Empowering Conversational AI is a Trip to Mars: Progress and Future of Open Domain Human-Computer Dialogues

Rui Yan [1,2] **and Wei Wu** [3]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Wangxuan Institute of Computer Technology, Peking University
[3] Microsoft Corporation
ruiyan@ruc.edu.cn, wuwei@microsoft.com

Abstract

Dialogue systems powered by conversational artificial intelligence (AI) have never been so popular. Interacting with computer through languages reveals a more natural interface to give orders and acquire information—just like human communication. Due to promising potential as *virtual assistants* and/or *social bots*, major NLP, AI and even Search & Mining communities are explicitly calling-out for contributions of conversational studies.

Learning towards real conversational intelligence is a trip to Mars; perhaps we are yet on Earth. We have achieved substantial progress from recent research outputs. Still we have major obstacles to overcome. In this paper, we present an overview of progress and look forward to future trends so as to shed light on possible directions towards success.

Overview of Conversational Studies

> "Twenty minutes of small talk with a computer isn't just a moonshot. It's a trip to Mars."

Starting from 1960s, research for conversational artificial intelligence (Conversational AI) has never been so popular as in recent years. Dialogue systems have great potential and commercial value (e.g., *personal assistants*, *agent systems*, and *social bots*), as real world applications.

There are two mainstream types of dialogue systems. Task-oriented dialogue systems are designed for helping people complete specific tasks such as question-answering (Ferrucci et al. 2013), bus information inquiry (Raux et al. 2005), etc. Non-task-oriented dialogue systems, a.k.a., chit-chat dialogue systems, aim to engage users in open domain human-machine conversation for entertainment and/or emotional companionship, namely social bot or chatbot. In this paper, we mainly focus on the second type of dialogue system but we will make a comparison in details.

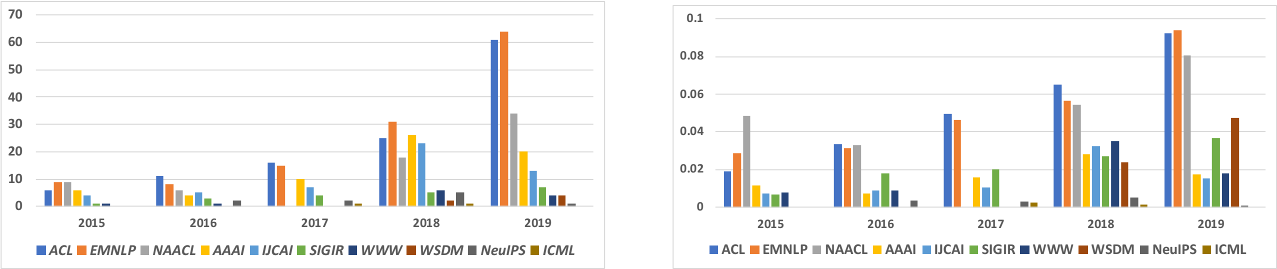
Dialogue systems are expected to take over as the main marketing and communication channel across numerous major industries. The large-scaled messaging capabilities powered by the automatic intelligent conversational agents are predicted to bring revolutions in business operation and managements. Therefore, companies, capitals, and even government officials, regard research and development in dialogue systems and products as the key driving force towards new profits. Apple Inc. launched Siri, the personal virtual assistant for every individual user. Google released several relevant products, from Google Assistant, to Google Home, and now Douplex. Amazon holds Alexa Prize Challenge every year to encourage cutting-edge research into daily dialogues. There are many more companies, big names or startups, working towards conversational AI.

Take Microsoft as an example. The chatbot dialogue system named Microsoft XiaoIce was first released to Chinese users in 2014, and then was launched in Japan (Rinna), U.S. (Zo), India (Ruuh), and Indonesia (Rinna). The products now have attracted more than 800 million users all over the world, and the technologies behind have powered a great number of business applications such as various third party official accounts and IoT (Internet-of-Things) devices. Until 2018, users from the 5 countries have finished more than 30 billion conversations with XiaoIce; and on average, each conversation lasts up to 23 turns. The promising user data indicate impressive popularity of the dialogue system based product in the real-world scenarios.

The need from industry stimulates research in academia. Major conferences and journals are now explicitly calling out for research articles of conversational studies. We conducted preliminary statistics about the number of accepted papers published at conversational AI related conferences within recent 5 years. The target venues for our investigation include ACL, EMNLP, NAACL, AAAI, IJCAI, NeurIPS, SIGIR, ICML, and others. We use a simple filter to identify the number of research papers about conversational studies, using the query keywords as "conversation" or "dialogue" or "response". The result is visualized in Figure 1.

It is obvious that the absolute (total) number of accepted papers about conversational studies and dialogue systems steadily keeps going up, as illustrated in Figure 1(a). The phenomenon may indicate that dialogue research becomes more and more popular, especially in the *Natural Language Processing* (NLP) community. Considering the recent AI boost which results in expansion in almost all research areas, we rule out such a factor by counting the ratio of accepted papers about dialogue systems against all accepted papers.

We have similar observations in Figure 1(b). The ratio of dialogue papers is also growing up in recent years, especially in major NLP conferences (ACL, NAACL, and EMNLP).



(a). Growth measured in paper numbers

(b). Growth measured in percentage

Figure 1: Statistics: the number of accepted dialogue papers published at conversational AI related conferences: (a) y-axis denotes the number of accepted papers in absolute numbers; (b) y-axis denotes the ratio of accepted dialogue papers against all accepted papers per venue per year.

Based on the observation, we assume that the communities draw the consensus that dialogue systems and conversational studies are developing fast, although not fully mature. This area now attracts more and more attention from researchers from both industry and academia. Hence, it is important to provide a brief summary report in this research area indicating 1) where are we now and 2) where we are going.

Preliminary: Dialogue Systems

Before diving into details of the non-task-oriented dialogue systems, let us first inspect the big picture of conversational AI. Task-oriented dialogue systems and non-task-oriented dialogue systems are different in various aspects, but in some way, the two are also related. Specifically, we compare the two types of dialogue systems as follows:

Domain of conversation. Conversations in task-oriented systems are often restricted within task-specific domains, while conversations in non-task-oriented dialogue often span over open domain topics without any constraint, although there can be special designs for specific domains.

User intent. In a task-oriented dialogue system, user intent is framed within a pre-defined task dependent schema, and every turn of a conversation can be expressed with a command-like semantic form such as $\texttt{request}(\texttt{restaurant}; \texttt{foodtype} = \texttt{Thai})$. For non-task-oriented dialogue systems, the user intent is much more diverse and complicated due to the nature of open domain conversations. Therefore, it is difficult to convey user intent with rigid logic forms, and a common practice is just to tag utterances with high-level dialogue acts (Jurafsky, Shriberg, and Biasca 1997) such as *statement, question, backchannel, etc.* or with high-level topics (Li et al. 2017) such as *tourism, politics, etc.*

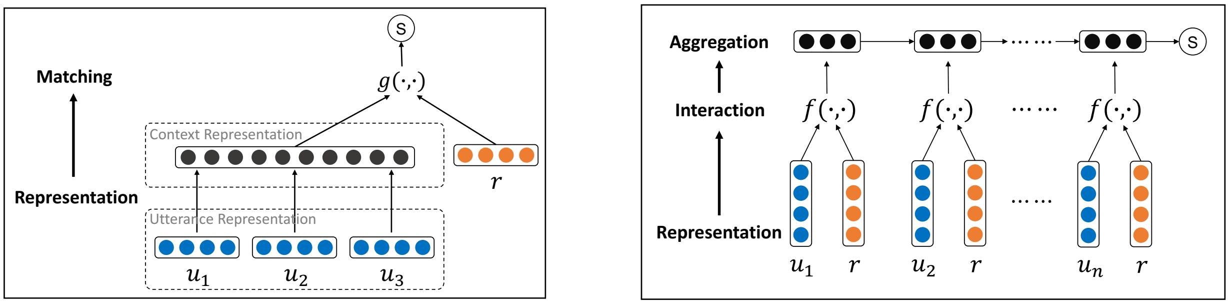
Architecture. Task-oriented dialogue systems are often built using a pipeline method with modules such as language understanding, dialogue management, and natural language generation (Gao et al. 2019). Each module can be independently constructed with rules or data-driven approaches. Although there are studies on end-to-end task-oriented dialogue systems (Wen et al. 2017; Bordes, Boureau, and Weston 2016), the modular approach is still favored by the mainstream community. Non-task-oriented dialogue systems, on the other hand, benefit from advances of neural machine learning and a large amount of human conversations available on the Web, and are often built in an end-to-end way with data-driven approaches.

Evaluation. Task-oriented dialogue systems are often evaluated with task completion rate (i.e., the ratio of dialogues that successfully solve users' problems in the end) (Williams and Zweig 2016) or the (average) number of turns to complete a task (Dhingra et al. 2017), although each of the modules can also be independently evaluated. On the contrary, non-task-oriented dialogue systems lack of standard evaluation criteria. Automatic evaluation is proven weakly correlated with human judgment (Liu et al. 2016; Yan 2018), while human evaluation is expensive and too subjective to reproduce. In spite of the challenges, existing work either evaluates quality of responses with perplexity (Xing et al. 2018), or includes word similarity with references (Liu et al. 2016; Tao et al. 2018; Tong et al. 2018). Moreover, it is feasible to just recruit human volunteers to conduct system-level evaluation according to their experience in talking with the conversational AI (Chen et al. 2018a; Fang et al. 2017).

Due to the aforementioned difference, the overlap between the research efforts on the two types of dialogue systems is generally not significantly large. Still there are some questions to be clarified. The first question is:

Are the two types of dialogues completely distinct? We believe the answer is a big "NO". In fact, task-oriented dialogues and chit-chat are not just black and white. For a conversational product to be successful, e.g., Amazon Alexa, the system needs task-oriented dialogues to help users to fulfill tasks, and also needs chit-chat as transitions from one task to another so as to connect task completion sessions with smooth coherence. Usually we see task-oriented dialogues and chit-chat dialogues interleave in real human-machine conversations. Therefore, we envision that with the two types of systems individually becoming better and better, an important research topic in the future is how to encapsulate task-oriented dialogues and chat-like dialogues in order to optimize user experience in real scenarios.

Another question is: **which type of dialogues is more useful?** We believe there is no clear answer to this question. Intuitively, task-oriented dialogue systems are helpful, and thus useful in a variety of ways. The main challenge lies in



(a). Framework I

(b). Framework II

Figure 2: Two matching frameworks: (a) embedding-based matching; and (b) interaction-based matching.

how to scale up to multiple domains for different tasks. With the release of datasets such as MultiWOZ (Budzianowski et al. 2018), there is rapid progress towards overcoming the challenge (Wu et al. 2019a; Chen et al. 2019). On the other hand, the motivation of building a system for chit-chat seems "debatable", but what is rational is actual and what is actual is rational. Famous products, like Alexa and Siri, are hybrid of task- and non-task dialogues. The reason could be 1) to coordinate task sessions, or 2) chatbots provide social companionship and emotional comforts to certain users.

Where Have We Been

Problem Formalization

Given a sequence of utterance $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ as the conversation history, where $n \geq 1$ and $\{c_i\}_{i=1}^{n}$ are arrayed in a temporal order, [1] the problem of human-machine dialogue can be formalized as seeking a function $\mathcal{F}(\cdot|\mathcal{C})$ such that a response r can be predicted as c_{n+1} in a machine-turn. Depending on how r is acquired, existing studies can be categorized into retrieval methods and generation methods. In retrieval methods, r is selected from a bunch of candidates $\{r_i\}_{i=1}^{m}$ which are often retrieved from an index of existing human conversations. Thus, $\mathcal{F}(\cdot|\mathcal{C})$ is defined as a matching function $g(\mathcal{C}, r)$ that measures how likely a candidate r_i to be a proper response after $\mathcal{C}$. The learning of $g(\cdot,\cdot)$ needs supervision, and is often performed with a set of triples $\{(y_i, \mathcal{C}_i, r_i)\}_{i=1}^{N}$ where y_i is a (binary) label indicating the matching degree between $\mathcal{C}_i$ and r_i (Fu et al. 2020a,b). In generation methods, r is synthesized by organizing tokens or words. Then $\mathcal{F}(\cdot|\mathcal{C})$ is defined as a conditional language model $P(r|\mathcal{C})$ as a distribution of r in the language space. Estimation of $P(r|\mathcal{C})$ only requires human conversation sessions $\{(\mathcal{C}_i, r_i\}_{i=1}^{N}$ without any annotation.

Retrieval Methods

A retrieval-based system for open domain dialogues is built upon the success of modern search engines (Ji, Lu, and Li 2014). Aside from indexing and ranking which have been well studied in search, a problem of great interest is how to effectively learn a matching model $g(\cdot,\cdot)$. With the advances of neural representation learning, matching accuracy has been greatly improved on benchmarks. On the other hand, the gap between offline static test and online conversation experience still exists.

While existing work defines $g(\cdot,\cdot)$ with various neural networks, the architectures can be classified into two frameworks. Figure 2(a) illustrates the architecture of Framework I. The framework works in a "representation → matching" procedure. The key idea is to simultaneously embed $\mathcal{C}$ and r into a space and then define $g(\mathcal{C}, r)$ as similarity of the embedding. Since $\mathcal{C}$ consists of multiple utterances, embedding of $\mathcal{C}$ is conducted in a hierarchical manner with $\{c_i\}_{i=1}^{n}$ firstly represented as vectors and then the utterance vectors abstracted as a vector of $\mathcal{C}$. Then, the research effort is paid to definitions of the embedding of $\mathcal{C}$ and r with conversational features designed with human expertise (Wang et al. 2013) or neural structures such as recurrent neural networks (Lowe et al. 2015; Kadlec, Schmid, and Kleindienst 2015; Yan, Zhao, and E. 2017) and convolutional neural networks (Hu et al. 2014; Zhou et al. 2016; Yan, Song, and Wu 2016; Yan and Zhao 2018a; Fu et al. 2020b).

Figure 2(b) shows the architecture of Framework II. Different from Framework I where matching between $\mathcal{C}$ and r happens at the last step, Framework II follows a "representation-interaction-aggregation" paradigm and lets each $c_i \in \mathcal{C}$ interact with r through a function $f(\cdot,\cdot)$ at the beginning. $\{f(c_i, r)\}_{i=1}^{n}$ are finally aggregated as $g(\mathcal{C}, r)$. The flexibility of the framework lies in definitions of the representations, $f(\cdot,\cdot)$, and the aggregation operation. Utterances and responses can be represented via an RNN model, a CNN model, a Transformer model (Vaswani et al. 2017), or a hybrid of these models. $f(\cdot,\cdot)$ is usually defined with either a (2D) CNN structure (Hu et al. 2014) or an attention mechanism (Wang and Jiang 2016). Aggregation is often achieved through an RNN in order to capture sequential relationship among $\{c_i\}_{i=1}^{n}$ (Wu et al. 2017, 2019b), but can also be replaced with a CNN (Zhou et al. 2018c) or a multi-layer perceptron (Wu et al. 2017). Because matching happens at an utterance-level rather than at a context-level, Framework II is able to capture more matching information in a context-response pair than Framework I. Models proposed in recent years, such as sequential matching network (SMN) (Wu et al. 2017), deep attention matching network (DAM)

[1] The research of open domain dialogues starts from a single-turn assumption where $n = 1$, but now focuses on a more natural multi-turn assumption where $n > 1$.

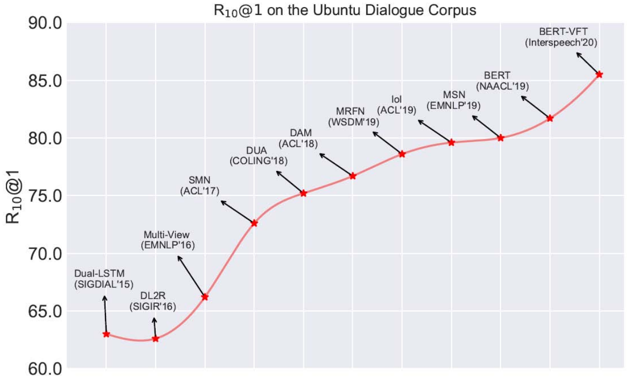


Figure 3: Evolution trajectory of matching models: a view of state-of-the-art leader board of Ubuntu Dialogue Corpus.

(Zhou et al. 2018c), deep utterance aggregation (Zhang et al. 2018b), multi-representation fusion network (MRFN) (Tao et al. 2019a), interaction-over-interaction (IoI) (Tao et al. 2019b), and multi-hop selector network (MSN) (Yuan et al. 2019), generally belong to Framework II as special cases.

One appealing feature of retrieval methods is that there are well-established benchmarks and evaluation methods for model comparison. To be specific, given $\mathcal{C}$ and $\{(r_i, y_i)\}_{i=1}^{m}$ as a context and a bunch of response candidates with labels, one can rank $\{r_i\}_{i=1}^{m}$ according to $\{g(\mathcal{C}, r_i)\}_{i=1}^{m}$, and then evaluate the performance of $g(\cdot,\cdot)$ with methods in learning to rank (Liu et al. 2009) based on $\{y_i\}_{i=1}^{m}$. Widely applied benchmark datasets include the Ubuntu Dialogue Corpus (UDC) (Lowe et al. 2015), Douban Conversation Corpus (Douban) (Wu et al. 2017), and E-Commerce Dialogue Corpus (E-Commerce) (Zhang et al. 2018b). Among the datasets, UDC, due to its largest size of the test set, is often treated as a ruler with the metric $R_{10}@1$ (i.e., the fraction of contexts where the only positive response is ranked at the top position among the 10 candidates) for model comparison as a leader board. Figure 3 illustrates the performance of matching models with respect to the metric on UDC.

On one hand, it seems to be optimistic that the leap from Dual-LSTM to BERT-VFT (Whang et al. 2020) paves the way for intelligent human-machine conversation; on the other hand, performance of the models in real human-machine conversation is actually misled by the *static* test sets, since the data (e.g., UDC) are automatically constructed and thus contain noise and many easy patterns. The original test set of these benchmark datasets are generally automatically built with response retrieval. Researchers recruit human annotators to label every context-response pair with $\{1, 0\}$ where 1 means that the response appropriately replies to the context from the human judges' perspective. Unfortunately, state-of-the-art matching architectures, such as SMN and DAM, generally suffer from dramatic performance drop given such a conversation experience based test.

Therefore, an important task in the future could be bridging the gap between performance on automatically built static test sets and performance in real human-machine conversations. A recent shift from designing more sophisticated neural matching architectures to pursuing robust learning approaches for the existing architectures (Wu et al. 2018; Feng et al. 2019) seems to take a step towards the problem, but what the emerging studies have achieved (e.g., $\sim$2% improvement) is still far from enough.

Generation Methods

At the other end of the spectrum, generation methods aim to recover $P(r|\mathcal{C})$ from observed conversations. The idea of generation for open domain dialogues with data-driven approaches is inspired by the research on machine translation, spanning over the age of statistical machine translation (Ritter, Cherry, and Dolan 2011) and the age of neural machine translation (Shang, Lu, and Li 2015). Similar to machine translation, modeling of $P(r|\mathcal{C})$ is based on the encoder-decoder architecture (Sutskever, Vinyals, and Le 2014; Vaswani et al. 2017). The encoder first transforms $\mathcal{C}$ into a sequence of hidden vectors $\mathcal{H} = (h_1, \ldots, h_{n'})$, and then the decoder predicts the words (or tokens) in r one by one from a sequence of latent states $\mathcal{S} = (s_1, \ldots, s_{m'})$ by attending to the relevant parts of $\mathcal{H}$ (Shang, Lu, and Li 2015). Common implementations of encoder-decoder include LSTM, GRU, and the Transformer structures (Vaswani et al. 2017). The research of generation methods expands from the encoder-decoder architecture, by defining and learning $P(r|\mathcal{C})$ from various dimensions. Hereby we illustrate typical directions for generative dialogue systems.

Relevance to Context. The most basic but fundamental requirement for dialogue system is to maintain relevance throughout the conversation session. Intuitively, the contexts can be modeled as a simple representation concatenation to the current query representation in a non-hierarchical way (Sordoni et al. 2015). Later, dialogues are modeled as word-level for each utterance sentence and then an utterance-level to keep the global information passed through turns, which is a hierarchical structure (Serban et al. 2016, 2017; Xing et al. 2018; Tian et al. 2017). Variational methods have been introduced to model contexts with the latent variable in addition to structures (Serban et al. 2017). Memory-based networks are incorporated, which supports human-like operations such as reading, writing and updating (Sukhbaatar et al. 2015; Graves, Wayne, and Danihelka 2014; Yan and Zhao 2018b; Chen et al. 2018b).

For all these methods, efforts have been paid to identify important information from previous utterances, and then to fuse the information into the decoding process. In this way, the generated responses are relevant to the contexts.

One-to-Many Diversity. Given a particular query utterance, there can be multiple different responses, which are totally dissimilar to each other, but they are all appropriate to respond the query utterance. Such a phenomenon is known as "one-to-many" diversity in conversations. However, trivial and non-committal responses take up a large portion in conversational data because these responses are universal enough to respond many utterances (Li et al. 2016a). Hence, the dialogue systems are likely to learn the frequently used patterns and output generic responses.

A simple solution is to lower the probability and weights of generic utterances to penalize them (Li et al. 2016a). Another intuitive way is to re-rank the candidate tokens to generate during the decoding phase from the N-best candidate token lists (Li, Monroe, and Jurafsky 2016). Determinantal Point Process (DPP) is a diversity-oriented metric to balance between quality and diversity to generate diverse responses by diversifying the beam search process (Song et al. 2018). CVAE learns latent variables to depict a distribution over potential conversational intents and generates diverse responses accordingly (Zhao, Zhao, and Eskenazi 2017; Serban et al. 2017). A recent study reveals the correlation between one query and multi-references with 2-step CVAE model for diverse response generation (Qiu et al. 2019).

For these models, a strength is that they manage to increase diversity in generation while a weakness is shortage of intrinsic understanding how diversity is expressed in dialogues, which results in lack of interpretability for diversity.

Human-Like Factors. Human-like factors are to the interest of a wide spectrum of audience. We illustrate some human factors under investigation.

Generally, people have unique ways to express utterances in conversations, namely "*persona*". The persona models have been proposed with speaker-level vector representation (Li et al. 2016b) or multi-modal distributions over utterance and speakers (Chan et al. 2019). To distinguish the roles of speakers and addressees is necessary since a speaker needs to express differently to different addressees (Ouchi and Tsuboi 2016), especially for group conversations (Le et al. 2019; Hu et al. 2019). Interactive recurrent neural network structures have been proposed to model speaker and addressee representations so that the conversation is characterized by interactions (Zhang et al. 2018a).

"*Emotions*" are the unique expressions by humans: happy, sad, angry, etc. Perception and expression of emotions are key factors to human-like dialogues. Researchers tried end-to-end learning to equip systems with the ability to perceive emotions and then express emotions such as emoji (Zhou and Wang 2018). Emotional Chatting Machine decides which emotion to express and then incorporates an internal and external emotion states for balancing semantic information and emotions dynamically (Zhou et al. 2018a; Wei et al. 2019). Emotion is further formulated as a transition network with controlling functions in dialogues (Qiu et al. 2020).

Now we revisited current progress of generation methods. The **Pros** of various research objectives indicate vibrant studies towards human-like conversational intelligence from multiple aspects. However, the **Cons** are also prominent: unlike the retrieval methods, there is *no benchmark test set* or *universal evaluation metrics* for all research directions. Researchers generally conduct experiments on characteristics-featured data sets and evaluate based on proposed objectives.

Next Steps to Mars

Pre-training

Encouraged by the breakthrough from BERT (Devlin et al. 2019), the research community began to realize the power of pre-training, and there is a surge of interest on pre-training for NLP tasks in both language understanding (Joshi et al. 2019; Liu et al. 2019; Yang et al. 2019) and language generation (Radford et al. 2018; Song et al. 2019; Dong et al. 2019). In terms of open domain dialogues, the outstanding performance of the systems from Team Lost in Conversation and Team Hugging Face (Wolf et al. 2019) on the second conversational intelligence challenge (convAI2) sheds light on the future of natural human-machine dialogues with pre-training techniques such as the GPT-3 model by Brown et al. (2020). In fact, there is a clear trend that pre-trained models are being used to improve various sub-tasks in dialogues such as response selection (Whang et al. 2019; Xu et al. 2021), dialogue act prediction (Mehri et al. 2019), and response generation (Zhang et al. 2019; Zhao et al. 2020b).

A recent report from Microsoft (Zhang et al. 2019) seems to indicate that pre-training is able to solve some long-standing problems such as "generic responses" and "lack of commonsense" in open domain dialogue generation and achieve human parity on response quality. Therefore, in near future, just like in other language tasks, pre-training → fine-tuning may become the new paradigm of building dialogue systems, especially when one cannot collect enough dialogues (e.g., for some specific domain or in multiple modalities) via end-to-end training. At the same time, some new problems may surface: 1) how to handle the huge cost from big pre-trained models in an online environment; 2) how to interpret the mechanism of the pre-trained models and thus engineers can effectively debug the systems as before.

"Knowledge is Power"

In human conversations, utterances are often grounded on external knowledge, such as commonsense from a knowledge base, documents, tables, etc. It is weird for a dialogue system to say "the sun rises from the west every day". The sentence is absolutely correct in grammar, but violates commonsense. It is believed to be rather essential to equip dialogue systems with knowledge grounding towards better conversational experience.

Knowledge-grounded utterance generation is firstly investigated for Knowledge-Based Question-and-Answering (KB-QA) (He et al. 2017). In dialogues, a Tri-LSTM model was proposed to use commonsense knowledge as external memories to facilitate LSTMs to encode commonsense assertions in order to enhance response selection (Young et al. 2018). Ghazvininejad et al. (2018) extend the traditional encoder-decoder model by considering both dialogue history and external "facts" from Wikipedia for response generation. Beyond triplets from the knowledge base, knowledge graph is also incorporated into response generation by dynamic querying and integration with the graph information (Zhou et al. 2018b). In addition to knowledge graph, many researchers are dedicated to utilizing Web knowledge for response generation or response selection (Hua et al. 2020). Zhou et al. (2018) release a data set where human conversations are grounded in a set of movie-related documents from Wikipedia. Dinan et al. (2019) further release another document-grounded data set with Wiki articles covering broader topics. Knowledge-augmented responses are generated based on these datasets (Zhao et al. 2020a).

Yet, there are **pain points** for current knowledge-aware dialogue systems. The existing knowledge, either knowledge base or knowledge graph, is too sparse for daily conversations. People can talk about anything in dialogues but definitely we do not have everything available in the prerequisite knowledge repository. Another problem is that in its current form, knowledge reasoning is also a bottleneck. To this end, we expect that a universal schema to extract knowledge from dialogue contents and to build the knowledge repository on-the-fly will be the key to success for knowledge-aware dialogues. The knowledge shall be extracted, updated (with accumulation and reasoning), and then be fused into future dialogues dynamically when applicable.

Multi-Modal Dialogues

Humans converse based on multiple channels of senses. Intelligent conversational agents are expected to be capable of leveraging multi-modal signals, such as textual information, visual information, emotional information, audio information, etc. in their interaction with humans. Challenges to multi-modal conversation are generally two-fold: 1) how to model the multi-modal information in response generation/selection; and 2) how to effectively learn such a model with data-driven approaches. Some recent studies have touched the first challenge: Mostafazadeh et al. (2017) ground open domain response generation by images; Huber et al. (2018) leverage conversation context, visual sentiment, and facial expression for response generation. Gao et al. (2020) propose to combine response learning with texts and visual stickers as well as emoji.

A second challenge, however, is still left open. Since neural models, although powerful in terms of representation capability, are data-hungry, but multi-modal conversation data by nature are more difficult to obtain than single-modal data. As a remedy, some researchers work on dataset construction with crowd-sourcing (Shuster et al. 2018). However, such datasets are usually small in scale (e.g., several thousands). Therefore, semi-supervised learning and unsupervised learning could be the major concern in future research of multi-modal dialogues. After all, there is abundant plain text, tables, images, audios, and videos on Web, and the only problem is that they are not naturally collated.

Future Applications

With the rapid progress of conversational intelligence, we believe that the potential of dialogue systems is beyond what we have witnessed so far on social bots and virtual assistants. In this section, we illustrate some promising scenarios where open domain dialogues could be useful. Some of them have surfaced a bit, and we also believe that there are more to come in the society. Here, we highlight some directions:

Conversational Search. Since 5-6 years ago, big search players, such as Google and Microsoft, have been working on how to make their search service more conversational. For example, Google allowed users to speak their search on Chrome in 2013. Open domain dialogues, especially after they are well powered by knowledge, could greatly enhance the experience of conversational search by re-shaping it as a multi-turn question-answering and/or information seeking process in multi-domain.

Conversational Recommendation. Information provision is not totally passive anymore. Agents can proactively recommend relevant information to users during proper timing of conversation based on understanding on users' interest and intentions. The systems are even able to transfer knowledge from one user to others in a privacy-safe way. Conversational recommendation is likely going to act as information exchange in people's daily communication.

Internet-of-Things (IoT). With the success of smart speakers, e.g., Amazon Echo and Google Home, it seems no doubt that the physical world could become more connected with conversational intelligence in the future. No matter task commands or information requests, all we need to do is just to speak. People will embrace a smarter life with advanced dialogue technologies in which casual chats make things happen in a natural way.

Entertainments. Personalized and informative chat will change the way we entertain. Games will become more immersive when people can interact with characters in them rather than just experiencing what has been designed; virtual idols will be able to sing, dance, and talk to everyone; kids can make friends with their robots, just like Hiro and Baymax in Sci-Fi movies. Although intelligent robots seem to be far away, we will eventually have them in our daily life, and the smart speaker indicates just a beginning.

To sum up, we are likely to have more industries and markets that conversational AI will play an important role and make a big change. Researchers and practitioners are striving to improve the intelligence of dialogues systems and make it more inviting in reality.

Conclusions

We have witnessed a rapid surge of conversational studies in the past few years, especially the dialogue systems in the open domain. The research community of conversational AI is expanding and companies are making great efforts to develop conversational products due to the great potential value: conversational research proceeds with prosperity.

In this paper, we have systematically summarized the overview of current progress, where we have been and where we are going in the future. We are entering the AI era whereby large-scale data become more easily available and learning techniques become more powerful. We may stand at the entrance of future success in more advanced dialogue systems. Although we still face bottlenecks and obstacles to improve conversational AI, there is a reason for us to be optimistic about the future of dialogue systems when more efforts are devoted and key problems are solved.

Acknowledgments

We thank all anonymous reviewers for their valuable comments, and we thank Chongyang Tao and Ran Le for their help to prepare the manuscript. The work is supported by National Key R&D Program of China (2020YFB1406702), NSFC Grant No. 61876196 and Beijing Outstanding Young Scientist Program No. BJJWZYJH012019100020098. Rui Yan is supported as a young fellow at Beijing Academy of Artificial Intelligence (BAAI).

References

Bordes, A.; Boureau, Y.-L.; and Weston, J. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683* .

Brown, T.; and et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS'20*, 1–25.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP'18*, 5016–5026.

Chan, Z.; Li, J.; Yang, X.; Chen, X.; Hu, W.; Zhao, D.; and Yan, R. 2019. Modeling Personalization in Continuous Space for Response Generation via Augmented Wasserstein Autoencoders. In *EMNLP-IJCNLP'19*, 1931–1940.

Chen, C.-Y.; Yu, D.; Wen, W.; Yang, Y. M.; Zhang, J.; Zhou, M.; Jesse, K.; Chau, A.; Bhowmick, A.; Iyer, S.; et al. 2018a. Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *Proc. Alexa Prize* .

Chen, H.; Ren, Z.; Tang, J.; Zhao, Y. E.; and Yin, D. 2018b. Hierarchical variational memory network for dialogue generation. In *WWW'18*, 1653–1662.

Chen, W.; Chen, J.; Qin, P.; Yan, X.; and Wang, W. Y. 2019. Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention. In *ACL'19*, 3696–3709.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL'19*, 4171–4186.

Dhingra, B.; Li, L.; Li, X.; Gao, J.; Chen, Y.-N.; Ahmed, F.; and Deng, L. 2017. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access. In *ACL'17*, 484–495.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098* .

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. *arXiv:1905.03197* .

Fang, H.; Cheng, H.; Clark, E.; Holtzman, A.; Sap, M.; Ostendorf, M.; Choi, Y.; and Smith, N. A. 2017. Sounding board–university of washington's alexa prize submission. *Alexa prize proceedings* .

Feng, J.; Tao, C.; Wu, W.; Feng, Y.; Zhao, D.; and Yan, R. 2019. Learning a Matching Model with Co-teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems. In *ACL'19*, 3805–3815.

Ferrucci, D.; Levas, A.; Bagchi, S.; Gondek, D.; and Mueller, E. T. 2013. Watson: beyond jeopardy! *Artificial Intelligence* 199: 93–105.

Fu, Z.; Cui, S.; Ji, F.; Zhang, J.; Chen, H.; Zhao, D.; and Yan, R. 2020a. Query-to-Session Matching: Do NOT Forget History and Future during Response Selection for Multi-Turn Dialogue Systems. In *CIKM'20*, 365–374.

Fu, Z.; Cui, S.; Shang, M.; Ji, F.; Zhao, D.; Chen, H.; and Yan, R. 2020b. Context-to-Session Matching: Utilizing Whole Session for Response Selection in Information-Seeking Dialogue Systems. In *KDD'20*, 1605–1613.

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

Gao, S.; Chen, X.; Liu, C.; Liu, L.; Zhao, D.; and Yan, R. 2020. Learning to Respond with Stickers: A Framework of Unifying Multi-Modality in Multi-Turn Dialog. In *WWW'20*, 1138–1148.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *AAAI'18*, 5110–5117.

Graves, A.; Wayne, G.; and Danihelka, I. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* .

He, S.; Liu, C.; Liu, K.; and Zhao, J. 2017. Generating Natural Answers by Incorporating Copying and Retrieving Mechanisms in Sequence-to-Sequence Learning. In *ACL'17*, 199–208.

Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional neural network architectures for matching natural language sentences. In *NIPS'14*, 2042–2050.

Hu, W.; Chan, Z.; Liu, B.; Zhao, D.; Ma, J.; and Yan, R. 2019. GSN: a graph-structured network for multi-party dialogues. In *IJCAI'19*, 5010–5016.

Hua, K.; Feng, Z.; Tao, C.; Yan, R.; and Zhang, L. 2020. Learning to Detect Relevant Contexts and Knowledge for Response Selection in Retrieval-based Dialogue Systems. In *CIKM'20*, 525–534.

Huber, B.; McDuff, D.; Brockett, C.; Galley, M.; and Dolan, B. 2018. Emotional dialogue generation using image-grounded language models. In *SIGCHI'18*, 277.

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

Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529* .

Jurafsky, D.; Shriberg, E.; and Biasca, D. 1997. Switchboard-damsl labeling project coder's manual. *Technická Zpráva* 97–02.

Kadlec, R.; Schmid, M.; and Kleindienst, J. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753* .

Le, R.; Hu, W.; Shang, M.; You, Z.; Bing, L.; Zhao, D.; and Yan, R. 2019. Who Is Speaking to Whom? Learning to Identify Utterance Addressee in Multi-Party Conversations. In *EMNLP-IJCNLP'19*, 1909–1919.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL'16*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A Persona-Based Neural Conversation Model. In *ACL'16*, 994–1003.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562* .

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP'17*, 986–995.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP'16*, 2122–2132.

Liu, T.-Y.; et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3(3): 225–331.

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

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL'15*, 285–294.

Mehri, S.; Razumovsakaia, E.; Zhao, T.; and Eskenazi, M. 2019. Pretraining Methods for Dialog Context Representation Learning. *arXiv preprint arXiv:1906.00414* .

Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G.; and Vanderwende, L. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In *IJCNLP'17*.

Ouchi, H.; and Tsuboi, Y. 2016. Addressee and Response Selection for Multi-Party Conversation. In *EMNLP'16*, 2133–2143.

Qiu, L.; Li, J.; Bi, W.; Zhao, D.; and Yan, R. 2019. Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References. In *ACL'19*, 3826–3835.

Qiu, L.; Shiu, Y.; Lin, P.; Song, R.; Liu, Y.; Zhao, D.; and Yan, R. 2020. What If Bots Feel Moods? In *SIGIR'20*, 1161–1170.

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

Raux, A.; Langner, B.; Bohus, D.; Black, A. W.; and Eskenazi, M. 2005. Let's Go Public! Taking a spoken dialog system to the real world. In *Ninth European conference on speech communication and technology*.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-Driven Response Generation in Social Media. In *EMNLP'11*, 583–593.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI'16*, 3776–3783.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI'17*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL'15*, 1577–1586.

Shuster, K.; Humeau, S.; Bordes, A.; and Weston, J. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945* .

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *International Conference on Machine Learning*, 5926–5936.

Song, Y.; Yan, R.; Feng, Y.; Zhang, Y.; Zhao, D.; and Zhang, M. 2018. Towards a Neural Conversation Model with Diversity Net Using Determinantal Point Processes. In *AAAI'18*, 5932–5939.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *NAACL-HLT'15*, 196–205.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, 2440–2448.

Sutskever, I.; Vinyals, O.; and Le, Q. 2014. Sequence to sequence learning with neural networks. In *NIPS'14*.

Tao, C.; Mou, L.; Zhao, D.; and Yan, R. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI'18*, 722–729.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 2019a. Multi-Representation Fusion Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *WSDM'19*, 267–275. ACM.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 2019b. One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues. In *ACL'19*, 1–11.

Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; and Zhao, D. 2017. How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models. In *ACL'17*, 231–236.

Tong, X.; Fu, Z.; Shang, M.; Zhao, D.; and Yan, R. 2018. One "Ruler" for all languages: multi-lingual dialogue evaluation with adversarial multi-task learning. In *IJCAI-ECAI'18*, 4432–4438.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS'17*, 5998–6008.

Wang, H.; Lu, Z.; Li, H.; and Chen, E. 2013. A Dataset for Research on Short-Text Conversations. In *EMNLP'13*, 935–945. URL http://aclweb.org/anthology/D13-1096.

Wang, S.; and Jiang, J. 2016. Learning Natural Language Inference with LSTM. In *NAACL-HLT'16*, 1442–1451.

Wei, W.; Liu, J.; Mao, X.; Guo, G.; Zhu, F.; Zhou, P.; and Hu, Y. 2019. Emotion-aware Chat Machine: Automatic Emotional Response Generation for Human-like Emotional Interaction. In *CIKM'19*, 1401–1410. ACM.

Wen, T.; Vandyke, D.; Mrkšíc, N.; Gašíc, M.; Rojas-Barahona, L.; Su, P.; Ultes, S.; and Young, S. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL'17*, 438–449.

Whang, T.; Lee, D.; Lee, C.; Yang, K.; Oh, D.; and Lim, H. 2019. Domain Adaptive Training BERT for Response Selection. *arXiv preprint arXiv:1908.04812* .

Whang, T.; Lee, D.; Lee, C.; Yang, K.; Oh, D.; and Lim, H. 2020. An effective domain adaptive post-training method for bert in response selection. In *Proc. Interspeech*.

Williams, J. D.; and Zweig, G. 2016. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269* .

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

Wu, C.-S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; and Fung, P. 2019a. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *ACL'19*, 808–819.

Wu, Y.; Wu, W.; Li, Z.; and Zhou, M. 2018. Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. In *ACL'18*, 420–425.

Wu, Y.; Wu, W.; Xing, C.; Xu, C.; Li, Z.; and Zhou, M. 2019b. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Computational Linguistics*, volume 45, 163–197.

Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *ACL'17*, 496–505.

Xing, C.; Wu, Y.; Wu, W.; Huang, Y.; and Zhou, M. 2018. Hierarchical recurrent attention network for response generation. In *AAAI'18*, 5610–5617.

Xu, R.; Tao, C.; Jiang, D.; Zhao, X.; Zhao, D.; and Yan, R. 2021. Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues. In *AAAI'21*.

Yan, R. 2018. "Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI. In *IJCAI-ECAI'18*, 5520–5526.

Yan, R.; Song, Y.; and Wu, H. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR'16*, 55–64. ACM.

Yan, R.; and Zhao, D. 2018a. Coupled context modeling for deep chit-chat: towards conversations between human and computer. In *KDD'18*, 2574–2583.

Yan, R.; and Zhao, D. 2018b. Smarter Response with Proactive Suggestion: A New Generative Neural Conversation Paradigm. In *IJCAI-ECAI'18*, 4525–4531.

Yan, R.; Zhao, D.; and E., W. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *SIGIR'17*, 685–694.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237* .

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting End-to-End Dialog Systems with Commonsense Knowledge. In *AAAI'18*.

Yuan, C.; Zhou, W.; Li, M.; Lv, S.; Zhu, F.; Han, J.; and Hu, S. 2019. Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots. In *EMNLP-IJCNLP'19*, 111–120.

Zhang, R.; Lee, H.; Polymenakos, L.; and Radev, D. 2018a. Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs. In *AAAI'18*, 5690–5697.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv preprint arXiv:1911.00536* .

Zhang, Z.; Li, J.; Zhu, P.; Zhao, H.; and Liu, G. 2018b. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In *COLING'18*, 3740–3752.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL'17*.

Zhao, X.; Wu, W.; Tao, C.; Xu, C.; Zhao, D.; and Yan, R. 2020a. Low-Resource Knowledge-Grounded Dialogue Generation. In *ICLR'20*.

Zhao, X.; Wu, W.; Xu, C.; Tao, C.; Zhao, D.; and Yan, R. 2020b. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *EMNLP'20*, 3377–3390.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018a. Emotional chatting machine: emotional conversation generation with internal and external memory. In *AAAI'18*, 730–738.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018b. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI-ECAI'18*, 4623–4629.

Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yu, D.; Tian, H.; Liu, X.; and Yan, R. 2016. Multi-view Response Selection for Human-Computer Conversation. In *EMNLP'16*, 372–381.

Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W. X.; Yu, D.; and Wu, H. 2018c. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *ACL'18*, 1118–1127.

Zhou, X.; and Wang, W. Y. 2018. MojiTalk: Generating Emotional Responses at Scale. In *ACL'18*, 1128–1137.