURT. Many recent works have pointed out that fine-tuning PLMs with task-related data and objectives can achieve much better performance than simply using the contextualized representations calculated from PLMs. Under this background, Sellam *et al.* (2020) build an automatic evaluation metric upon the BERT model. They first launch a pre-training process with three different strategies to obtain synthetic pre-training data and nine relevant pre-training objectives. Then, the pre-trained model is fine-tuned on supervised data to perform the evaluation of text generation.

4.9.3 Learning-based Metrics.

ADEM. Lowe *et al.* (2016) first propose to treat the evaluation process as a next utterance prediction task which is based on the context-response manner. Then, Lowe *et al.* (2017) propose the automatic dialogue evaluation model (ADEM), which consists of a pre-trained hierarchical RNN encoder and a liner mapping layer. More concretely, the hierarchical RNN encoder is pretrained with the encoding-decoding utterance generation task on a large-scale corpus. Then, ADEM is further trained on a few labeled data for predicting the evaluation score,

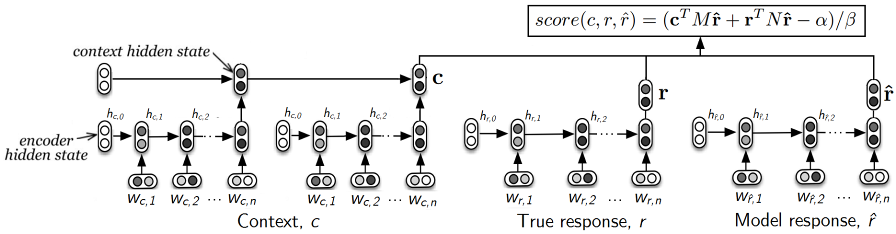

Figure 4.11: An illustration of ADEM (Lowe *et al.*, 2017).

written as

$$score(c, r, \hat{r}) = (c^T M\hat{r} + r^T N\hat{r} - \alpha)/\beta \tag{4.6}$$

where c, r, $\hat{r}$ represent the encoded hidden states of context, reference, and generated response candidate, respectively. $M, N \in R^n$ are learnable parameters to map the output into [1, 5]. The model structure of ADEM is presented in Figure 4.11.

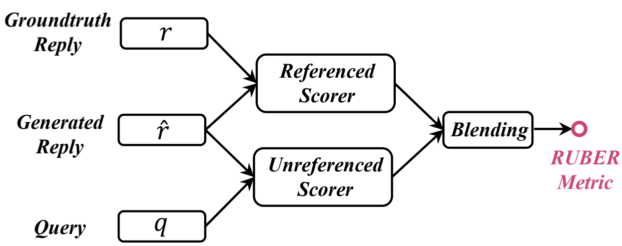

Figure 4.12: An overview of RUBER (Tao *et al.*, 2018).

RUBER. As presented in Figure, 4.12 is an unsupervised metric, considering both the matching degree of candidate-reference (i.e., referenced score) and candidate-query (i.e., unreferenced score). Different from the ADEM, RUBER does not use the human-labeled dataset. The referenced score model is trained with the supervision signal from a variant of the extrema metric. The unreferenced score model is trained to distinguish between negative and positive samples. The referenced

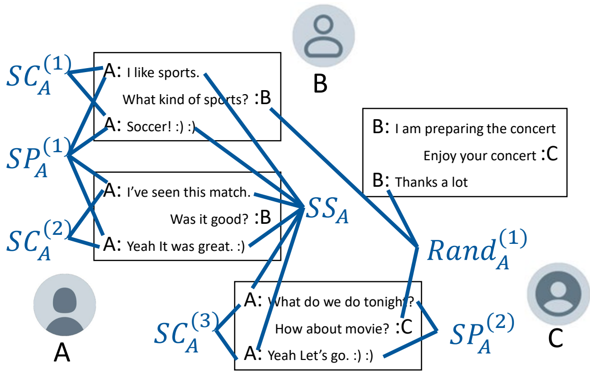

Figure 4.13: An illustration of SSREM (Bak and Oh, 2020).

and unreferenced scores are then normalized to a bounded range, i.e., (0,1). Such an unsupervised learning-based metric achieves a high correlation with human annotation. Ghazarian *et al.* (2019) enhance the RUBER metric with the contextualized representation from the BERT.

PONE. The quality of negative samples is crucial when training a discriminative unsupervised evaluation metric. However, the randomly sampled negative samples in the learning-based metrics are easy to distinguish from the positive references. To improve these metrics, Lan *et al.* (2020) introduce several strategies to obtain high-quality negative samples and more feasible positive samples.

Bak and Oh (2020) also demonstrate that the negative sampling strategy is critical for a learn-based metric like RUBER. Unlike uniformly sampling negative samples from the whole corpus, they propose constructing speaker-aware negative samples with four different difficulty levels to avoid the easily fitting problem. Figure 4.13 presents an example of obtaining different types of negative samples. The easiest negative samples ($Rand_A$) are randomly selected from utterances of other speakers (i.e., B and C) except for the current one, i.e., speaker A. The second type (SS_A) refers to the utterances of the current speaker A. The third category (SP_A) represents utterances of the current speaker A in conversations with the same partner, i.e., B or C. The most challenging negative samples are the utterances of the current speaker A from the same dialogue session.

Sai *et al.* (2020) construct a dataset that provides multiple positive responses for each data sample and find that such a dataset can enhance previous learning-based metrics in both performance and robustness. However, the enhanced learning-based metrics are still vulnerable to adversarial examples. Further experiments with the augmentation of PLMs also yield the same phenomenon, i.e., these metrics are easily attacked by adversarial examples.

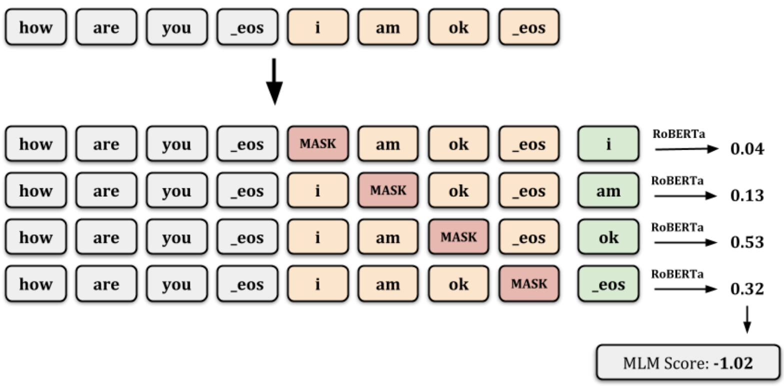

Figure 4.14: The MLM sub-metric in USR (Mehri and Eskenazi, 2020b).

USR. Mehri and Eskenazi (2020b) propose an unsupervised reference-free method for evaluating dialogue generation. They first introduce the MLM sub-metric to evaluate the naturalness of the generated responses from the perspective of language modeling, which is shown in Figure 4.14. Then, they extend the context-response matching component of RUBER with the enhancement of RoBERTa and different types of context information, e.g., facts, and knowledge, to construct the retrieval sub-metric. These two sub-metrics are combined by a regression model to evaluate the response from five different perspectives.

Others. Tong *et al.* (2018) propose jointly learning multi-lingual data by adversarial multi-task learning for enhancing the hidden representation of the dialogue utterance. Huang *et al.* (2020a) enhance the evaluation model by the knowledge graph. Mehri and Eskenazi (2020a) also build an unsupervised evaluation metric upon PLMs, i.e., DialoGPT. Chan *et al.* (2021) introduce a self-supervised evaluation method to enhance one-to-many evaluation in the latent space.

4.10 Summary

This section presents typical generation-based chit-chat techniques to address the challenges of diversity, context modeling, knowledge utilization and grounding, human factors learning, and performance evaluation. It can be observed that: (1) many of recent research still focuses on designing better evaluation metrics since there are no reliable and cheap metrics for evaluating chit-chat conversations, (2) except for consistency between context (dialogue history, extra knowledge), persona, emotion and response in retrieval-based chit-chat systems, generation-based chit-chat systems further encounter the challenge of generating controllable and expected content without ethical issues.

5

Ensemble-Based Chit-Chat Systems

As introduced in the previous two sections, there are two main kinds of chit-chat models: retrieval-based frameworks and generation-based solutions. The retrieval-based models collect a large number of conversations (query-response pairs). As shown in the upper part of Figure 5.1, they first recall a small response set from the conversations through a fast recalling manner. Then, the retrieval-based models rank the small response set to select the best response utilizing a complex but effective method. Different from retrieval-based methods, generation-based models trained on conversations can create a response in the fashion of word by word. To make full use of the advantages of both the retrieval-based and generation-based paradigms, the ensemble methods construct candidate responses from the two kinds of models in which the best response from the candidate responses is selected.

As presented in Section 3, the retrieval-based models select responses from a large conversation set where the responses in the conversations are written by humans literally. Thus, the obtained responses are fluent, natural, and reliable in practice. However, the retrieval-based models fail when lacking appropriate responses in the conversation set. The merits of these real-world human conversations guarantee the effectiveness of

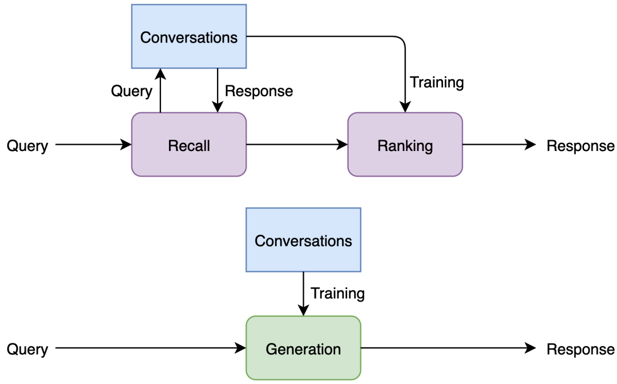

Figure 5.1: Comparison between retrieval and generation based models. The upper part represents retrieval-based methods, while the lower part demonstrates generation-based approaches.

retrieved-based methods, but meanwhile, it constitutes a bottleneck for the line of response-retrieval models.

Generation-based models involve generating responses given the query. Different from simple copy and reuse of existing human utterances, the generation models can learn to "create" responses by sampling words from a pre-defined vocabulary under the constraint of the conditional language model learned by the encoder-decoder framework. The sampled response space is, to some extent, unlimited (Yan, 2018) and thus, the generation-based models can handle more complex queries. However, in real-world applications, the generation models can not always guarantee to generate qualified responses. Sometimes the responses are short and meaningless, and sometimes they are diverse but unrelated to the query.

The pipeline of the two models are drawn in Figure 5.1. The conversations in the retrieval-based models is treated as the searching source for the recall process, and the ranking model is trained using the conversations. As for the generation-based models, the conversations are used to train the generation model only.

Considering the advantages and disadvantages, researchers combine the two kinds of methods into one "model" and name it as "ensemble model". Generally, ensemble models absorb the advantage of both models through reranking the results from retrieval and generation-based models, feeding the retrieved response into generation models, treating the retrieved responses as prototype editing, helping each other through adversarial training. And the results show that the ensemble approach is appealing in performance (Chen *et al.*, 2017).

5.1 Integration and Reranking Based Ensemble

The ensemble model is first introduced by Song *et al.* (2018), which is based on integration and reranking. As demonstrated in Figure 5.2, this type of model shares a similar pipeline of: 1) retrieving a small candidate response set from a large conversation set, 2) feeding the retrieved response into the generation model, and 3) reranking the retrieved and generated responses. The first step to retrieve the candidate response set is the same as the retrieval-based models.

The retrieved response is fed into the generation-based model to decrease the probability of universal replies, owning to the additional condition (retrieved response). Besides, the retrieved responses are written by humans, which makes the response a good guide to the generation model. To feed the retrieved response into the generated model, most of the researchers build a multi-encoder generation model. One encoder encodes the query, and the other encoder encodes the retrieved response. The retrieved responses are diverse and can be utilized to answer the query from different perspectives and in different ways. In view of this, Song *et al.* (2018) and Yang *et al.* (2019a) feed multiple retrieved responses into the generation model, instead of one response (Zhuang *et al.*, 2017). An attention mechanism is applied to assign weights for the retrieved responses. The representations of query and retrieved responses are used to initialize the decoder. The generation model leverages the same training strategy to typical response generation models. Besides, Song *et al.* (2018) utilize a copy mechanism to improve the generated response with helpful words that appeared in the retrieved response.

After obtaining the responses from the retrieval and generation-based models, the ensemble model reranks the responses to select the most appropriate one as the final response. The generation model may create an excellent response or a meaningless response. So it is necessary to de-

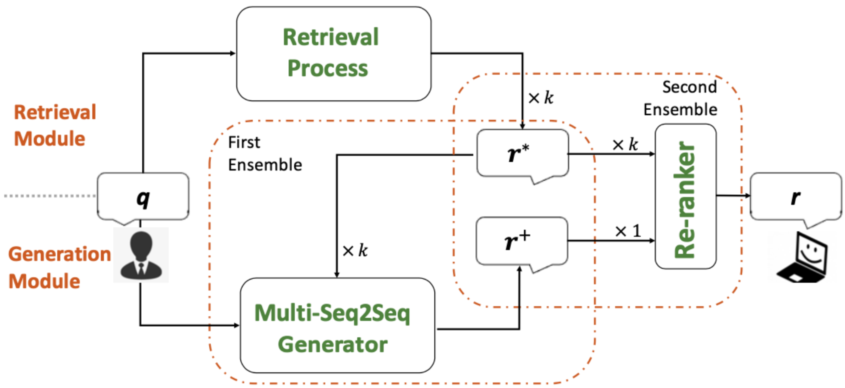

Figure 5.2: An example of integration and reranking-based ensemble model (Song *et al.*, 2018).

sign the reranking model to select the final response. There are different ways to rerank the responses: 1) constructing conversation-related features including term similarity, entity similarity, topic similarity, length, fluency (Song *et al.*, 2018) and ranking the response using the gbdt or xgboost classifier, and 2) training a neural network-based classifier model to evaluate the responses. For example, Cai *et al.* (2019b) train an interactive matching model to predict the matching degree between a query and a response. They build an interaction matrix based on the pairwise similarity of words within a query and the correlated response candidate. Then matrix is fed into a convolutional neural network (CNN) to predict the final ranking result.

This kind of ensemble model mainly helps the generation model with the retrieved response and rerank the responses from both the retrieval and generation-based models. Feeding the retrieved response into the generation model improves the quality of the generated responses. And the reranking process ensures the quality of the final response.

5.2 Template and Prototype-Based Ensemble

Template and prototype-based ensemble can be also categorised as generation-based models. In this survey, we treat chit-chat systems that involve both retrieval and generation components as ensemble methods, leaving pure retrieval-based and generation-based approaches in the previous two sections. When writing a paragraph of text, human tends to find a similar existing text as the template and produce the new text by editing the selected template text. The template helps people to build the skeleton of the paragraph, and it is only needed to incorporate the new contents into the skeleton. In order to facilitate the text generation model to produce fluent dialogue response, researchers propose the prototype-based generation methods, which firstly retrieve a similar dialogue response as the prototype and then edit the prototype by considering the current dialogue context semantics. Different from the prototype editing-based dialogue generation methods, it will be more straightforward to use the retrieved text as the final response to users. It is obvious that the text may have many irrelevant words and facts, which will confuse the user. Thus, the prototype-based model is a trade-off between retrieval-based and generation-based methods, and this method can give fluent and consistent responses.

Guu *et al.* (2018) firstly propose a prototype editing-based text generation method, where the generation process involves two different steps. To generate a sentence, they first sample a random prototype sentence from a pre-set training set and then inject an edit vector obtained by random sampling to edit the prototype sentence into a new sentence using an attention mechanism. Following this line, Wu *et al.* (2019b) bring the prototype-then-edit paradigm into the dialogue generation task, which first retrieves a prototype from a pre-defined index and then edits the retrieved prototype based on its context discrepancy with the target response. For the language modeling task in Guu *et al.* (2018), the editing vector is randomly sampled since their primary target is to build a generative language model by producing human-like sentences. In contrast, Wu *et al.* (2019b) take both retrieved dialogue context and current dialogue context into consideration when they revise a prototype response.

Previous methods use the latent edit vector to produce the new text. In contrast to these methods, Cai *et al.* (2019a) propose a skeleton-to-response paradigm in which the skeleton is extracted from the retrieved dialogues. However, when the retrieved response is irrelevant to the input query, the performance will drop sharply. A possible reason is that both the useful and useless information are mixed in the dense vector space, which is uninterpretable and uncontrollable. Cai *et al.* (2019a) employ the skeleton generator to extract a response skeleton by detecting and removing unwanted words in a retrieved response, and use the response generator to add query-specific details to the generated skeleton for query-to-response generation. Since the skeleton is produced by extracting and removing words, this method uses the reinforcement learning method to update the model parameters. However, previous skeleton-and-generation methods encounter the challenge of precisely extracting a skeleton and efficiently training a response generator based on the retrieved skeleton. Cai *et al.* (2019b) present a dialogue generation framework that utilizes an interpretable matching component to extract the skeleton and trains a separate generator to perform response generation based on the retrieved skeleton. One novel characteristic of this model is that the training of the skeleton extractor and the response generator is decoupled, yet they work cooperatively under the help of a retrieval system. Since there is no explicit response skeleton in general query-response pairs for training, they propose to employ an interpretable matching model for matching skeleton extraction.

So far, the introduced methods all concentrate on the chit-chat without incorporating any knowledge into the dialogue model, resulting in the model generating frequently occurred responses, e.g., "I don't know", etc. However, incorporating knowledge is also a difficult task for dialogue generation, where knowledge facts selection for a given context is still an obstacle. The widely acknowledged entity name matching method tends to retrieve uncorrelated facts. To address this challenge, Wu *et al.* (2020c) present a knowledge selection method and a knowledge-aware generative approach, named Prototype-KR and Prototype-KRG, respectively. The Prototype-KR component is responsible for retrieving and ranking relevant facts from the pre-retrieved prototype dialogues of a given query, based on the observation that knowledge facts in similar dialogues are also closely related. The Prototype-KRG module then completes the response generation process conditioned on the retrieved and ranked knowledge facts. Inspired by human writers that a textual draft is usually polished many times, Zhang *et al.* (2020b) introduce a polishing strategy to integrate the retrieval techniques into response

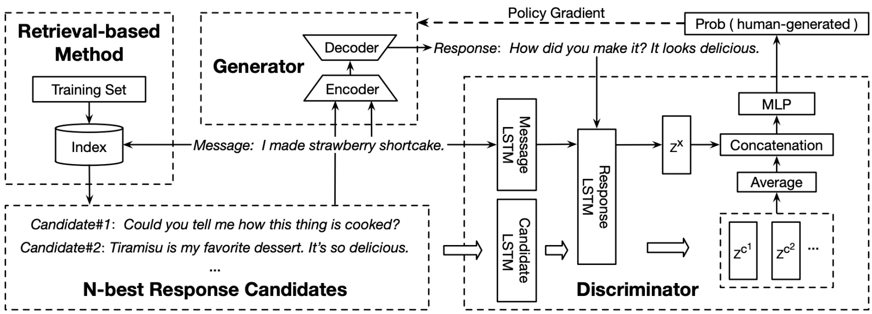

Figure 5.3: An overview of the adversarial based approach (Zhu *et al.*, 2019).

generation. Concretely, they use the retrieved response as a reference to polish the generated response draft, which is the same as human writers that take associated materials for reference to improve their draft.

Most previous works usually focus on generating fluent responses using the prototype-based method. However, the expression capability of dialogue systems towards a given style in producing conversations has a direct influence on their usability and user experience. Su *et al.* (2020b) propose a new prototype-to-style architecture to achieve stylistic dialogue generation. The proposed method first uses an information retrieval system to extract a prototype from the retrieved responses. Then, a stylistic response generator is introduced to create a stylistic response based on the extracted prototype and the given language style.

5.3 Adversarial-Based Methods

Most prior deep learning dialogue models approximate such a goal by creating the target response conditioned on the given dialogue history under the training objective of the likelihood maximization. Although these methods achieve success in many generation tasks and some of the methods achieve the state-of-the-art performance in benchmark dialogue datasets, such oversimplified training objectives will lead to some problems, e.g., generating dull, generic, repetitive, and short-sighted responses (Li *et al.*, 2017a).

A good dialogue model is supposed to produce content indistinguishable from human responses, i.e., the training objective of dialogue models should resemble the Turing test. Researchers borrow the idea of adversarial training (Goodfellow *et al.*, 2014) in dialogue systems, in which two models are trained in an adversarial fashion, including a generator to approximate the probability of creating a dialogue response and a discriminator to evaluate whether the constructed response is human-like. Intuitively, the simple idea is to combine the retrieval-based and generation-based dialogue systems under the generator-discriminator framework.

Unlike previous ensemble methods that mainly focus on a specific component, i.e., retrieval or generation, Zhang *et al.* (2019b) present an adversarial training scheme to mutually enhance each component of an ensemble dialogue model, in which the ensemble model comprises two generators and a discriminator. A sequence-to-sequence generator targets creating response candidates, and the other generator is designed to obtain hard negative samples for confusing the discriminator. The discriminator is trained to evaluate whether a sample pair is from the true data or adversarial candidates. Following this work, Yu *et al.* (2019a) propose a hybrid approach for open-domain dialogue generation. This model combines the advantages of retrieval methods and generative methods. The system aims to produce sequences that are indistinguishable from human-created sentences. Therefore, they employ adversarial training to the generative model and use reinforcement learning to optimize the model that involves non-differential modules. Zhu *et al.* (2019) propose an adversarial response generation framework enhanced by a retrieval-based method, where the overview model architecture is shown in Figure 5.3. Distinct from existing approaches, this method leverages the retrieved N-best response candidates to augment the training of both the encoder-decoder generator and the discriminator.

5.4 Summary

By comparing the above three types of "ensemble" chit-chat systems, we can conclude that: (1) **integration and re-ranking based ensemble can achieve better performance than either retrieval-based or generation-based method in the high probability** since these ensemble methods can pick up the best response from both retrieval-based and generation-based methods. The retrieved candidates serve as prototypes for the generation-based component. In turn, generation-based modules can provide more candidates, resembling data augmentation. (2) **Prototype retrieval based ensemble can achieve better performance than only utilizing generation-based module** with the augmentation of related human conversations. However, it is difficult to distinguish good or bad between this kind of ensemble and retrieval-based method, since retrieved response with high-quality might not need revision, while poor retrieved response can deteriorate the generation process. (3) **Adversarial training is an effective add-on for the above-mentioned two types of ensemble** insomuch as adversarial training can propagate back more loss signal of obtaining human-like responses.

6

Connecting Chit-Chat with Tasks

In the early years, dialogue systems were designed either for completing specific tasks or serving as entertainment tools with chit-chat conversations and non-limited topics. With the shared back-bone neural models and the evolution of user preferences, the boundary between task-oriented dialogue and chit-chat has been much more blurred in recent years. The task-oriented system can output task-independent content while chit-chat conversation can achieve an intended purpose. Since previous sections have thoroughly discussed and reviewed most of the representative chit-chat model frameworks in the research community and industry applications, we further discuss the connection between existing chit-chat systems and tasks in this section to sketch out the whole landscape of conversational AI. Generally, dialogue systems serve as an interface for users to interact with computers by human language. For task-oriented dialogues, chit-chat skills are useful in detecting user intent, making recommendations, and improving user experience for better engagement. We start this section with a brief review of completing tasks with dialogue systems, which has been systematically studied and compared in previous surveys (Gao *et al.*, 2019a), following with the connections between traditional task-specific systems and chit-chat ones. We then present the most important and successful combination paradigm of dialogue system and tasks in the IR community, i.e., combing conversation with search and recommendation. Besides, we also briefly review other emerging and representative cases of connecting dialogue systems with tasks, including conversational question answering and conversational machine reading. Moreover, we also discuss the possible research direction that unifies the framework of both chit-chat and task-oriented systems in the paradigm of pre-trained language models. This section ends with a discussion of better interleaving of chit-chat and task-oriented dialogues, bottlenecks to overcome, and future applications.

6.1 Linking Task-Driven Systems with Chit-Chat

6.1.1 Overview of Task-Oriented Dialogue Systems

Building task-oriented dialogue systems has been a long-range goal of AI community, and great efforts have been made by researchers from both academia and industry. To promote the development of this field, Gao *et al.* (2019a) sketched out the landscape of conversational AI in the last few years, mainly focusing on task-driven systems and neural approaches. As shown in Figure 6.1, the typical neural architecture of a task-oriented system consists of four modules to process text information, including intent understanding, dialogue state tracking, dialogue policy, and response generation. Before discussing the connections between chat-driven dialogue and task-specific systems, we first summarize and review each of the four modules with the notations and conversation examples from Daniel and James (2020).

Intent Understanding. Intent understanding component mainly involves **slot-filling**, **domain** and **intent classification**. Take the following user utterance for example (Daniel and James, 2020):

```
“O  O     O   O    O   B-DES  I-DES      O   B-DEPTIME  B-DEPTIME  O”
“I  want  to  fly  to  San    Francisco  on  Monday     afternoon  please”
```

the domain and intent of this utterance are computed as AIRLINE and SHOWFLIGHT, respectively, by a neural classifier that can process utterance, e.g., the combination of the BERT model and a feed-forward

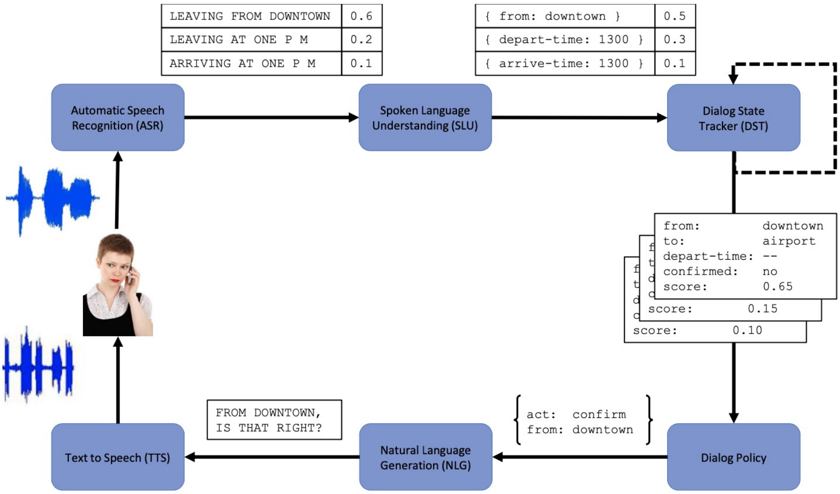

Figure 6.1: A representative framework of task-driven spoken dialogue system (Williams *et al.*, 2016).

layer. The slot filling task normally contains a sequence labelling process with predefined tags, say BIO tags, to predict the slots and conduct filler string extraction for each slot. As shown by the above example, the slot DEPTIME is tagged by the sequence labeler, and the filler string `San Francisco` is extracted for filling the slot.

Dialogue State Tracking. The function of dialogue state tracking is to compute the current dialogue state that includes both fillers of each slot and the most recent user dialogue act. Herein, the dialogue act corresponds to the interactive function of the turn or sentence, which is designed for each particular task. Take the restaurant recommendation system (Young *et al.*, 2010) as an example, the dialogue act "`REQUEST(a,b=x,...)`" refers to inquiry value for each given `b=x` in which `b` can be a slot name and `x` is the corresponded filler string. The slot-filling and dialogue act detection are launched jointly by the dialogue state tracker to process each user utterance, e.g., the example from Mrkšić *et al.* (2017) that

"`I'm looking for a cheaper restaurant`"

is processed as

"`inform(price=cheap)`"

where `inform` refers to dialogue act, and `price` and `cheap` are slot and filler, respectively. The newly coming user utterance will be converted with the constraints of the entire state of the frame at this point (the fillers of each slot). For instance, the next user utterance following the above-mentioned user turn is to convert

"`Thai food, somewhere downtown`"

into

"`inform(price=cheap, food=Thai, area=centre)`"

which is the simplest dialogue state tracker, and more sophisticated models can be found in Gao *et al.* (2019a).

Dialogue Policy. With the obtained dialogue state, i.e., the most recent dialogue act and slot-fillers, the dialogue policy predict the following dialogue act to be taken. Given the dialogue act sequences from dialogue system (A) and a user (U) before the current turn i of the conversation, the dialogue policy is supposed to predict the next dialogue action A_i, formulated by:

$$\hat{A}_i = \underset{A_i \in A}{\operatorname{argmax}} P(A_i | A_1, U_1, \ldots, A_{i-1}, U_{i-1}) \tag{6.1}$$

If the dialogue state is simplified by merely maintaining the set of slot-fillers, the computation of next dialogue action A_i is to maximize:

$$\hat{A}_i = \underset{A_i \in A}{\operatorname{argmax}} P(A_i | Frame_{i-1}, A_{i-1}, U_{i-1}) \tag{6.2}$$

where $Frame_i$ refers to the current state of the frame with slots filled by current fillers. A_{i-1} and U_{i-1} are the last turn dialogue acts of the system and user, respectively. The probability function of $P(\cdot)$ is parameterized by neural networks.

Low-Resource Setting in Generating Tasks Dialogues. Once the next dialogue action is computed by dialogue policy, the language generation module is introduced. Normally, natural language generation (NLG) comprises two stages, i.e., content planning and sentence realization. In application, content planning can be completed by dialogue policy component, and thus the natural language generation part mainly focuses on creating utterance sentences based on the dialogue act and slot fillers, which is exemplified by the following case from Nayak *et al.* (2017):

"`recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french)`
1 Au Midi is in Midtown and serves French food.
2 There is a French restaurant in Midtown called Au Midi."
where the latter two sentences are generated by the NLG module conditioned on the given dialogue action `recommend()` and these slot fillers.

As high-quality training data is expensive and difficult to collect for various sorts of problems, NLG task is usually under the low-resource setting. To address this limitation, the delexicalization strategy is presented, where slot values are replaced by placeholders to represent the correlated slots, as shown in the following (also from Nayak *et al.* (2017)):

"`recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french)`
1) `restaurant_name` is in `neighborhood` and serves `cuisine` food.
2) There is a `cuisine` restaurant in `neighborhood` called `restaurant_name`."
Thus, the NLG process is decomposed by first obtaining the delexicalized sentences for model training and then utilizing these data to train an encoder-decoder neural network for generating delexicalized sentences. These placeholders are replaced by the correlated slot values to output the final utterance.

6.1.2 Completing Tasks and Chit-Chat Dialogues

In real-world applications, chit-chat conversations typically consist of chat-driven contents alternating with task-oriented utterances. As shown in Table 6.1, the given synthetic conversation case represents many realistic conversation situations. The user starts a conversation with a chatbot in the demands of emotional engagement, and then the bot attempts to show empathy and divert the attention of the user, following with the task-oriented sub-dialogue of recommending and obtaining a first-person shooter video game running on an iPad. The

Table 6.1: An illustration of chat-driven conversation case in which the conversation starts with chit-chat and then shifts to goal-oriented dialogue.

Types	Roles	Utterances
Chat	User:	Hi ANONYMOUS-Bot, I feel a little bit depressed today?
	Bot:	What happened? Is anything wrong?
	User:	My SIGIR submission is got rejected. It is five months of hard work.
	Bot:	I am sorry to hear that. Do you know the reason for it?
	User:	Not yet. I just want to escape from this situation right now.
	Bot:	Yeah, you really should take a break.
Task	User:	OK, any recommendations?
	Bot:	How about play games at home, considering the outbreak of COVID-19.
	User:	That sounds okay! I used to play shooter video games.
	Bot:	Do you prefer first-person or third-person?
	User:	I have played first-person a lot. Third-person would be better.
	Bot:	You could try PUBG or PUBG mobile.
	User:	I have never played PUBG mobile before, and I just bought an iPad.
	Bot:	You can download it from App Store, and it's free.
	User:	Thank you! I can't wait to try this game.

task-oriented sub-dialogue will be completed by the aforementioned framework. With both the capabilities of dealing with chat-driven and task-driven situations and not a single one that can be omitted, the chatbot can obtain a better user experience and higher loyalty. Besides the chat-centric scenarios, the chat-driven conversation can, in turn, facilitate task-driven conversations. For instance, introducing more task-irrelevant and chit-chat contents in the iteration turn of goal-oriented user-machine conversation could make the dialogue system appear more intelligent rather than a machine that can merely complete specific tasks, resulting in more user engagement and trust. Many other situations also need chat-driven conversation to assist task-driven dialogues, e.g., recommendation, question answering, which will be illustrated in the next few sections.

Apart from the alternative appearance of task-oriented contents and chat-driven utterances, chit-chat and task-specific systems also encounter some common challenges, including but are not limited to conversational implicature, dialogue structure modeling, grounding, multi-turn and dynamic context modeling, and low-resource text generation. Besides, there is an emerging trend of transmitting from end-to-end neural framework to multi-stage pipeline for chat-driven systems, which resembles main-stream task-oriented systems in many stages, e.g., intend understanding (domain and intent classification, keywords extraction), dialogue analysis (discourse analysis, structure extraction), content planning, natural language generation. In turn, researches that complete task-oriented dialogues with end-to-end models are not rare recently. Furthermore, task-oriented and chat-driven systems share many back-bone models in the deep neural age such as encoder-decoder frameworks (Sutskever *et al.*, 2014; Vaswani *et al.*, 2017), and pre-trained language models (Devlin *et al.*, 2018; Zhang *et al.*, 2020f).

6.2 Conversational Search and Recommendation

Search and recommendation are the two main ways for humans to obtain information from the Internet, and they are also the two areas where AI is most widely used in practice. The purpose of search and recommendation tasks is to find suitable items or websites for users. The only difference is that the task of search is conducted by the user, i.e., user-initiative, where the user will post an explicit query, while the recommendation task is performed by the system, and the user will not give direct instruction.

With the consistent development of dialogue technology on the traditional single-round static search and recommendation system, many works on conversational search and recommendation systems have been aroused (Zhang *et al.*, 2021; Yang *et al.*, 2021a; Li *et al.*, 2021b). For example, compared with the traditional recommendation system, the conversational recommendation system can inquire about the user's attribute preferences and the attitude of the product through natural language and understand the user's feedback through natural language. Through multiple rounds of human-computer interaction, it will be more conducive for the system to accurately understand the user's preferences and find more suitable products.

A conversational search/recommendation system needs not only to recommend items based on the context of human-machine conversations but also to generate replies like a dialogue system. In order to enable the conversational search/recommendation system to have the above-

Conversation initiated by user `ID=AQGUDK0MSQ95L`
U: Can you find me a `tablet` on Amazon?
S: Sure, any requirement on the `network`?
U: Built-in free wireless data network.
S: Any preference on the `memory`?
U: 2GB of internal memory as well as a microSD expansion slot for additional memory.
S: Any preference on the `battery`?
U: Battery is removable and user-replaceable.
Result: Product `ID=1400532620`.

Figure 6.2: An example of template-based conversational recommendation session (Zhang *et al.*, 2018d).

mentioned functions, in recent years, researchers have explored from two perspectives, where the first is dialogue understanding and item retrieval, and the second is response generation.

6.2.1 Dialogue Understanding and Item Retrieval

In conversational search and recommendation, the system needs to understand the context of the conversation and predict which product a user prefers. Specifically, this task can be defined as: given the dialogue context information $\{d_1, d_2, \cdots, d_n\}$, the task is to predict the user's desired product i_k, where d_k is a round of human-computer interaction dialogue content, in which the content could be a sentence composed of natural language, or a recommended product i_k which belongs to set of all products $\{i_n\}_{n=1}^{N}$.

In early studies, researchers limited the natural language in the interaction process with templates and did not allow users or machines to use free natural language to communicate. For example, Zhang *et al.* (2018d) constrained the natural language interaction between humans and machines as the form shown in Figure 6.2. In the conversation session, the user's inquires and replies to the machine can only be in a given format, and it is the same for the questions and replies of the machine. Although this template-based interactive mode realizes conversational search/recommendation, there are many limitations and

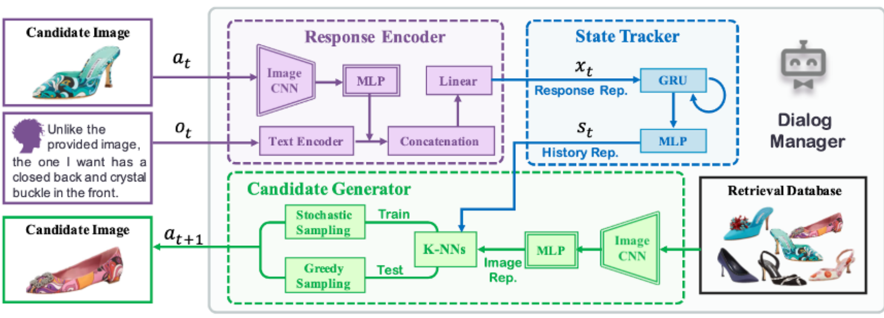

Figure 6.3: The proposed model framework in Guo *et al.* (2018).

inconveniences in practice.

In recent years, researchers have tried to remove the template-based restrictions and allow users to communicate with the system in unrestricted natural language, thereby improving the convenience and user experience of the search/recommendation system during conversations. Guo *et al.* (2018) consider the conversational recommendation task in the e-commerce shopping scenario. In their setting, each user has one desired item. The system interacts with the user in multiple rounds and recommends an item for the user in each round. Given the recommended item, the user gives feedback to describe the difference between the recommended item and the desired one. Based on the user's feedback, the machine will re-recommend the product until the machine recommends the expected product for the user. Specifically, the model consists of a dialogue encoder module that encodes each round of human-computer interaction dialogue, a state tracking module that integrates the dialogue content of each round, and a candidate generation module that generates candidates and retrieves items based on the dialogue state.

As shown in Figure 6.3, for the product a_t recommended by the machine to the user in the t_{th} interaction, and the text feedback information o_t which describes the difference between the product and the product that the user wants, the dialogue encoder module calculates the dialogue representation x_t. After that, the state tracker uses a GRU network to update the present dialogue state, which takes the t_{th} round dialogue representation x_t as input and calculates the updated historical representation s_t up to time t. Taking the historical representation s_t through time t as the query, the candidate generator calculates the similarity between s_t and each product representation and uses the K-NN method to generate a candidate set. According to the similarity calculated by the model, the item with the largest similarity $a_{(t+1)}$ is recommended to the user in round t+1.

During training, the authors pre-train a user simulator with additional data on the correlated image caption task to get the user's text feedback information after each step of product recommendation from the model. The related image caption task is to generate descriptions about differences given a target image and a candidate image. The pre-trained image caption model will be fixed during training and used as a user simulator to generate text feedback o_t in each round of dialogue. In addition, since the K-NN operation used to generate the candidate in the candidate generator module and the sampling operation are non-differentiable, the authors adopt a two-stage training method that consists of a supervised learning stage and a reinforcement learning process. In the supervised learning phase, the randomly initialized model is used to interact with the user simulator for multiple rounds by maximizing the probability of retrieving the user's desired item while minimizing the probability of retrieving a randomly negative item for each interaction turn. In the reinforcement learning phase, the model is initialized with the model obtained in the supervised learning phase. In each round of interaction with the user simulator, the reward is represented by the probability of recommending the desired item predicted by the model. The model is trained with the model-based reinforcement learning algorithm.

During inference, based on the probability distribution calculated by the candidate generator of the model in each round of interaction, the model selects the item with the largest score as the recommendation result until the item that the user desires is correctly recommended.

Although the method proposed by Guo *et al.* (2018) could interactively recommend items based on the user's natural language feedback in multiple turns and finally find the product that the user desires, there are still some shortcomings. The most important issue in the method proposed by Guo *et al.* (2018) is that, when the model interacts with users and recommends products in each round, it does not consider whether there is a deviation or a conflict between the products in this round and the content mentioned in the previous rounds of user feedback. To address this problem and avoid conflicts between the recommended products and the user's previously mentioned requirements, researchers have improved on the method proposed by Guo *et al.* (2018) and verified their approaches on the same recommendation data set.

In Yu *et al.* (2019b), the researchers pre-trained an MLP to predict the similarity between the current round of recommended product and the information related to the item attributes and item features mentioned by users in previous rounds of interaction. In accordance with Guo *et al.* (2018), Yu *et al.* (2019b) also apply reinforcement learning to train the product recommendation module. For the reward signal, in addition to the reward that measures the accuracy of recommended items, Yu *et al.* (2019b) incorporate the prediction of the pre-trained MLP as the reward. Zhang *et al.* (2020d) adopt the idea of adversarial learning and introduce a discriminator to determine whether the currently recommended product conflicts with what the user has said before. For each iteration in a training step, the model is first updated according to the recommendation loss and then updated based on the backward gradient calculated by the discriminator. After updating the model parameters at each training iteration, the discriminator is updated based on the correlated loss signal. In this way, the model simultaneously finds the products that users want through multiple rounds of interaction, and the recommended products in each round of interaction will not conflict with previous user feedback.

Knowledge Enhanced Approach. To correctly recommend products to users, one important issue of the conversational recommendation system is that the model must have a sufficient understanding of the characteristics of each product itself. Researchers have proposed to incorporate knowledge graphs to model attribute information on items. Meanwhile, knowledge graphs not only allow the model to have an understanding of the product but also can be used to model other entity information mentioned in user-system interactions to help the machine better understand the dialogue context.

Sarkar *et al.* (2020) proposed to introduce DBpedia as an external knowledge to improve the performance of movie recommendation tasks. Aiming at the contextual information of the interaction between the user and the machine, this paper uses each entity mentioned in the dialogue to obtain sub-graphs from the knowledge graph through multi-hop propagation or the page-rank algorithm. The model calculates the graph embedding of these entities in the sub-graph and uses the attention mechanism to integrate the embedding to calculate the user representation. Based on the user representation that incorporates knowledge information, the model can effectively recommend items to the user. Experimental results also confirm that the recommendation accuracy is significantly improved by incorporating the knowledge graph compared to only using the training data set.

User Memory Modeling. To improve the accuracy of recommendations, researchers not only consider incorporating external knowledge to learn more strong representations items, but also focus on how to model the user's historical behavior information.

Xu *et al.* (2020a) propose to model each user's historical interaction item-sequences and the attribute information. This information is modeled in a graph structure, which is denoted as a user memory graph. In the process of interacting with users and recommending products, the model first uses the BERT model to encode the chat content information in the current session to obtain a vector representation and then uses an inference module based on R-GCN. The content and preference information in the user memory graph are used to infer user preferences and make product recommendations. In addition to predicting the expected products, the model also predicts information such as slot, dialogue act, and product attributes of the user's chat content and uses multi-objective learning to enhance the model performance further.

PLMs in Conversational Search/Recommendation. In addition to studying how to introduce external resources such as knowledge graphs and user historical behaviors to improve the recommendation accuracy of dialogue recommendation tasks, some researchers also focus on applying pre-trained language models in dialogue recommendation.

Penha and Hauff (2020) propose to detect whether the pre-trained language model has learned the knowledge to solve the dialogue recommendation task. For BERT and Roberta, they first verify whether the models obtained directly after pre-training without fine-tuning have an understanding of the characteristics and content of the product, as well as the similarity information of products. Then, the product recommendation performance was explored. For understanding the characteristics of a product, i.e., the knowledge of the content, the authors draw on the masked language model task in BERT for probing, i.e., designing a problem similar with Devlin *et al.* (2018) to evaluate the model performance in terms of product characteristic understanding. As for the ability to recommend similar products based on specific products, i.e., generalization ability, the authors explore from two perspectives. On the one hand, the authors borrow the idea of the next sentence prediction task in BERT model training. A template as "If you liked The Hobbit, [SEP] you will also like Lord of the Rings" is designed, and the probability that the model retrieves the correct sentence correctly is calculated. On the other hand, for sentences such as "It gives a brilliant picture of three bright young people." and "The Brothers Karamazov.", they use the vector representation from the position of the CLS token calculated by the BERT model to compute the similarity for testing whether the model can correctly distinguish these products. Experimental results show that the pre-trained model has already captured some knowledge about books, movies, and music, without a fine-tuning process on specific data.

Besides, the authors also investigate the performance of the BERT model with an extra fine-tuning stage on the dialogue recommendation dataset, thereby constructing a search-style dialogue recommendation model. Experimental results demonstrate that directly launching fine-tuning on the recommendation data can lead to several notable issues. It is easy to be attacked by adversarial samples. The researchers also found that if the masked language model for product features and the next sentence prediction for product relevance are added at the same time for fine-tuning BERT, the introduction of external knowledge to the model can be relieved to a certain extent. The model simply learns to capture the shortcomings of the pattern and significantly improves the accuracy of the recommendation.

6.2.2 Response Generation

In a conversational search/recommendation system, the model not only needs to correctly recommend products according to the content of the conversation but also needs to interact fluently with users, especially for natural language. Interacting with the user in human language, on the one hand, allows the machine to ask user preferences for a better understanding of user intentions, and on the other hand, brings a better user experience by recommending customized products according to each dialogue sentence. To build a conversational search/recommendation system that can not only interact with users in natural language but also recommend products for users, it is necessary to make the model have the functions of reply generation and item retrieval at the same time, and it must be able to make accurate judgments during the interaction, i.e., whether the response should be generated at present or product recommendation should be made.

Li *et al.* (2018) explore the field of movie recommendation and builds the first conversational recommendation system that can interact with users in natural language. The model comprises an LSTM-based HRED encoder to encode the dialogue context information and an LSTM decoder for reply generation. In addition, the authors use the movieLens data set to pre-train a recommendation module based on denoising auto-encoder. In order to enable the model to determine whether to output a movie or a word at each step when generating a reply, the model introduces a switch mechanism in the decoder, i.e., a gating mechanism is used to determine the current decoding or recommendation. If the model judges that it is to decode, it will select a word in the vocabulary for output. Otherwise, the model first maps the hidden state in the decoder to a semantic space about all movies, i.e., calculating a vector with the dimension of the number of movies. The value of each dimension is between 0-1, and the vector is used as the input of the recommendation module to predict the recommendation result. In the training process, the model adopts the teacher forcing strategy. During the test, the model determines whether it is to decode

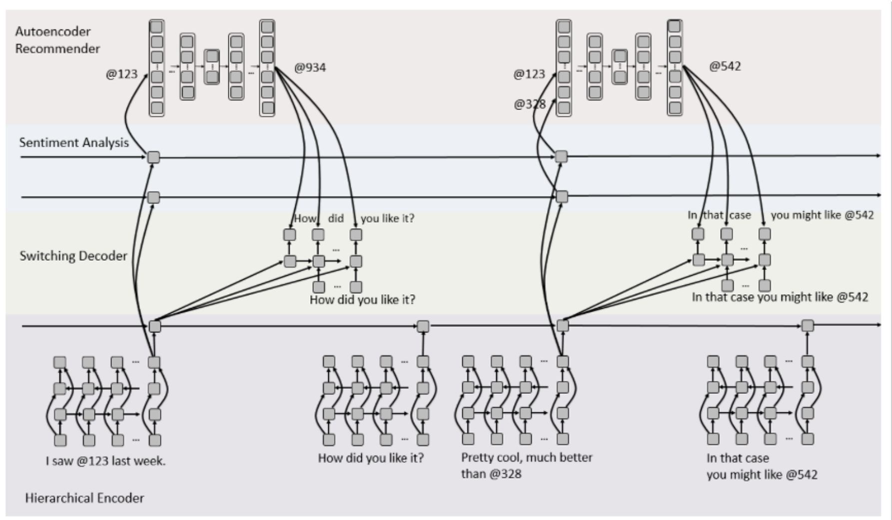

Figure 6.4: The model framework in Li *et al.* (2018).

or recommend according to the results of the switch gate at each step of the decoding.

Although in Li *et al.* (2018), the model can generate response and product recommendation at the same time and can automatically determine whether it should be decoding or recommendation during operation, this method is only limited to a given entire context and let the model generate a sentence with the recommended product reply to this mission scenario. However, in a real dialogue recommendation scenario, the machine needs to be able to conduct a complete dialogue with the user from beginning to end and recommend products in the dialogue, so it is not enough to only consider the given context for response generation and product recommendation. Based on this shortcoming, Kang *et al.* (2019b) explored how to allow the machine to conduct a complete dialogue with the user from beginning to end. To achieve this goal, a two-stage training method combining supervised learning and reinforcement learning is adopted in the paper. In the supervised learning stage, the author uses actual corpus to train two symmetrical models in the teacher forcing mode. In the reinforcement learning stage, the author lets the two models chat from beginning to end until they find the desired product. During the training process, the authors introduce rewards related to response quality and recommendation accuracy and use policy gradients to update the model. In this way, after the reinforcement learning stage, the model that plays the role of a machine can interact with multiple rounds of natural language from start to finish and predict the products the user wants during the interaction. Similar to the works in item retrieval, in the area of how to better construct a conversational search/recommendation that can both generate replies and retrieve items, researchers also focus on how to base on external resources such as external knowledge and historical user behavior to improve performance on recommendation accuracy and response quality.

Knowledge Enhanced Approaches. In recent years, many works have been conducted on various forms of external knowledge, including knowledge graphs, knowledge about product characteristics/attributes, and knowledge about text topics.

In the studies based on knowledge graph, Chen *et al.* (2019) and Liao *et al.* (2019) use the commodity and other entity information that appear in the chat context to obtain the corresponding sub-graphs from DBPedia, calculate the graph representation of these entities, and use them in the recommendation module of the model. Compared with the representation of the product only learned from the training set, the product graph representation calculated by introducing the knowledge graph is more conducive to capturing the attributes of the product itself and the relationship between products. Therefore, it can bring better recommendation performance. At the same time, when the model is generating a reply, it will also take the vector obtained by integrating the representation of the product in the context through the attention mechanism as part of the input. In this way, the model can also be assisted by the knowledge graph when replying. On this basis, Zhou *et al.* (2020a) not only consider the knowledge graph DBPedia related to commodity attributes and relationships but also consider the knowledge graph of word level. The conceptnet is introduced in the article, from which the sub-graphs corresponding to the words appearing in the context are obtained, and the graph representation of these words is calculated. Through such a modeling method that considers both item-related knowledge and word-related knowledge, the model has been further improved.

Focusing on the relevant knowledge of product attributes and characteristics, Liao *et al.* (2018) use the category tree of the product and the picture information to pre-train a matching model so as to learn the high-quality representation of each product with attribute and feature information and use these representations. In the recommendation and response generation, the accuracy of the recommendation and the richness of the content of the response generation can be improved.

In terms of topic-related knowledge, Zhou *et al.* (2020a) introduce topic-related information for each sentence in the chat context and cut it into the model training process. In addition to letting the model do response generation and product recommendation tasks, it also adds a topic prediction task. That is, which aspect of the sentence should be said in the current interaction of the prediction model. In this way, the model can be more specific about the current context information, which content should be generated in response to, thereby improving the performance of the model.

Personalized Information Modeling. In addition to incorporating external resources to enhance the conversational search/recommendation system, enhancing conversational search and recommendation by modeling user's personalized information is also a popular research direction. In conversational search and recommendation tasks, the user's personalized information mainly includes two aspects, i.e., user profile and user history behavior. Zhou *et al.* (2020b) model the two aspects of personalized preference information jointly. In this paper, the user's historical review information is represented by the user's profile for the movie recommendation scenario. The user's historically watched movie records are used as the user's profile. For the user profile, the model uses a Profile-Bert module as the encoder. The authors apply the ideas in the sequential recommendation for historical behavior information and use the SASSeq encoder module to encode historical behavior information. With the vector representation encoded from the two parts of personalized information, i.e., the historical behavior and user profile, the response generation and product recommendation

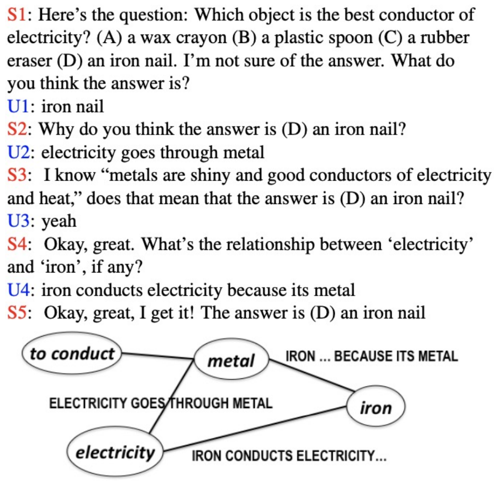

Figure 6.5: A real conversation from Hixon *et al.* (2015), and the task is formulated as generating the bottom knowledge graph based on the upper conversation utterances.

performance could be improved.

6.3 Conversational Question Answering

Many efforts have also been devoted to combining question answering with chit-chat conversation task (Qu *et al.*, 2020). Most of this research can be placed into three groups:1) seeking to leverage extra conversation task to assist question answering (QA), i.e., QA-centric setting; 2) detecting and addressing question answering sub-tasks from chat-driven conversations to enhance user experience of chatbots; and 3) completing question answering in a conversation-like manner.

QA-Centric. Hixon *et al.* (2015) presented a representative system to enhance the question-answering task by introducing an open conversation task. Through chatting with users about specific questions, the system can automatically construct knowledge graphs and, in turn, improve the performance of the QA task. Figure 6.5 presents a typical user dialogue that aligns utterances and knowledge graphs. Specifically, the authors collected 107 science questions of the 4-th grade New York Regents exam (Clark *et al.*, 2014), where each question is paired with four possible answers. The conversation task is initialized to ask users to choose an answer candidate for a specific question and present their explanations for their answers. Then, two different dialogue strategies are introduced to keep the conversation going until a knowledge graph is constructed that can support the answer of the user. There are three types of knowledge graphs at different levels built from dialogues, including utterance-level, dialogue-level, and global knowledge graph. To obtain utterance-level knowledge graph (uKG) whose nodes are all concepts in an utterance, two constraints are used to prune edges, in which only salient relations can be reserved. The obtained uKGs are then merged into a dialogue knowledge graph (dKG) with a sentence alignment strategy. Finally, all dKGs are added to the global knowledge graph with the enhancement of a relation filter.

Conversation-Centric. Wang *et al.* (2018) study the task of asking questions in chit-chat conversation systems, which is different from the traditional question generation task of question answering, including question patterns, topic scopes, and diversity. The authors find that good questions mainly consist of interrogatives, topic words, and ordinary words, where interrogatives lexicalize questioning patterns, while topic words and ordinary words capture topic transition and grammatical constraints, respectively, as shown in Figure 6.6(a). Then, two typed decoders are devised to generate more meaningful questions in the large-scale chit-chat conversation system. Wang *et al.* (2019b) focus on improving the semantic coherence between generated question and the paired post and answer, and meanwhile preventing generating dull questions, as shown in Figure 6.6(b). Laban *et al.* (2020) further build a chatbot system to create conversations with a user about the news.

Conversation-Like QA. Information seeking and gathering through conversation is essential for humans, which involves a sequence of interconnected questions and answers (Reddy *et al.*, 2019). To study the conversation-like question answering task, the authors introduce a dataset named CoQA for building and evaluating conversational QA systems. Figure 6.7 presents an example conversation from the CoQA dataset. Most works of conversational QA follow a similar setting from

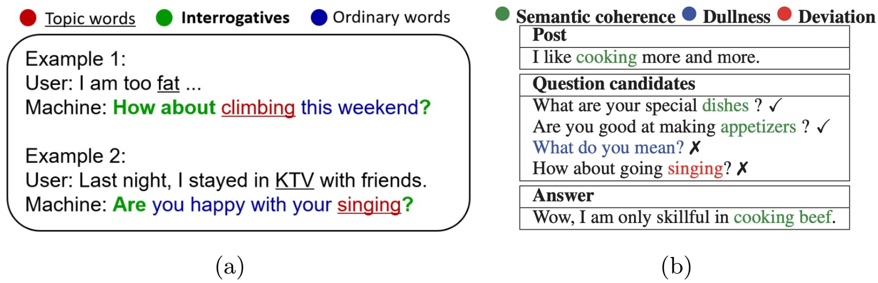

Figure 6.6: Conversation examples from chat-centric conversational question answering (Wang *et al.*, 2018; Wang *et al.*, 2019b). The left part illustrates the different composition patterns of questions in large-scale chit-chat conversation systems, while the right shows that question generation in chit-chat conversation requires semantic coherence between generated question and the corresponded post and answer and needs to avoid dull questions.

then on. Shen *et al.* (2019b) propose a multi-task learning approach, comprising a semantic parsing module enhanced by pointer and a type-aware entity detection method, to achieve conversational question answering on the large-scale knowledge base. Pan *et al.* (2019) propose a reinforced dynamic reasoning network to address the challenging conversational question generation task, where each question is generated conditioned on a passage and a conversation history. Kaiser *et al.* (2020) explore the setting of conversational question answering on passages with a word proximity network from large-scale corpora. Baheti *et al.* (2020) utilize data augmentation to enhance the fluency of generated answer response in conversation QA. Kundu *et al.* (2020) investigate the sub-task of identifying follow-up questions in conversational QA. Other perspectives of conversation QA such as feedback-weighted learning (Campos *et al.*, 2020), the comparison of learning to reason and exploiting patterns (Verma *et al.*, 2020), informative and specific question generation (Qi *et al.*, 2020), the intricate relationship between question reformulation and answer selection (Vakulenko *et al.*, 2020), and multi-task learning with dynamic task weighting (Kongyoung *et al.*, 2020) have also been studied recently.

As one of the typical conversational QA tasks, many existing works have explored conversational machine reading comprehension (MRC)

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 6.7: A conversation from CoQA dataset (Reddy *et al.*, 2019), where Q_i, A_i, R_i refer to question, answer, and rationale, respectively.

from different angles (Gupta *et al.*, 2020), including interpretation of natural language rules (Saeidi *et al.*, 2018), entailment-driven extracting and editing (Zhong and Zettlemoyer, 2019), the combination of response generation and machine reading in creating contentful responses (Qin *et al.*, 2019), enhancing conversational MRC with multi-perspective convolutional cube (Zhang, 2019), multiple-choice reading comprehension (Sun *et al.*, 2019), the incorporation of pre-trained language models (Ohsugi *et al.*, 2019), conversation flow utilization via graph neural network (Chen *et al.*, 2020b), and coarse-to-fine reasoning with explicit memory (Gao *et al.*, 2020).

6.4 Connections in the Era of Pre-trained Language Models

In the past few years, with the rapid growth of pre-trained language models (PLMs) (Devlin *et al.*, 2019; Liu *et al.*, 2019b; Lan *et al.*, 2019; Radford *et al.*, 2018a; Radford *et al.*, 2019; Brown *et al.*, 2020; Lewis *et al.*, 2020; Raffel *et al.*, 2020), the paradigm of NLP has shifted dramatically. The huge changes brought by PLMs have made various NLP tasks, including chit-chat, more closely linked. Therefore, in this section, we introduce the latest research progress of PLMs, their applications on dialogue systems, and potential directions of dialogue systems in the paradigm of PLMs, which may be a lot of help to the IR community.

6.4.1 The Paradigm of Pre-trained Language Models

Considering its great power, PLMs received great attention from Natural Language Processing (NLP) community and develop rapidly (Zhang and Li, 2021). PLMs are usually stacks of multiple Transformer (Vaswani *et al.*, 2017) layers. At the pre-training stage, PLMs are learned on large-scale textual corpora with unsupervised objectives, e.g., masked language modeling (MLM) (Devlin *et al.*, 2019; Liu *et al.*, 2019b; Lan *et al.*, 2019), casual language modeling (CLM) (Radford *et al.*, 2018a; Radford *et al.*, 2019; Brown *et al.*, 2020), denoising (Lewis *et al.*, 2020; Raffel *et al.*, 2020). By using these unsupervised objectives, PLMs can acquire abundant syntactic (Hewitt and Manning, 2019), linguistic (Jawahar *et al.*, 2019), semantic (Yenicelik *et al.*, 2020) and world knowledge (Petroni *et al.*, 2019), which confirms the effectiveness of pre-train tasks. By simply using the supervised data of downstream tasks to train PLMs, the knowledge contained in the PLMs can easily be adapted to the downstream tasks at the fine-tuning stage. Until now, pre-training and then fine-tuning have achieved state-of-the-art performance on almost all NLP tasks, which shows it has become the dominant paradigm of the NLP community. However, even though conventional fine-tuning methods have achieved huge success, the gap between pre-training and fine-tuning in data size and training objectives restrict the capabilities of PLMs, which needs a better way to make full use of PLMs.

Before the age of PLMs, to achieve better performance, a lot of work focused on reformulating the format of one task to another in the NLP community, e.g., reformulating text classification task (Yang *et al.*, 2018) or relation exaction (Zeng *et al.*, 2018) as sequence generating task, summarization task as question answering task (McCann *et al.*, 2018), and parsing task as language modeling task (Charniak *et al.*, 2016). In the Paradigm of PLMs, since effectiveness of pre-training tasks, reformulating the downstream tasks to the format of pre-training tasks, which is named prompt-based learning[1], has excellent potential and gains more and more attention. To make it clearer, we take a simple sentiment classification case: Given a sentence *"This movie is great!"* as input, prompt-based learning first append a prompt to the sentence, which makes the sentence become *"This movie is great! The sentiment of this sentence is [MASK]."* Then, the expected classification outputs are extracted from the MLM head's word predicting probability at the position of [MASK]. Specifically, each label's score is achieved by the predicting probability of its preset corresponding label word (e.g., 'positive', 'negative'). Due to the great power of PLMs, prompt-based learning has incredible progress: Schick and Schütze (2021a) and Schick and Schütze (2021b) reformulate text classification, entailment, and question answering as mask language modeling task, which achieves remarkable performance at the few-shot setting; Brown *et al.* (2020) concatenate examples with labels to the end of input (called in-context learning), which do not need to calculate the gradient of parameters and achieved promising results; Raffel *et al.* (2020) reformulate all NLP tasks to sequence generation tasks and propose a seq2seq PLM (T5) to address all the NLP tasks; to further improve the generalization ability and zero-shot performance of PLMs, Wei *et al.* (2021) and Sanh *et al.* (2021) propose multitask prompted training, which reformulates a large number of downstream tasks' supervised data as the format of pre-trained tasks and then finetune PLMs on these large-scale transformed data.

With the deepening of the research, the capability and generalization ability of PLMs increase rapidly and bring considerable changes in the NLP community. As Figure 6.8 shows, with the rapid development of PLMs (since 2018), the trend of reformulating and unifying in natural language processing tasks accelerated abruptly. Since the link between different NLP tasks becomes close, it is possible to unify chit-chat and

[1]More related work and other perspectives can be found in the recent surveys by Liu *et al.* (2021) and Sun *et al.* (2021)

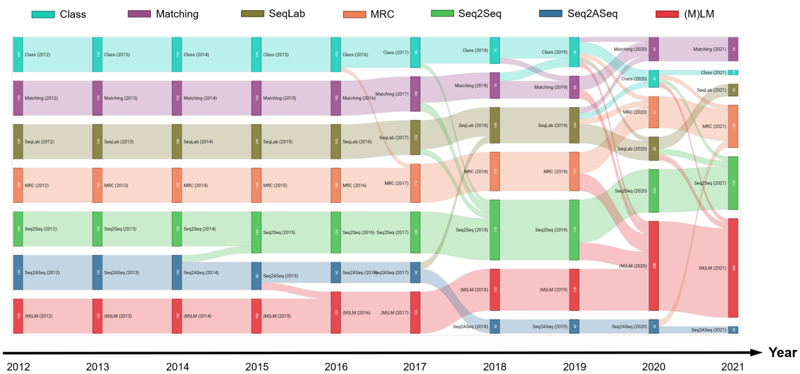

Figure 6.8: Sankey diagram to depict the trend reformulating and unifying in natural language processing tasks (Sun *et al.*, 2021).

others tasks.

6.4.2 PLMs for Dialogue Systems

Since the great potential of PLMs, applying PLMs to dialogue systems gains more and more attention from conversation AI communities. Due to the different characteristics of chit-chat and task-oriented tasks, the difficulty and problems of applying PLMs are also different.

Chit-Chat. Due to the distribution difference between dialogue corpora and plain texts, fine-tuning PLMs on dialogue corpus directly is not as effective as other NLP tasks. Therefore, there are a lot of works focusing on constructing dialogue-specific PLMs. Due to the high annotation cost of realistic dialogue corpus, researchers usually collect comment chains from social media (Reddit, Twitter, etc) instead. Zhang *et al.* (2020f) release DialoGPT, which is inherited from GPT-2 (Radford *et al.*, 2019) and further trained on dialogue corpus. Adiwardana *et al.* (2020) propose Meena, which uses a seq2seq model and surpass the chit-chat system that has a complex handcrafted framework. Roller *et al.* (2021) release Blender, which also employ a standard seq2seq transformer architecture has three variants with 90M, 2.7B, and 9.4B parameters. Bao *et al.* (2020a) propose PLATO, which is carried out

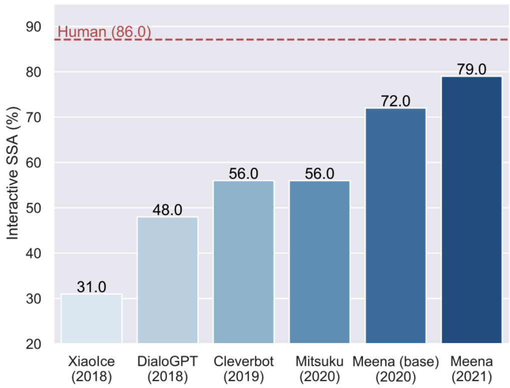

Figure 6.9: Manual evaluation on dialogue systems (Han *et al.*, 2021).

latent recognition task on large-scale dialogue corpus to address one-to-many problems. Bao *et al.* (2020b) introduce curriculum learning and scale-up PLATO to PLATO-2, which has three model sizes: 1.6B, 314M, and 93M. Bao *et al.* (2021) present the PLATO-XL with up to 11B parameters, which conducts multi-party aware pre-training. Qi *et al.* (2021) extend ProphetNet to chit-chat dialogue tasks and propose two models (ProphetNet-Dialog-En and ProphetNet-Dialog-Zh). Wang *et al.* (2020a) release CDialGPT, which is trained on the 12M Chinese chit-chat conversations. Zhou *et al.* (2021a) build EVA, which is a chit-chat dialogue system and contains a pre-trained dialogue model with 2.8B parameters. As shown in Figure 6.9, with the growth of data size and parameters, the PLMs achieve better performance on chit-chat tasks.

Task-Oriented Dialogue. Different from chit-chat, the task-oriented dialogue has explicit goals, which usually needs a modularized pipeline for more interpretability and controllability. Besides, the annotating cost of task-oriented dialogue is higher than chit-chat. Therefore it is more challenging to apply PLMs into task-oriented dialogue. Faced with the complexity of the pipeline and the data scarcity problem, Ham *et al.* (2020) and Hosseini-Asl *et al.* (2020) treat the inputs and outputs of all modules as single sequences and then use PLMs to optimize the modules in an end-to-end method (framework is shown in Figure 6.10);

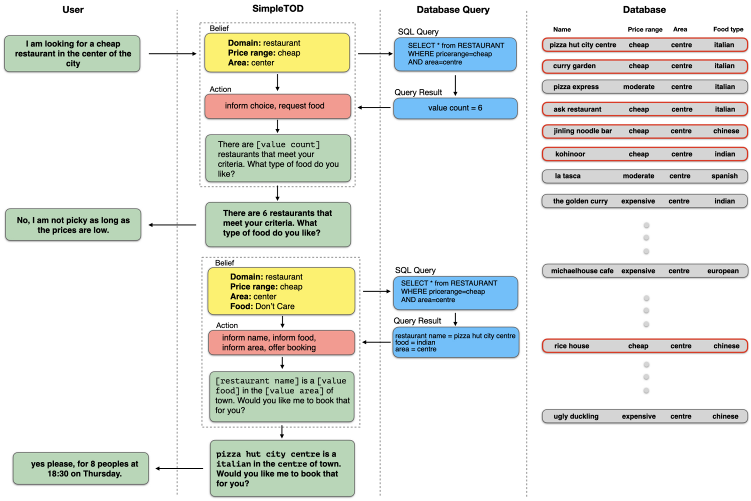

Figure 6.10: Framework of SimpleTOD (Hosseini-Asl *et al.*, 2020).

Lin *et al.* (2020b) and Peng *et al.* (2020) leverage PLMs to jointly learn DST and dialogue response generation; Yang *et al.* (2021b) fine-tune PLMs on dialogue session level instead of dialogue turn level; Su *et al.* (2021) utilize specific prompts to reformulate the format of sub-tasks of task-oriented dialogue and introduce a multi-task pre-training strategy that uses heterogeneous dialogue corpora; Wu *et al.* (2020b) release TOD-BERT, which is an adaptation of BERT trained on multiple task-oriented datasets with different domains and has a clear advantage on few-shot experiments; Mi *et al.* (2021) propose a self-training approach, which iteratively labels the most confident data from unlabeled dialogue corpus and achieves remarkable performance at few-shot setting; He *et al.* (2021) employ a semi-supervised method to explicitly learn dialogue policy and introduce a consistent regularization term to better use unlabeled data. Overall, it has great potential to use PLMs to unify different sub-tasks of task-oriented dialogue.

6.4.3 Towards Unified Dialogue System

As stated in the previous section, with the rapid growth of PLMs, unifying different tasks becomes a trend. Along with this trend, it is a promising direction of constructing a unified dialogue system that uses one model to handle chit-chat, task-oriented dialogue, and even work as knowledge bases, which may transfer the huge changes brought by the PLMs to the IR community.

Unifying Chit-Chat and Task-Oriented Dialogue. With the help of powerful PLMs, it is more and more possible to unify chit-chat and task-oriented dialogue. Zhao *et al.* (2021) propose a unified dialogue system (UniDS), which unifies chit-chat and task-oriented in a schema. Concretely, based on the end-to-end PLM frameworks for task-oriented dialogue, they see chit-chat as a type of dialogue policy and fine-tune the PLMs on the mixed data of chit-chat and task-oriented dialogue. Because of the capability to switch between two types of dialogues, UniDS is more robust than previous approaches. With large-scale pre-training and 11B parameters, PLATO-XL (Bao *et al.*, 2021) achieve SOTA results on both chit-chat and task-oriented dialogue datasets, which is strong enough to be the foundation model of conversational AI. Madotto *et al.* (2021) create a chatbot, i.e., the Few-Shot Bot (FSB), which can be adapted to handle different dialogue tasks without training, whose framework is shown in Figure 6.11. To be more specific, based on in-context learning (Brown *et al.*, 2020), they concatenate examples at the start and propose a strategy to select appropriate task-specific prompt to transform the format of the dialogue context. The most significant advantage of FSB is that it is based on general PLMs other than dialogue-specific PLMs, which avoids the annotation cost of dialogue corpus. With the exploration of unifying chit-chat and task-oriented dialogue, it is more and more possible that construct a unified dialogue system.

Language Model as Knowledge Bases. To explore whether PLMs can work as knowledge bases, Petroni *et al.* (2019) propose a benchmark named LAMA and use 'fill-in-blank' problems to probe the world knowledge from PLMs, which shows PLMs has learned abundant world knowledge during pre-training.

To test the ability of the PLMs as a knowledge base, Roberts *et al.* (2020), Lewis *et al.* (2021), and Wang *et al.* (2021) explore closed-book

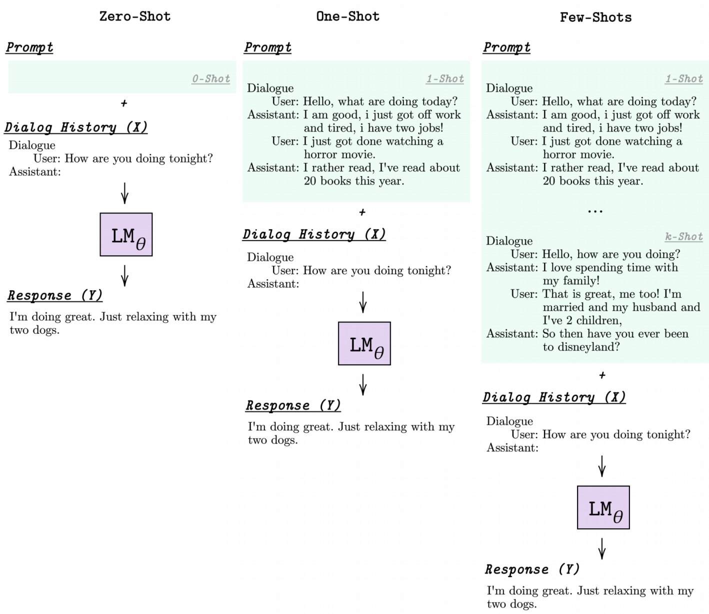

Figure 6.11: Framework of FSB (Madotto *et al.*, 2021).

QA tasks, where models need to answer the question without the help of an external knowledge base. In the dialogue tasks, there is also some pioneering work to explore using PLMs as knowledge bases to generate responses. Tuan *et al.* (2020) fine-tune a language model as a knowledge generation model; Xu *et al.* (2021) inject knowledge into lightweight adaptors named KnowExpert and leverage these knowledge to generate informative response so as to avoid suffering the slow speed of the retrieval process, whose framework is shown in Figure 6.12; Zhou *et al.* (2021b) utilize PLMs to generate relevant knowledge explicitly at first and then generate a response, which is named 'think before talk'. As response generative approaches are more and more critical in dialogue systems, with tremendous progress made by utilizing PLMs as knowledge bases, it is also a promising direction that uses generative approaches to get relevant knowledge passages.

In conclusion, because of the great potentials of PLMs, constructing a unified dialogue system, which is beyond chit-chat, becomes more and more hopeful.

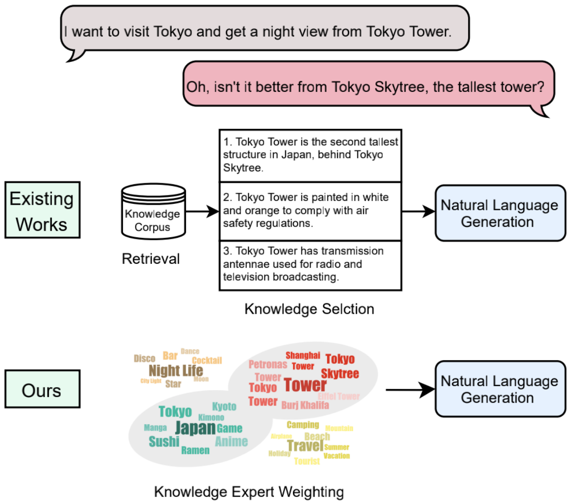

Figure 6.12: Illustration of KnowExpert (Xu *et al.*, 2021).

6.5 Discussions

We can conclude that chit-chat links tasks in two aspects. From the perspective of system combination, the chit-chat module is a vital supporting component of the task-oriented system, which can improve user experience in completing tasks. From the side of techniques, it is also possible to build a unified framework that can support both chit-chat conversations and tasks in the era of large-scale pre-trained language models. It can be imaged that, interleaving chit-chat and task-oriented dialogues will be more prevalent in the near future. Through combining the merits of chit-chat and traditional task-specific conversations, conversational AI systems will be competent for various types of tasks in a more natural and intelligent way. However, there are some obstacles to overcome. The first challenge is the lack of enough training examples that have both task-specific and chit-chat annotations. Another problem is the generalization of task-specific knowledge, such as knowledge graphs, and domain-specific document texts.

7

Conclusions

7.1 Current Progress

In witness of the resurgent and rapid ascend of chit-chat systems, many works based on deep learning have been presented in the last few years. In this monograph, we survey these recently released papers from varied aspects, and the main progress can be summarized as follows:

- Current chit-chat systems mainly leverage three different types of solutions, e.g., retrieval-based methods, generation-based approaches, and the ensemble of these two types of solutions.
- The most challenging problems that we are encountering in building chit-chat systems are long-range context modeling, one-to-diversity, human factors learning and fusion, knowledge and grounding, and the combination of pre-trained language models. Recent studies have made substantial progress on these problems.
- We sketch the landscape of conversational AI in the age of deep learning from the perspective of chit-chat.
- We discussed the connections between chit-chat dialogue systems and conventional task-oriented conversation systems, as well as newly emerging IR tasks.

7.2 On-Going Struggles and Possible Future Trends

Through analyzing existing research on chit-chat dialogue systems, we can outline the ongoing struggles and possible future trends, as listed below.

The Intrinsic Challenges of Chit-Chat. Although tremendous progress has been made in recent years in utilizing deep neural models to construct chit-chat systems, some intrinsic challenges of chit-chat are still not completely solved, even for state-of-the-art models. Among which, the most salient problems are:

- As stated by Huang *et al.*, 2020b, consistency is crucial for chatbots to gain long-term confidence and trust. With recent strong neural methods, there are still some deficiencies to respond consistently given the dialogue context and present consistent behaviors, which requires modeling long-range context information, long-term dialogue history, user profile, and the personas of the chatbots.
- One-to-diversity is a nagging problem of chit-chat conversation. As discussed in Section 4.3, researchers have tried several methods from different aspects (i.e., data manipulation, generation frameworks, training objectives, and leveraging extra resources) to mitigate the one-to-diversity problem. Owing to the lack of one-to-diversity data and the variability of chit-chat, one-to-diversity is still an open question.
- Another challenge for chit-chat systems is how to achieve a better understanding of the dialogues such as context, semantics, structure, and discourse. With a better understanding of the dialogues, the performance of dialogue systems will be enhanced.
- In addition, the evaluation of the chit-chat system is still an open challenge since chit-chat conversations are intricate and difficult to formulate. It is non-trivial to devote more efforts to design better evaluation methods, especially for generation-based systems.

More Sophisticated Requirements in Applications. With the fast evolution of conversational AI systems and the closer links between chit-chat dialogues and goal-oriented tasks, chatbots have to meet more sophisticated requirements in applications, which will pose new challenges for IR and NLP researchers.

- As demonstrated in Section 6, the line between chit-chat and goal-oriented tasks has become increasingly blurred. Chit-chat systems will be expected to detect user needs and complete these goals in real-time. In turn, goal-oriented tasks also need more pertinent skills of chit-chat to achieve the pre-set goals.
- Commonsense knowledge learning and utilizing of chit-chat systems are still in the preliminary stage. Further efforts are needed to build commonsense-aware chit-chat systems.
- With the development of mobile internet and applications, chit-chat systems require processing multi-modal information and dealing with heterogeneous data.
- Besides, with the wider application of conversational AI systems, it is non-trivial to pay more attention to the ethical issues and possible bias in human-machine conversations.

New Paradigm Based on Pre-Trained Language Models. Starting from BERT, pre-trained language models have changed the phase of NLP and IR fields. Various pre-trained models have been introduced in chit-chat systems, either as the backbone of generation models or providing context-aware vector representation for the context-response matching of retrieval-based systems, and achieved impressive improvements. Recent research further reveals that pre-trained language models are still underexplored, and thus tapping the potential of pre-trained language models for chit-chat systems is a valuable direction.

- One of the merits of large-size PLMs is their generalization ability, which can transfer self-trained knowledge to enhance chit-chat conversation modeling. Introducing PLMs can easily achieve domain adaptation and topic shifts.

- Another promising attribute of PLMs is their few-shot/zero-shot capabilities. For instance, GPT-3 does not even need to fine-tune model parameters to complete various tasks, which only needs to give a prompt or a few demonstration cases. Following this line, various prompt-based methods have been proposed recently for different classification tasks. It is predictable that prompt-based methods for chit-chat conversations will occur in the near future to better leverage the few-shot/zero-shot capabilities of PLMs.

- We also have to say that PLMs suffer from model efficiency and data efficiency to improve generalization ability and prevent over-fitting on small datasets. Thus, how to speed up PLMs and improve data efficiency beyond all doubt are crucial for chit-chat systems.

Acknowledgements

We are grateful to the editors and the anonymous reviewers for their constructive comments and suggestions. We also thank Zhangming Chan, Zhenxin Fu, Jiazhan Feng, Shen Gao, Chang Liu, Ran Le, Yue Wang, and Xueliang Zhao for their comments and suggestions in preparing the first version of this survey. This work was supported by the National Natural Science Foundation of China (NSFC Grant No. 62122089 & No. 61876196) and Beijing Outstanding Young Scientist Program (No. BJJWZYJH012019100020098). Juntao Li is the corresponding author.

References

Adiwardana, D., M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, *et al.* (2020). "Towards a human-like open-domain chatbot". *arXiv preprint arXiv:2001.09977*.

Akama, R., K. Inada, N. Inoue, S. Kobayashi, and K. Inui. (2017). "Generating stylistically consistent dialog responses with transfer learning". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 408–412.

Alamri, H., V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, *et al.* (2019). "Audio visual scene-aware dialog". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7558–7567.

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. (2015). "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

Bahdanau, D., K. Cho, and Y. Bengio. (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015*.

Baheti, A., A. Ritter, J. Li, and W. B. Dolan. (2018). "Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3970–3980.

Baheti, A., A. Ritter, and K. Small. (2020). “Fluent Response Generation for Conversational Question Answering”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 191–207.

Bak, J. and A. Oh. (2019). “Variational hierarchical user-based conversation model”. 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)*. 1941–1950.

Bak, J. and A. Oh. (2020). “Speaker Sensitive Response Evaluation Model”. *arXiv preprint arXiv:2006.07015*.

Banerjee, S. and A. Lavie. (2005). “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*. 65–72.

Bao, S., H. He, F. Wang, H. Wu, and H. Wang. (2020a). “PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 85–96.

Bao, S., H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Guo, Z. Liu, and X. Xu. (2020b). “Plato-2: Towards building an open-domain chatbot via curriculum learning”. *arXiv preprint arXiv:2006.16779*.

Bao, S., H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Wu, Z. Guo, H. Lu, X. Huang, *et al.* (2021). “Plato-xl: Exploring the large-scale pre-training of dialogue generation”. *arXiv preprint arXiv:2109.09519*.

Bradeško, L. and D. Mladenić. (2012). “A survey of chatbot systems through a loebner prize competition”. In: *Proceedings of Slovenian language technologies society eighth conference of language technologies*. Institut Jožef Stefan Ljubljana, Slovenia. 34–37.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* (2020). “Language models are few-shot learners”. *arXiv preprint arXiv:2005.14165*.

Cai, D., Y. Wang, W. Bi, Z. Tu, X. Liu, W. Lam, and S. Shi. (2019a). “Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory”. In: *NAACL*.

Cai, D., Y. Wang, W. Bi, Z. Tu, X. Liu, and S. Shi. (2019b). "Retrieval-guided Dialogue Response Generation via a Matching-to-Generation Framework". In: *EMNLP*.

Cai, H., H. Chen, Y. Song, Z. Ding, Y. Bao, W. Yan, and X. Zhao. (2020a). "Group-wise Contrastive Learning for Neural Dialogue Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 793–802.

Cai, H., H. Chen, Y. Song, C. Zhang, X. Zhao, and D. Yin. (2020b). "Data Manipulation: Towards Effective Instance Learning for Neural Dialogue Generation via Learning to Augment and Reweight". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6334–6343.

Cai, Y., M. Zuo, Q. Zhang, H. Xiong, and K. Li. (2020c). "A Bichannel Transformer with Context Encoding for Document-Driven Conversation Generation in Social Media". *Complexity*. 2020.

Campos, J. A., K. Cho, A. Otegi, A. Soroa, E. Agirre, and G. Azkune. (2020). "Improving Conversational Question Answering Systems after Deployment using Feedback-Weighted Learning". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2561–2571.

Cao, K. and S. Clark. (2017). "Latent Variable Dialogue Models and their Diversity". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 182–187.

Cao, Y., W. Bi, M. Fang, and D. Tao. (2020). "Pretrained Language Models for Dialogue Generation with Multiple Input Sources". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 909–917.

Chan, Z., J. Li, X. Yang, X. Chen, W. Hu, D. Zhao, and R. Yan. (2019). "Modeling personalization in continuous space for response generation via augmented wasserstein autoencoders". 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)*. 1931–1940.

Chan, Z., L. Liu, J. Li, H. Zhang, D. Zhao, S. Shi, and R. Yan. (2021). "Enhancing the open-domain dialogue evaluation in latent space". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4889–4900.

Chang, J., R. He, L. Wang, X. Zhao, T. Yang, and R. Wang. (2019). "A Semi-Supervised Stable Variational Network for Promoting Replier-Consistency in Dialogue Generation". 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)*. 1920–1930.

Charniak, E. *et al.* (2016). "Parsing as language modeling". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2331–2336.

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

Chen, H., Z. Ren, J. Tang, Y. E. Zhao, and D. Yin. (2018a). "Hierarchical variational memory network for dialogue generation". In: *Proceedings of the 2018 World Wide Web Conference*. 1653–1662.

Chen, Q., J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. (2019). "Towards knowledge-based recommender dialog system". *arXiv preprint arXiv:1908.05391*.

Chen, X., F. Meng, P. Li, F. Chen, S. Xu, B. Xu, and J. Zhou. (2020a). "Bridging the Gap between Prior and Posterior Knowledge Selection for Knowledge-Grounded Dialogue Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3426–3437.

Chen, Y., L. Wu, and M. J. Zaki. (2020b). "GraphFlow: Exploiting Conversation Flow with Graph Neural Networks for Conversational Machine Comprehension".

Chen, Y.-N., A. Celikyilmaz, and D. Hakkani-Tur. (2018b). "Deep learning for dialogue systems". In: *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*. 25–31.

Chu, E., P. Vijayaraghavan, and D. Roy. (2018). "Learning Personas from Dialogue with Attentive Memory Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2638–2646.

Clark, P., N. Balasubramanian, S. Bhakthavatsalam, K. Humphreys, J. Kinkead, A. Sabharwal, and O. Tafjord. (2014). "Automatic construction of inference-supporting knowledge bases". In: *4th Workshop on Automated Knowledge Base Construction (AKBC)*. Citeseer.

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

Colombo, P., W. Witon, A. Modi, J. Kennedy, and M. Kapadia. (2019). "Affect-Driven Dialog Generation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3734–3743.

Copple, K. L. (2008). *Bringing AI to life: putting today's tools and resources to work*. na.

Csáky, R., P. Purgai, and G. Recski. (2019). "Improving Neural Conversational Models with Entropy-Based Data Filtering". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5650–5669.

Cui, Z., Y. Li, J. Zhang, J. Cui, C. Wei, and B. Wang. (2020). "Focus-Constrained Attention Mechanism for CVAE-based Response Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2021–2030.

Daniel, J. and H. M. James. (2020). "Speech and Language Processing". *Online Draft*.

Das, A., S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. (2017a). "Visual dialog". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 326–335.

Das, A., S. Kottur, J. M. Moura, S. Lee, and D. Batra. (2017b). "Learning cooperative visual dialog agents with deep reinforcement learning". *arXiv preprint arXiv:1703.06585*.

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

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

Dinan, E., S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. (2018). “Wizard of wikipedia: Knowledge-powered conversational agents”. *arXiv preprint arXiv:1811.01241*.

Du, J., W. Li, Y. He, R. Xu, L. Bing, and X. Wang. (2018). “Variational autoregressive decoder for neural response generation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3154–3163.

Du, W. and A. W. Black. (2019). “Boosting dialog response generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 38–43.

Dziri, N., E. Kamalloo, K. Mathewson, and O. R. Zaiane. (2019). “Augmenting Neural Response Generation with Context-Aware Topical Attention”. In: *Proceedings of the First Workshop on NLP for Conversational AI*. 18–31.

Engonopoulos, N., C. Teichmann, and A. Koller. (2018). “Discovering User Groups for Natural Language Generation”. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 171–179.

Feng, S., X. Ren, H. Chen, B. Sun, K. Li, and X. Sun. (2020). “Regularizing Dialogue Generation by Imitating Implicit Scenarios”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6592–6604.

Forgues, G., J. Pineau, J.-M. Larchevêque, and R. Tremblay. (2014). “Bootstrapping dialog systems with word embeddings”. In: *Nips, workshop*. Vol. 2.

Galley, M., C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. (2015). "deltableu: A discriminative metric for generation tasks with intrinsically diverse targets". *arXiv preprint arXiv:1506.06863*.

Gan, Z., Y. Cheng, A. Kholy, L. Li, J. Liu, and J. Gao. (2019). "Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6463–6474.

Gao, J. (2017). "An introduction to deep learning for natural language processing". *International Summer School on Deep Learning, Bilbao*.

Gao, J., M. Galley, L. Li, *et al.* (2019a). "Neural Approaches to Conversational AI". *Foundations and Trends® in Information Retrieval*. 13(2-3): 127–298.

Gao, J., C. Xiong, P. Bennett, and N. Craswell. (2022). "Neural Approaches to Conversational Information Retrieval". *arXiv preprint arXiv:2201.05176*.

Gao, J., W. Bi, X. Liu, J. Li, and S. Shi. (2019b). "Generating multiple diverse responses for short-text conversation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 6383–6390.

Gao, J., W. Bi, X. Liu, J. Li, G. Zhou, and S. Shi. (2019c). "A Discrete CVAE for Response Generation on Short-Text Conversation". 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)*. 1898–1908.

Gao, X., S. Lee, Y. Zhang, C. Brockett, M. Galley, J. Gao, and W. B. Dolan. (2019d). "Jointly Optimizing Diversity and Relevance in Neural Response Generation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1229–1238.

Gao, X., Y. Zhang, S. Lee, M. Galley, C. Brockett, J. Gao, and B. Dolan. (2019e). "Structuring Latent Spaces for Stylized Response Generation". 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).* 1814–1823.

Gao, Y., C.-S. Wu, S. Joty, C. Xiong, R. Socher, I. King, M. Lyu, and S. C. Hoi. (2020). "Explicit memory tracker with coarse-to-fine reasoning for conversational machine reading". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 935–945.

Gehring, J., M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. (2017). "Convolutional Sequence to Sequence Learning". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org. 1243–1252.

Ghazarian, S., J. T.-Z. Wei, A. Galstyan, and N. Peng. (2019). "Better automatic evaluation of open-domain dialogue systems with contextualized embeddings". *arXiv preprint arXiv:1904.10635.*

Ghazvininejad, M., C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. (2018). "A knowledge-grounded neural conversation model". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. (2014). "Generative Adversarial Nets". In: *NIPS.*

Gopalakrishnan, K., B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tür, and A. A. AI. (2019). "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations." In: *INTERSPEECH.* 1891–1895.

Gu, J.-C., T. Li, Q. Liu, Z.-H. Ling, Z. Su, S. Wei, and X. Zhu. (2020a). "Speaker-aware bert for multi-turn response selection in retrieval-based chatbots". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* 2041–2044.

Gu, J.-C., Z.-H. Ling, X. Zhu, and Q. Liu. (2019a). "Dually Interactive Matching Network for Personalized Response Selection in Retrieval-Based Chatbots". 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)*. 1845–1854.

Gu, J.-C., Z. Ling, Q. Liu, Z. Chen, and X. Zhu. (2020b). "Filtering before Iteratively Referring for Knowledge-Grounded Response Selection in Retrieval-Based Chatbots". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1412–1422.

Gu, X., K. Cho, J.-W. Ha, and S. Kim. (2019b). "DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BkgBvsC9FQ.

Guo, D., C. Xu, and D. Tao. (2019a). "Image-question-answer synergistic network for visual dialog". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10434–10443.

Guo, D., H. Wang, and M. Wang. (2019b). "Dual Visual Attention Network for Visual Dialog." In: *IJCAI*. 4989–4995.

Guo, D., H. Wang, H. Zhang, Z.-J. Zha, and M. Wang. (2020). "Iterative context-aware graph inference for visual dialog". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10055–10064.

Guo, X., H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. S. Feris. (2018). "Dialog-based interactive image retrieval". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 676–686.

Gupta, P., S. Mehri, T. Zhao, A. Pavel, M. Eskenazi, and J. P. Bigham. (2019). "Investigating evaluation of open-domain dialogue systems with human generated multiple references". *arXiv preprint arXiv: 1907.10568*.

Gupta, S., B. P. S. Rawat, and H. Yu. (2020). "Conversational Machine Comprehension: a Literature Review". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2739–2753.

Guu, K., T. Hashimoto, Y. Oren, and P. Liang. (2018). "Generating Sentences by Editing Prototypes". *Transactions of the Association for Computational Linguistics*. 6: 437–450.

Ham, D., J.-G. Lee, Y. Jang, and K.-E. Kim. (2020). "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 583–592.

Han, X., Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, L. Zhang, W. Han, M. Huang, *et al.* (2021). "Pre-trained models: Past, present and future". *AI Open*.

He, K., X. Zhang, S. Ren, and J. Sun. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

He, T. and J. Glass. (2020). "Negative Training for Neural Dialogue Response Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2044–2058.

He, W., Y. Dai, Y. Zheng, Y. Wu, Z. Cao, D. Liu, P. Jiang, M. Yang, F. Huang, L. Si, *et al.* (2021). "GALAXY: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-Supervised Learning and Explicit Policy Injection". *arXiv preprint arXiv:2111.14592*.

Hendricks, L. A., S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. (2016). "Deep compositional captioning: Describing novel object categories without paired training data". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–10.

Hewitt, J. and C. D. Manning. (2019). "A structural probe for finding syntax in word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4129–4138.

Hixon, B., P. Clark, and H. Hajishirzi. (2015). "Learning knowledge graphs for question answering through conversational dialog". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 851–861.

Hosseini-Asl, E., B. McCann, C.-S. Wu, S. Yavuz, and R. Socher. (2020). "A simple language model for task-oriented dialogue". *arXiv preprint arXiv:2005.00796*.

Hu, B., Z. Lu, H. Li, and Q. Chen. (2014). "Convolutional neural network architectures for matching natural language sentences". In: *NIPS*.

Hu, W., Z. Chan, B. Liu, D. Zhao, J. Ma, and R. Yan. (2019). "GSN: A Graph-Structured Network for Multi-Party Dialogues".

Hu, Z., J. E. F. Tree, and M. Walker. (2018). "Modeling linguistic and personality adaptation for natural language generation". In: *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*. 20–31.

Hua, K., Z. Feng, C. Tao, R. Yan, and L. Zhang. (2020). "Learning to Detect Relevant Contexts and Knowledge for Response Selection in Retrieval-based Dialogue Systems". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 525–534.

Huang, C., O. R. Zaiane, A. Trabelsi, and N. Dziri. (2018). "Automatic dialogue generation with expressed emotions". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 49–54.

Huang, L., Z. Ye, J. Qin, L. Lin, and X. Liang. (2020a). "GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems". *arXiv preprint arXiv:2010.03994*.

Huang, M., X. Zhu, and J. Gao. (2020b). "Challenges in building intelligent open-domain dialog systems". *ACM Transactions on Information Systems (TOIS)*. 38(3): 1–32.

Hutchens, J. L. (1996). "How to pass the Turing test by cheating". *School of Electrical, Electronic and Computer Engineering research report TR97-05. Perth: University of Western Australia.*

Jawahar, G., B. Sagot, and D. Seddah. (2019). "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3651–3657.

Ji, Z., Z. Lu, and H. Li. (2014). “An information retrieval approach to short text conversation”. *arXiv preprint arXiv:1408.6988*.

Jiang, B., W. Zhou, J. Yang, C. Yang, S. Wang, and L. Pang. (2020). “PEDNet: A Persona Enhanced Dual Alternating Learning Network for Conversational Response Generation”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 4089–4099.

Jiang, S., P. Ren, C. Monz, and M. de Rijke. (2019). “Improving neural response diversity with frequency-aware cross-entropy loss”. In: *The World Wide Web Conference*. 2879–2885.

Jiang, S. and M. de Rijke. (2018). “Why are Sequence-to-Sequence Models So Dull? Understanding the Low-Diversity Problem of Chatbots”. In: *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. 81–86.

Jung, J., B. Son, and S. Lyu. (2020). “AttnIO: Knowledge Graph Exploration with In-and-Out Attention Flow for Knowledge-Grounded Dialogue”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3484–3497.

Kaiser, M., R. Saha Roy, and G. Weikum. (2020). “Conversational Question Answering over Passages by Leveraging Word Proximity Networks”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2129–2132.

Kang, G.-C., J. Lim, and B.-T. Zhang. (2019a). “Dual Attention Networks for Visual Reference Resolution in Visual Dialog”. 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)*. 2024–2033.

Kang, D., A. Balakrishnan, P. Shah, P. Crook, Y.-L. Boureau, and J. Weston. (2019b). “Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue”. *arXiv preprint arXiv:1909.03922*.

Khan, K., G. Sahu, V. Balasubramanian, L. Mou, and O. Vechtomova. (2020). “Adversarial Learning on the Latent Space for Diverse Dialog Generation”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 5026–5034.

Khayrallah, H. and J. Sedoc. (2020). "SMRT Chatbots: Improving Non-Task-Oriented Dialog with Simulated Multiple Reference Training". *arXiv preprint arXiv:2011.00547*.

Kim, H., H. Tan, and M. Bansal. (2020). "Modality-balanced models for visual dialogue". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 8091–8098.

Ko, W.-J., G. Durrett, and J. J. Li. (2019). "Linguistically-informed specificity and semantic plausibility for dialogue generation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3456–3466.

Ko, W.-J., A. Ray, Y. Shen, and H. Jin. (2020). "Generating Dialogue Responses from a Semantic Latent Space". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4339–4349.

Kong, J., Z. Zhong, Y. Cai, X. Wu, and D. Ren. (2020). "TSDG: Content-aware Neural Response Generation with Two-stage Decoding Process". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2121–2126.

Kongyoung, S., C. Macdonald, and I. Ounis. (2020). "Multi-Task Learning using Dynamic Task Weighting for Conversational Question Answering". In: *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*. 17–26.

Kottur, S., J. M. Moura, D. Parikh, D. Batra, and M. Rohrbach. (2018). "Visual coreference resolution in visual dialog using neural module networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 153–169.

Kottur, S., X. Wang, and V. Carvalho. (2017). "Exploring Personalized Neural Conversational Models." In: *IJCAI*. 3728–3734.

Kundu, S., Q. Lin, and H. T. Ng. (2020). "Learning to Identify Follow-Up Questions in Conversational Question Answering". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 959–968.

Laban, P., J. Canny, and M. A. Hearst. (2020). "What's The Latest? A Question-driven News Chatbot". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 380–387.

Lan, T., X.-L. Mao, W. Wei, X. Gao, and H. Huang. (2020). "Pone: A novel automatic evaluation metric for open-domain generative dialogue systems". *ACM Transactions on Information Systems (TOIS)*. 39(1): 1–37.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. (2019). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*.

Lei, W., X. He, M. de Rijke, and T.-S. Chua. (2020). "Conversational Recommendation: Formulation, Methods, and Evaluation". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2425–2428.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

Lewis, P., P. Stenetorp, and S. Riedel. (2021). "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1000–1008.

Li, J. and X. Sun. (2018). "A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 678–683.

Li, J., M. Galley, C. Brockett, J. Gao, and W. B. Dolan. (2016a). "A Diversity-Promoting Objective Function for Neural Conversation Models". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 110–119.

Li, J., M. Galley, C. Brockett, G. Spithourakis, J. Gao, and W. B. Dolan. (2016b). "A Persona-Based Neural Conversation Model". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 994–1003.

Li, J., W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. (2016c). "Deep Reinforcement Learning for Dialogue Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1192–1202.

Li, J., W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky. (2017a). "Adversarial Learning for Neural Dialogue Generation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2157–2169.

Li, J., Z. Zhang, H. Zhao, X. Zhou, and X. Zhou. (2020a). "Task-specific Objectives of Pre-trained Language Models for Dialogue Adaptation". *arXiv preprint arXiv:2009.04984*.

Li, J., C. Liu, C. Tao, Z. Chan, D. Zhao, M. Zhang, and R. Yan. (2021a). "Dialogue history matters! Personalized response selection in multi-turn retrieval-based chatbots". *ACM Transactions on Information Systems (TOIS)*. 39(4): 1–25.

Li, J., L. Qiu, B. Tang, D. Chen, D. Zhao, and R. Yan. (2019). "Insufficient data can also rock! learning to converse using smaller data with augmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 6698–6705.

Li, J. and R. Yan. (2018). "Overview of the NLPCC 2018 shared task: Multi-turn human-computer conversations". In: *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer. 446–451.

Li, L., C. Xu, W. Wu, Y. Zhao, X. Zhao, and C. Tao. (2020b). "Zero-Resource Knowledge-Grounded Dialogue Generation". *arXiv preprint arXiv:2008.12918*.

Li, M., S. Roller, I. Kulikov, S. Welleck, Y.-L. Boureau, K. Cho, and J. Weston. (2020c). "Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4715–4728.

Li, Q., H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen. (2020d). "EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 4454–4466.

Li, R., S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. (2018). "Towards deep conversational recommendations". *arXiv preprint arXiv:1812.07617*.

Li, S., S. Feng, D. Wang, K. Song, Y. Zhang, and W. Wang. (2020e). "EmoElicitor: An Open Domain Response Generation Model with User Emotional Reaction Awareness". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*. 3637–3643.

Li, Y., H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. (2017b). "Dailydialog: A manually labelled multi-turn dialogue dataset". *arXiv preprint arXiv:1710.03957*.

Li, Y., S. A. Hayati, W. Shi, and Z. Yu. (2021b). "DEUX: An Attribute-Guided Framework for Sociable Recommendation Dialog Systems". *arXiv preprint arXiv:2105.00825*.

Lian, R., M. Xie, F. Wang, J. Peng, and H. Wu. (2019). "Learning to Select Knowledge for Response Generation in Dialog Systems".

Liao, L., Y. Ma, X. He, R. Hong, and T.-s. Chua. (2018). "Knowledge-aware multimodal dialogue systems". In: *Proceedings of the 26th ACM international conference on Multimedia*. 801–809.

Liao, L., R. Takanobu, Y. Ma, X. Yang, M. Huang, and T.-S. Chua. (2019). "Deep conversational recommender in travel". *arXiv preprint arXiv:1907.00710*.

Lin, C.-Y. (2004). "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 74–81.

Lin, X., W. Jian, J. He, T. Wang, and W. Chu. (2020a). "Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 41–52.

Lin, Z., A. Madotto, G. I. Winata, and P. Fung. (2020b). "MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3391–3405.

Lintean, M. C. and V. Rus. (2012). "Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics." In: *Flairs conference*. 244–249.

Litman, D., S. Singh, M. S. Kearns, and M. Walker. (2000). "NJFun-a reinforcement learning spoken dialogue system". In: *ANLP-NAACL 2000 Workshop: Conversational Systems*.

Liu, C., K. Liu, S. He, Z. Nie, and J. Zhao. (2019a). "Incorporating Interlocutor-Aware Context into Response Generation on Multi-Party Chatbots". In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 718–727.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. (2021). "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing". *arXiv preprint arXiv:2107.13586*.

Liu, Q., Y. Chen, B. Chen, L. Jian-Guang, Z. Chen, B. Zhou, and D. Zhang. (2020). "You Impress Me: Dialogue Generation via Mutual Persona Perception". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1417–1427.

Liu, S., H. Chen, Z. Ren, Y. Feng, Q. Liu, and D. Yin. (2018a). "Knowledge diffusion for neural dialogue generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1489–1498.

Liu, Y., W. Bi, J. Gao, X. Liu, J. Yao, and S. Shi. (2018b). "Towards less generic responses in neural conversation models: A statistical re-weighting method". In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2769–2774.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. (2019b). "Roberta: A robustly optimized bert pretraining approach". *arXiv preprint arXiv:1907.11692*.

Liu, Z., Z.-Y. Niu, H. Wu, H. Wang, *et al.* (2019c). "Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs". 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).* 1782–1792.

Lowe, R., M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. (2017). "Towards an automatic turing test: Learning to evaluate dialogue responses". *arXiv preprint arXiv:1708.07149.*

Lowe, R., N. Pow, I. Serban, and J. Pineau. (2015). "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems". In: *SIGDIAL.* 285–294.

Lowe, R., I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. (2016). "On the evaluation of dialogue systems with next utterance classification". *arXiv preprint arXiv:1605.05414.*

Lu, J., A. Kannan, J. Yang, D. Parikh, and D. Batra. (2017). "Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 313–323.

Lu, J., X. Ren, Y. Ren, A. Liu, and Z. Xu. (2020). "Improving contextual language models for response retrieval in multi-turn conversation". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1805–1808.

Lu, Z. and H. Li. (2013). "A deep architecture for matching short texts". In: *NIPS.*

Luan, Y., C. Brockett, W. B. Dolan, J. Gao, and M. Galley. (2017). "Multi-Task Learning for Speaker-Role Adaptation in Neural Conversation Models". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 605–614.

Lubis, N., S. Sakti, K. Yoshino, and S. Nakamura. (2019). "Positive emotion elicitation in chat-based dialogue systems". *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 27(4): 866–877.

Ma, Z., R. Yang, B. Du, and Y. Chen. (2020). “A Control Unit for Emotional Conversation Generation”. *IEEE Access*. 8: 43168–43176.

Madotto, A., Z. Lin, G. I. Winata, and P. Fung. (2021). “Few-Shot Bot: Prompt-Based Learning for Dialogue Systems”. *arXiv preprint arXiv:2110.08118*.

Madotto, A., Z. Lin, C.-S. Wu, and P. Fung. (2019). “Personalizing dialogue agents via meta-learning”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5454–5459.

Majumder, B. P., H. Jhamtani, T. Berg-Kirkpatrick, and J. McAuley. (2020a). “Like Hiking? You Probably Enjoy Nature: Persona-grounded Dialog with Commonsense Expansions”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9194–9206.

Majumder, N., P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria. (2020b). “MIME: MIMicking Emotions for Empathetic Response Generation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8968–8979.

Mangrulkar, S., S. Shrivastava, V. Thenkanidiyoor, and D. A. Dinesh. (2018). “A context-aware convolutional natural language generation model for dialogue systems”. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 191–200.

Mazare, P.-E., S. Humeau, M. Raison, and A. Bordes. (2018). “Training Millions of Personalized Dialogue Agents”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2775–2779.

McCann, B., N. S. Keskar, C. Xiong, and R. Socher. (2018). “The natural language decathlon: Multitask learning as question answering”. *arXiv preprint arXiv:1806.08730*.

Mehri, S. and M. Eskenazi. (2020a). “Unsupervised evaluation of interactive dialog with dialogpt”. *arXiv preprint arXiv:2006.12719*.

Mehri, S. and M. Eskenazi. (2020b). “Usr: An unsupervised and reference free evaluation metric for dialog generation”. *arXiv preprint arXiv:2005.00456*.

Mi, F., W. Zhou, L. Kong, F. Cai, M. Huang, and B. Faltings. (2021). "Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1887–1898.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781*.

Misu, T., K. Georgila, A. Leuski, and D. Traum. (2012). "Reinforcement learning of question-answering dialogue policies for virtual museum guides". In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 84–93.

Mitchell, J. and M. Lapata. (2008). "Vector-based models of semantic composition". *NAACL-HLT*: 236–244.

Moghe, N., S. Arora, S. Banerjee, and M. M. Khapra. (2018). "Towards Exploiting Background Knowledge for Building Conversation Systems". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2322–2332.

Moon, S., P. Shah, A. Kumar, and R. Subba. (2019). "Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 845–854.

Mrkšić, N., D. Ó. Séaghdha, T.-H. Wen, B. Thomson, and S. Young. (2017). "Neural Belief Tracker: Data-Driven Dialogue State Tracking". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1777–1788.

Murahari, V., D. Batra, D. Parikh, and A. Das. (2020). "Large-scale pretraining for visual dialog: A simple state-of-the-art baseline". In: *European Conference on Computer Vision*. Springer. 336–352.

Nayak, N., D. Hakkani-Tür, M. A. Walker, and L. P. Heck. (2017). "To Plan or not to Plan? Discourse Planning in Slot-Value Informed Sequence to Sequence Models for Language Generation." In: *INTERSPEECH*. 3339–3343.

Niu, Y., H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen. (2019). "Recursive visual attention in visual dialog". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6679–6688.

Noh, H., P. H. Seo, and B. Han. (2016). "Image question answering using convolutional neural network with dynamic parameter prediction". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 30–38.

Ohsugi, Y., I. Saito, K. Nishida, H. Asano, and J. Tomita. (2019). "A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension". In: *Proceedings of the First Workshop on NLP for Conversational AI*. 11–17.

Olabiyi, O., A. Khazane, A. Salimov, and E. Mueller. (2019). "An Adversarial Learning Framework For A Persona-Based Multi-Turn Dialogue Model". In: *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. 1–10.

Pan, B., H. Li, Z. Yao, D. Cai, and H. Sun. (2019). "Reinforced Dynamic Reasoning for Conversational Question Generation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2114–2124.

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

Park, S., T. Whang, Y. Yoon, and H. Lim. (2020). "Multi-View Attention Networks for Visual Dialog". *arXiv preprint arXiv:2004.14025*.

Park, Y., J. Cho, and G. Kim. (2018). "A Hierarchical Latent Structure for Variational Conversation Modeling". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1792–1801.

Parthasarathi, P. and J. Pineau. (2018). "Extending Neural Generative Conversational Model using External Knowledge Sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 690–695.

Peng, B., C. Li, J. Li, S. Shayandeh, L. Liden, and J. Gao. (2020). "Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model". *arXiv preprint arXiv:2005.05298*.

Penha, G. and C. Hauff. (2020). “What does BERT know about books, movies and music? Probing BERT for Conversational Recommendation”. In: *Fourteenth ACM Conference on Recommender Systems*. 388–397.

Pennington, J., R. Socher, and C. D. Manning. (2014). “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. (2018). “Deep contextualized word representations”. *arXiv preprint arXiv:1802.05365*.

Petroni, F., T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. (2019). “Language Models as Knowledge Bases?” 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)*. 2463–2473.

Qi, P., Y. Zhang, and C. D. Manning. (2020). “Stay Hungry, Stay Focused: Generating Informative and Specific Questions in Information-Seeking Conversations”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 25–40.

Qi, W., Y. Gong, Y. Yan, C. Xu, B. Yao, B. Zhou, B. Cheng, D. Jiang, J. Chen, R. Zhang, *et al.* (2021). “ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation”. *arXiv preprint arXiv:2104.08006*.

Qian, Q., M. Huang, H. Zhao, J. Xu, and X. Zhu. (2018). “Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation.” In: *IJCAI*. 4279–4285.

Qin, L., M. Galley, C. Brockett, X. Liu, X. Gao, W. B. Dolan, Y. Choi, and J. Gao. (2019). “Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5427–5436.

Qiu, L., J. Li, W. Bi, D. Zhao, and R. Yan. (2019). “Are training samples correlated? learning to generate dialogue responses with multiple references”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3826–3835.

Qiu, L., Y. Shiu, P. Lin, R. Song, Y. Liu, D. Zhao, and R. Yan. (2020). "What If Bots Feel Moods?" In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1161–1170.

Qu, C., L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer. (2020). "Open-retrieval conversational question answering". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 539–548.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever. (2018a). "Improving Language Understanding by Generative Pre-Training". *Open AI, blog*.

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

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

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *Journal of Machine Learning Research*. 21(140): 1–67.

Rashkin, H., E. M. Smith, M. Li, and Y.-L. Boureau. (2018). "Towards empathetic open-domain conversation models: A new benchmark and dataset". *arXiv preprint arXiv:1811.00207*.

Rashkin, H., E. M. Smith, M. Li, and Y.-L. Boureau. (2019). "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5370–5381.

Reddy, S., D. Chen, and C. D. Manning. (2019). "Coqa: A conversational question answering challenge". *Transactions of the Association for Computational Linguistics*. 7: 249–266.

Ren, S., K. He, R. B. Girshick, and J. Sun. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *NIPS*.

Ritter, A., C. Cherry, and W. B. Dolan. (2011). "Data-driven response generation in social media". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 583–593.

Roberts, A., C. Raffel, and N. Shazeer. (2020). “How Much Knowledge Can You Pack into the Parameters of a Language Model?” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5418–5426.

Roller, S., E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, *et al.* (2021). “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 300–325.

Saeidi, M., M. Bartolo, P. Lewis, S. Singh, T. Rocktäschel, M. Sheldon, G. Bouchard, and S. Riedel. (2018). “Interpretation of Natural Language Rules in Conversational Machine Reading”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2087–2097.

Sai, A. B., A. K. Mohankumar, S. Arora, and M. M. Khapra. (2020). “Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining”. *Transactions of the Association for Computational Linguistics*. 8: 810–827.

Sanh, V., A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, *et al.* (2021). “Multitask prompted training enables zero-shot task generalization”. *arXiv preprint arXiv:2110.08207*.

Sankar, C., S. Subramanian, C. Pal, S. Chandar, and Y. Bengio. (2019). “Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 32–37.

Sarkar, R., K. Goswami, M. Arcan, and J. P. McCrae. (2020). “Suggest me a movie for tonight: Leveraging Knowledge Graphs for Conversational Recommendation”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 4179–4189.

Schatzmann, J., K. Weilhammer, M. Stuttle, and S. Young. (2006). “A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies”. *The knowledge engineering review*. 21(2): 97–126.

Schick, T. and H. Schütze. (2021a). "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 255–269.

Schick, T. and H. Schütze. (2021b). "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2339–2352.

Schwartz, I., S. Yu, T. Hazan, and A. G. Schwing. (2019). "Factor graph attention". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2039–2048.

Scialom, T., S. S. Tekiroğlu, J. Staiano, and M. Guerini. (2020). "Toward Stance-based Personas for Opinionated Dialogues". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2625–2635.

See, A., S. Roller, D. Kiela, and J. Weston. (2019). "What makes a good conversation? How controllable attributes affect human judgments". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1702–1723.

Sellam, T., D. Das, and A. P. Parikh. (2020). "BLEURT: Learning robust metrics for text generation". *ACL*.

Seo, P. H., A. M. Lehrmann, B. Han, and L. Sigal. (2017). "Visual Reference Resolution using Attention Memory for Visual Dialog". In: *NIPS*.

Serban, I. V., R. Lowe, P. Henderson, L. Charlin, and J. Pineau. (2018). "A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version". *Dialogue & Discourse*. 9(1): 1–49.

Shah, M., S. Mehri, and T. Srinivasan. (2020). "Reasoning Over History: Context Aware Visual Dialog". In: *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*. 75–83.

Shah, P., D. Hakkani-Tur, B. Liu, and G. Tur. (2018). "Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. 41–51.

Shang, L., Z. Lu, and H. Li. (2015). "Neural responding machine for short-text conversation". *arXiv preprint arXiv:1503.02364*.

Shawar, B. A. and E. S. Atwell. (2005). "Using corpora in machine-learning chatbot systems". *International journal of corpus linguistics*. 10(4): 489–516.

Shen, L. and Y. Feng. (2020). "CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 556–566.

Shen, L., Y. Feng, and H. Zhan. (2019a). "Modeling Semantic Relationship in Multi-turn Conversations with Hierarchical Latent Variables". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5497–5502.

Shen, T., X. Geng, Q. Tao, D. Guo, D. Tang, N. Duan, G. Long, and D. Jiang. (2019b). "Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base". 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)*. 2442–2451.

Shen, X., H. Su, W. Li, and D. Klakow. (2018). "Nexus network: Connecting the preceding and the following in dialogue generation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4316–4327.

Shen, X., H. Su, Y. Li, W. Li, S. Niu, Y. Zhao, A. Aizawa, and G. Long. (2017). "A Conditional Variational Framework for Dialog Generation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 504–509.

Shi, W. and Z. Yu. (2018). "Sentiment Adaptive End-to-End Dialog Systems". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 1509–1519.

Shum, M., S. Zheng, W. Kryscinski, C. Xiong, and R. Socher. (2020). "Sketch-Fill-AR: A Persona-Grounded Chit-Chat Generation Framework". In: *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI.* 118–131.

Simonyan, K. and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition". *arXiv preprint arXiv:1409.1556.*

Song, H., Y. Wang, W. Zhang, X. Liu, and T. Liu. (2020a). "Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 5821–5831.

Song, H., W.-N. Zhang, Y. Cui, D. Wang, and T. Liu. (2019a). "Exploiting Persona Information for Diverse Generation of Conversational Responses".

Song, H., W.-N. Zhang, J. Hu, and T. Liu. (2020b). "Generating persona consistent dialogues by exploiting natural language inference". In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 34. No. 05. 8878–8885.

Song, Y., C.-T. Li, J.-Y. Nie, M. Zhang, D. Zhao, and R. Yan. (2018). "An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems". In: *IJCAI.*

Song, Y., Z. Tian, D. Zhao, M. Zhang, and R. Yan. (2017). "Diversifying neural conversation model with maximal marginal relevance". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers).* 169–174.

Song, Z., X. Zheng, L. Liu, M. Xu, and X.-J. Huang. (2019b). "Generating responses with a specific emotion in dialog". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 3685–3695.

Sordoni, A., M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and W. B. Dolan. (2015). "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 196–205.

Su, H., X. Shen, S. Zhao, Z. Xiao, P. Hu, C. Niu, J. Zhou, *et al.* (2020a). "Diversifying Dialogue Generation with Non-Conversational Text". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7087–7097.

Su, Y., L. Shu, E. Mansimov, A. Gupta, D. Cai, Y.-A. Lai, and Y. Zhang. (2021). "Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System". *arXiv preprint arXiv:2109.14739*.

Su, Y., Y. Wang, S. Baker, D. Cai, X. Liu, A. Korhonen, and N. Collier. (2020b). "Prototype-to-Style: Dialogue Generation with Style-Aware Editing on Retrieval Memory". *ArXiv*. abs/2004.02214.

Suganuma, M., T. Okatani, *et al.* (2020). "Efficient Attention Mechanism for Visual Dialog that Can Handle All the Interactions Between Multiple Inputs". In: *16th European Conference on Computer Vision, ECCV 2020*. Springer Science and Business Media Deutschland GmbH. 223–240.

Sun, K., D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie. (2019). "Dream: A challenge data set and models for dialogue-based reading comprehension". *Transactions of the Association for Computational Linguistics*. 7: 217–231.

Sun, T., X. Liu, X. Qiu, and X. Huang. (2021). "Paradigm Shift in Natural Language Processing". *arXiv preprint arXiv:2109.12575*.

Sutskever, I., O. Vinyals, and Q. V. Le. (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing systems*. 3104–3112.

Takayama, J. and Y. Arase. (2020). "Consistent Response Generation with Controlled Specificity". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 4418–4427.

Tang, J., T. Zhao, C. Xiong, X. Liang, E. Xing, and Z. Hu. (2019). "Target-Guided Open-Domain Conversation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5624–5634.

Tao, C., J. Feng, C. Liu, J. Li, X. Geng, and D. Jiang. (2021a). "Building an Efficient and Effective Retrieval-based Dialogue System via Mutual Learning". *arXiv preprint arXiv:2110.00159*.

Tao, C., J. Feng, R. Yan, W. Wu, and D. Jiang. (2021b). "A survey on response selection for retrieval-based dialogues". In: *IJCAI*.

Tao, C., L. Mou, D. Zhao, and R. Yan. (2018). "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Tao, C., W. Wu, Y. Feng, D. Zhao, and R. Yan. (2020). "Improving Matching Models with Hierarchical Contextualized Representations for Multi-turn Response Selection". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1865–1868.

Tao, C., W. Wu, C. Xu, W. Hu, D. Zhao, and R. Yan. (2019a). "Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots". In: *WSDM*.

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

Tay, Y., A. T. Luu, and S. C. Hui. (2018a). "Hermitian Co-Attention Networks for Text Matching in Asymmetrical Domains". In: *IJCAI*. 4425–4431.

Tay, Y., L. A. Tuan, and S. C. Hui. (2018b). "Multi-cast Attention Networks". In: *SIGKDD*.

Thorat, S. A. and V. Jadhav. (2020). "A review on implementation issues of rule-based chatbot systems". In: *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

Tian, Z., W. Bi, D. Lee, L. Xue, Y. SONG, X. Liu, and N. L. Zhang. (2020). “Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 650–659.

Tian, Z., W. Bi, X. Li, and N. L. Zhang. (2019). “Learning to abstract for memory-augmented conversational response generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 3816–3825.

Tian, Z., R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao. (2017). “How to make context more useful? an empirical study on context-aware neural conversational models”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 231–236.

Tong, X., Z. Fu, M. Shang, D. Zhao, and R. Yan. (2018). “One" Ruler" for all languages: multi-lingual dialogue evaluation with adversarial multi-task learning”. *arXiv preprint arXiv:1805.02914.*

Tuan, Y.-L., Y.-N. Chen, and H.-y. Lee. (2019). “DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs”. *arXiv preprint arXiv:1910.00610.*

Tuan, Y.-L., W. Wei, and W. Y. Wang. (2020). “Knowledge Injection into Dialogue Generation via Language Models”. *arXiv preprint arXiv:2004.14614.*

Ueyama, A. and Y. Kano. (2020). “Diverse dialogue generation with context dependent dynamic loss function”. In: *Proceedings of the 28th International Conference on Computational Linguistics.* 4123–4127.

Vakulenko, S., S. Longpre, Z. Tu, and R. Anantha. (2020). “A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering”. In: *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI).* 7–16.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017). “Attention is all you need”. In: *Advances in neural information processing systems.* 5998–6008.

Verma, N., A. Sharma, D. Madan, D. Contractor, H. Kumar, and S. Joshi. (2020). "Neural Conversational QA: Learning to Reason vs Exploiting Patterns". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7263–7269.

Vinyals, O. and Q. Le. (2015). "A neural conversational model". *arXiv preprint arXiv:1506.05869*.

Vougiouklis, P., J. Hare, and E. Simperl. (2016). "A Neural Network Approach for Knowledge-Driven Response Generation". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 3370–3380.

Wallace, R. (2003). "The elements of AIML style". *Alice AI Foundation*. 139.

Wang, C., P. Liu, and Y. Zhang. (2021). "Can Generative Pre-trained Language Models Serve As Knowledge Bases for Closed-book QA?" In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics. 3241–3251. DOI: 10.18653/v1/2021.acl-long.251.

Wang, H., Z. Wu, and J. Chen. (2019a). "Multi-turn response selection in retrieval-based chatbots with iterated attentive convolution matching network". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1081–1090.

Wang, W., S. Feng, D. Wang, and Y. Zhang. (2019b). "Answer-guided and Semantic Coherent Question Generation in Open-domain Conversation". 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)*. 5069–5079.

Wang, X., W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. (2019c). "Persuasion for good: Towards a personalized persuasive dialogue system for social good". *arXiv preprint arXiv:1906.06725*.

Wang, Y., C. Liu, M. Huang, and L. Nie. (2018). "Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2193–2203.

Wang, Y., P. Ke, Y. Zheng, K. Huang, Y. Jiang, X. Zhu, and M. Huang. (2020a). "A large-scale chinese short-text conversation dataset". In: *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer. 91–103.

Wang, Y., S. Joty, M. Lyu, I. King, C. Xiong, and S. C. Hoi. (2020b). "VD-BERT: A Unified Vision and Dialog Transformer with BERT". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3325–3338.

Wei, J., M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. (2021). "Finetuned language models are zero-shot learners". *arXiv preprint arXiv:2109.01652*.

Weintraub, J. (1986). "History of the PC Therapist". URL: http://www.loebner.net/Prizef/weintraub-bio.html.

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

Whang, T., D. Lee, C. Lee, K. Yang, D. Oh, and H. Lim. (2019). "An effective domain adaptive post-training method for bert in response selection". *arXiv preprint arXiv:1908.04812*.

Whang, T., D. Lee, D. Oh, C. Lee, K. Han, D.-h. Lee, and S. Lee. (2020). "Do Response Selection Models Really Know What's Next? Utterance Manipulation Strategies for Multi-turn Response Selection". *arXiv preprint arXiv:2009.04703*.

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

Williams, J. D. (2007). "Partially observable Markov decision processes for spoken dialogue management". *PhD thesis*. University of Cambridge.

Wolf, T., V. Sanh, J. Chaumond, and C. Delangue. (2019). "Transfertransfo: A transfer learning approach for neural network based conversational agents". *arXiv preprint arXiv:1901.08149*.

Wu, B., M. Li, Z. Wang, Y. Chen, D. F. Wong, J. Huang, B. Wang, *et al.* (2020a). "Guiding Variational Response Generator to Exploit Persona". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 53–65.

Wu, C.-S., S. C. Hoi, R. Socher, and C. Xiong. (2020b). "TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 917–929.

Wu, S., Y. Li, D. Zhang, and Z. Wu. (2020c). "Improving Knowledge-Aware Dialogue Response Generation by Using Human-Written Prototype Dialogues". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1402–1411.

Wu, S., Y. Li, D. Zhang, Y. Zhou, and Z. Wu. (2020d). "Diverse and informative dialogue generation with context-specific commonsense knowledge awareness". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 5811–5820.

Wu, S., Y. Li, D. Zhang, Y. Zhou, and Z. Wu. (2020e). "Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*. 3766–3772.

Wu, W. and R. Yan. (2019a). "Deep Chit-Chat: Deep Learning for Chatbots". In: *Companion Proceedings of The 2019 World Wide Web Conference*. 1329–1329.

Wu, W. and R. Yan. (2019b). "Deep chit-chat: Deep learning for chatbots". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1413–1414.

Wu, W., Z. Guo, X. Zhou, H. Wu, X. Zhang, R. Lian, and H. Wang. (2019a). "Proactive human-machine conversation with explicit conversation goals". *arXiv preprint arXiv:1906.05572*.

Wu, Y., F. Wei, S. Huang, Y. Wang, Z. Li, and M. Zhou. (2019b). "Response generation by context-aware prototype editing". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 7281–7288.

Wu, Y., W. Wu, C. Xing, M. Zhou, and Z. Li. (2017). "Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots". In: *ACL*. 496–505.

Wu, Y., W. Wu, C. Xu, and Z. Li. (2018). "Knowledge enhanced hybrid neural network for text matching". In: *AAAI*.

Xing, C., Y. Wu, W. Wu, Y. Huang, and M. Zhou. (2018). "Hierarchical recurrent attention network for response generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Xu, H., S. Moon, H. Liu, B. Liu, P. Shah, and S. Y. Philip. (2020a). "User Memory Reasoning for Conversational Recommendation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 5288–5308.

Xu, J., H. Wang, Z. Niu, H. Wu, and W. Che. (2020b). "Knowledge graph grounded goal planning for open-domain conversation generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 9338–9345.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. (2015). "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. PMLR. 2048–2057.

Xu, R., C. Tao, D. Jiang, X. Zhao, D. Zhao, and R. Yan. (2020c). "Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues". *arXiv preprint arXiv:2009.06265*.

Xu, X., O. Dušek, I. Konstas, and V. Rieser. (2018a). "Better Conversations by Modeling, Filtering, and Optimizing for Coherence and Diversity". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3981–3991.

Xu, Y., E. Ishii, S. Cahyawijaya, Z. Liu, G. I. Winata, A. Madotto, D. Su, and P. Fung. (2021). "Retrieval-free knowledge-grounded dialogue response generation with adapters". *arXiv preprint arXiv:2105.06232*.

Xu, Z., N. Jiang, B. Liu, W. Rong, B. Wu, B. Wang, Z. Wang, and X. Wang. (2018b). "LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2070–2080.

Xu, Z., B. Liu, B. Wang, C.-J. Sun, X. Wang, Z. Wang, and C. Qi. (2017). "Neural response generation via gan with an approximate embedding layer". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 617–626.

Yan, R. (2018). ""Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI". In: *IJCAI*.

Yan, R., Y. Song, and H. Wu. (2016). "Learning to respond with deep neural networks for retrieval-based human-computer conversation system". In: *SIGIR*.

Yan, R. and W. Wu. (2021). "Empowering Conversational AI is a Trip to Mars: Progress and Future of Open Domain Human-Computer Dialogues". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 17. 15078–15086.

Yang, B., C. Han, Y. Li, L. Zuo, and Z. Yu. (2021a). "Improving Conversational Recommendation Systems' Quality with Context-Aware Item Meta Information". *arXiv preprint arXiv:2112.08140*.

Yang, L., J. Hu, M. Qiu, C. Qu, J. Gao, W. Croft, X. Liu, Y. Shen, and J. Liu. (2019a). "A Hybrid Retrieval-Generation Neural Conversation Model". *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Yang, P., X. Sun, W. Li, S. Ma, W. Wu, and H. Wang. (2018). "SGM: Sequence Generation Model for Multi-label Classification". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 3915–3926.

Yang, T., Z.-J. Zha, and H. Zhang. (2019b). "Making history matter: History-advantage sequence training for visual dialog". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2561–2569.

Yang, Y., Y. Li, and X. Quan. (2021b). "UBAR: Towards Fully End-to-End Task-Oriented Dialog System with GPT-2". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 16. 14230–14238.

Yang, Z., W. Wu, C. Xu, X. Liang, J. Bai, L. Wang, W. Wang, and Z. Li. (2020). "StyleDGPT: Stylized Response Generation with Pre-trained Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1548–1559.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. (2019c). "Xlnet: Generalized autoregressive pretraining for language understanding". *arXiv preprint arXiv:1906.08237*.

Yang, Z., X. He, J. Gao, L. Deng, and A. Smola. (2016). "Stacked attention networks for image question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

Yenicelik, D., F. Schmidt, and Y. Kilcher. (2020). "How does BERT capture semantics? A closer look at polysemous words". In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. 156–162.

Yih, W.-t., X. He, and J. Gao. (2015). "Deep learning and continuous representations for natural language processing". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. 6–8.

Yih, W.-t., X. He, and J. Gao. (2016). "Deep learning and continuous representations for natural language processing". In: *IJCAI*.

Young, S., M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. (2010). "The hidden information state model: A practical framework for POMDP-based spoken dialogue management". *Computer Speech & Language*. 24(2): 150–174.

Yu, H., A. Li, R. Jiang, Y. Jia, X. Zhao, and W. Han. (2019a). "HDGS: A Hybrid Dialogue Generation System using Adversarial Learning". *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*: 135–141.

Yu, T., Y. Shen, R. Zhang, X. Zeng, and H. Jin. (2019b). "Vision-language recommendation via attribute augmented multimodal reinforcement learning". In: *Proceedings of the 27th ACM International Conference on Multimedia*. 39–47.

Yuan, C., W. Zhou, M. Li, S. Lv, F. Zhu, J. Han, and S. Hu. (2019). "Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots". In: *EMNLP*.

Yuma, T., N. Yoshinaga, and M. Toyoda. (2020). "uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 199–206.

Zamani, H., J. R. Trippas, J. Dalton, and F. Radlinski. (2022). "Conversational Information Seeking". *arXiv preprint arXiv:2201.08808*.

Zeng, M., Y. Wang, and Y. Luo. (2019). "Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation". 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)*. 1267–1272.

Zeng, X., D. Zeng, S. He, K. Liu, and J. Zhao. (2018). "Extracting relational facts by an end-to-end neural model with copy mechanism". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 506–514.

Zhang, B. and X. Zhang. (2019). "Hierarchy response learning for neural conversation generation". 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)*. 1772–1781.

Zhang, H., Y. Lan, L. Pang, J. Guo, and X. Cheng. (2019a). "ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-turn Dialogue Generation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3721–3730.

Zhang, H., Z. Liu, C. Xiong, and Z. Liu. (2020a). "Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2031–2043.

Zhang, J., C. Tao, Z. Xu, Q. Xie, W. Chen, and R. Yan. (2019b). "EnsembleGAN: Adversarial Learning for Retrieval-Generation Ensemble Model on Short-Text Conversation". In: *SIGIR*.

Zhang, J., Y. Yang, C. Chen, L. He, and Z. Yu. (2021). "KERS: A Knowledge-Enhanced Framework for Recommendation Dialog Systems with Multiple Subgoals". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 1092–1101.

Zhang, L., Y. Yang, J. Zhou, C. Chen, and L. He. (2020b). "Retrieval-Polished Response Generation for Chatbot". *IEEE Access*. 8: 123882–123890.

Zhang, M. and J. Li. (2021). "A commentary of GPT-3 in MIT Technology Review 2021". *Fundamental Research*. 1(6): 831–833.

Zhang, R., Y. Zheng, J. Shao, X. Mao, Y. Xi, and M. Huang. (2020c). "Dialogue Distillation: Open-domain Dialogue Augmentation Using Unpaired Data". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3449–3460.

Zhang, R., T. Yu, Y. Shen, H. Jin, C. Chen, and L. Carin. (2020d). "Reward Constrained Interactive Recommendation with Natural Language Feedback". *arXiv preprint arXiv:2005.01618*.

Zhang, R., J. Guo, Y. Fan, Y. Lan, J. Xu, and X. Cheng. (2018a). "Learning to control the specificity in neural response generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1108–1117.

Zhang, S., E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. (2018b). "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2204–2213.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. (2019c). "Bertscore: Evaluating text generation with bert". *ICLR*.

Zhang, W., Y. Cui, Y. Wang, Q. Zhu, L. Li, L. Zhou, and T. Liu. (2018c). "Context-sensitive generation of open-domain conversational responses". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2437–2447.

Zhang, W., K. Song, Y. Kang, Z. Wang, C. Sun, X. Liu, S. Li, M. Zhang, and L. Si. (2020e). "Multi-Turn Dialogue Generation in E-Commerce Platform with the Context of Historical Dialogue". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1981–1990.

Zhang, X. (2019). "MC^2: Multi-perspective convolutional cube for conversational machine reading comprehension". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6185–6190.

Zhang, Y., S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan. (2020f). "DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.

Zhang, Y., X. Chen, Q. Ai, L. Yang, and W. B. Croft. (2018d). "Towards conversational search and recommendation: System ask, user respond". In: *Proceedings of the 27th acm international conference on information and knowledge management*. 177–186.

Zhang, Z., J. Li, P. Zhu, H. Zhao, and G. Liu. (2018e). "Modeling Multi-turn Conversation with Deep Utterance Aggregation". In: *COLING*. 3740–3752.

Zhao, T., R. Zhao, and M. Eskenazi. (2017). "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 654–664.

Zhao, W., M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger. (2019a). "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance". *arXiv preprint arXiv:1909.02622*.

Zhao, X., B. He, Y. Wang, Y. Li, F. Mi, Y. Liu, X. Jiang, Q. Liu, and H. Chen. (2021). "UniDS: A Unified Dialogue System for Chit-Chat and Task-oriented Dialogues". *arXiv preprint arXiv:2110.08032*.

Zhao, X., C. Tao, W. Wu, C. Xu, D. Zhao, and R. Yan. (2019b). "A document-grounded matching network for response selection in retrieval-based chatbots". In: *IJCAI*.

Zhao, X., W. Wu, C. Tao, C. Xu, D. Zhao, and R. Yan. (2019c). "Low-Resource Knowledge-Grounded Dialogue Generation". In: *International Conference on Learning Representations*.

Zhao, X., W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan. (2020a). "Knowledge-Grounded Dialogue Generation with Pre-trained Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3377–3390.

Zhao, Y., C. Xu, and W. Wu. (2020b). "Learning a Simple and Effective Model for Multi-turn Response Generation with Auxiliary Tasks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3472–3483.

Zheng, C., Y. Cao, D. Jiang, and M. Huang. (2020a). "Difference-aware Knowledge Selection for Knowledge-grounded Conversation Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 115–125.

Zheng, Y., G. Chen, M. Huang, S. Liu, and X. Zhu. (2019a). "Personalized dialogue generation with diversified traits". *arXiv preprint arXiv:1901.09672*.

Zheng, Y., R. Zhang, M. Huang, and X. Mao. (2020b). "A pre-training based personalized dialogue generation model with persona-sparse data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 9693–9700.

Zheng, Z., W. Wang, S. Qi, and S.-C. Zhu. (2019b). "Reasoning visual dialogs with structural and partial observations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6669–6678.

Zhong, P., C. Zhang, H. Wang, Y. Liu, and C. Miao. (2020). "Towards Persona-Based Empathetic Conversational Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6556–6566.

Zhong, V. and L. Zettlemoyer. (2019). "E3: Entailment-driven Extracting and Editing for Conversational Machine Reading". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2310–2320.

Zhou, H., M. Huang, T. Zhang, X. Zhu, and B. Liu. (2018a). "Emotional chatting machine: Emotional conversation generation with internal and external memory". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Zhou, H., P. Ke, Z. Zhang, Y. Gu, Y. Zheng, C. Zheng, Y. Wang, C. H. Wu, H. Sun, X. Yang, *et al.* (2021a). "EVA: An open-domain chinese dialogue system with large-scale generative pre-training". *arXiv preprint arXiv:2108.01547*.

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

Zhou, K., S. Prabhumoye, and A. W. Black. (2018c). "A Dataset for Document Grounded Conversations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 708–713.

Zhou, K., K. Zhang, Y. Wu, S. Liu, and J. Yu. (2019). "Unsupervised Context Rewriting for Open Domain Conversation". 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)*. 1834–1844.

Zhou, K., W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, and J. Yu. (2020a). "Improving conversational recommender systems via knowledge graph based semantic fusion". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1006–1014.

Zhou, K., Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen. (2020b). "Towards Topic-Guided Conversational Recommender System". *arXiv preprint arXiv:2010.04125*.

Zhou, L., J. Gao, D. Li, and H.-Y. Shum. (2020c). "The design and implementation of xiaoice, an empathetic social chatbot". *Computational Linguistics*. 46(1): 53–93.

Zhou, P., K. Gopalakrishnan, B. Hedayatnia, S. Kim, J. Pujara, X. Ren, Y. Liu, and D. Hakkani-Tur. (2021b). "Think Before You Speak: Using Self-talk to Generate Implicit Commonsense Knowledge for Response Generation". *arXiv preprint arXiv:2110.08501*.

Zhou, X. and W. Y. Wang. (2018). “MojiTalk: Generating Emotional Responses at Scale”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1128–1137.

Zhou, X., D. Dong, H. Wu, S. Zhao, D. Yu, H. Tian, X. Liu, and R. Yan. (2016). “Multi-view response selection for human-computer conversation”. In: *EMNLP*. 372–381.

Zhou, X., L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu. (2018d). “Multi-turn response selection for chatbots with deep attention matching network”. In: *ACL*.

Zhu, Q., L. Cui, W.-N. Zhang, F. Wei, and T. Liu. (2019). “Retrieval-Enhanced Adversarial Training for Neural Response Generation”. In: *ACL*.

Zhuang, Y., X. Wang, H. Zhang, J. Xie, and X. Zhu. (2017). “An Ensemble Approach to Conversation Generation”. In: *NLPCC*.

Zou, M., X. Li, H. Liu, and Z.-H. Deng. (2018). “Memd: A diversity-promoting learning framework for short-text conversation”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 1281–1291.