Open Questions for Next Generation Chatbots

Winson Ye
Computer Science Department
College of William and Mary
Williamsburg, VA
wye@email.wm.edu

Qun Li
Computer Science Department
College of William and Mary
Williamsburg, VA
liqun@cs.wm.edu

***Abstract*—Over the last few years, there has been a growing interest in developing chatbots that can converse intelligently with humans. For example, consider Microsoft's Xiaoice. It is a highly intelligent dialogue system that serves as both a social companion and a virtual assistant. Targeted towards Chinese users, Xiaoice is connected to 660 million online users and 450 million IoT devices. Because of the deep learning revolution, the field is moving quickly, so this survey aims to introduce newcomers to the most fundamental research questions for next generation neural dialogue systems. In particular, our analysis of the state of the art reveals the following 4 key research challenges: 1) knowledge grounding, 2) persona consistency, 3) emotional intelligence, and 4) evaluation. Knowledge grounding endows the chatbot with external knowledge to generate more informative replies. Persona consistency grants dialogue systems consistent personalities. We divide each fundamental research challenge into several smaller and more concrete research questions. For each fine grained research challenge, we examine state of the art approaches and propose future research directions.**

***Index Terms*—dialogue system, chatbot, virtual assistant, conversational response generation, conversational AI**

I. Introduction

Interest in artifically intelligent chatbots can be traced back to the Turing Test in the 1950s, where a computer that could fool a human into thinking that they were conversing with another human was deemed to have human-like intelligence. In modern times, we enjoy the ubiquitous presence of virtual assistants that can speak with human level fluency. These chatbots are referred to as goal oriented dialogue systems since they can only converse with the user in a restricted domain like reservation booking. A more ambitious variant is the open domain chatbot, which is designed to chit chat with the user on any topic, much like ELIZA. Spearheaded by the deep learning revolution, research on modern open domain chatbots has focused heavily on designing data driven systems rather than rule based ones. There are two types of open domain chatbots: retrieval based and generative based. Retrieval based chatbots aim to select the best response out of a repository of candidate responses. Generative based chatbots directly synthesize replies. To focus this survey, we only comment on generative models. Overall, the goal of developing better chatbots is to increase user engagement in existing use cases and pave the way for future applications. For example, Yan et al. [1] report that even in task oriented dialogue systems for scenarios like online shopping, 80% of the utterances are chit chat related. Future applications will incorporate emotionally intelligent chatbots in video games, social companions, or virtual therapists.

Since 2014, with the work of Sutskever et al. [2], seq2seq models have been the standard for neural dialogue generation. Seq2seq models solve sequence transduction problems. That is, given an input sequence X, how can the neural network generate the correct output sequence Y? For chatbots, X is the user input while Y is the AI's reply. To solve this problem, a seq2seq model uses a neural network consisting of an encoder and a decoder. The encoder transforms the input text into a feature vector while the decoder takes this high dimensional feature vector and generates the actual reply. The key idea is that the encoder is able to capture how related certain words are by projecting them into a high dimensional feature space. The decoder then takes this information and interprets it in order to generate a coherent response. In 2017, a new architecture dubbed the Transformer [3] has demonstrated state of the art results in various natural language processing (NLP) tasks such as machine translation. Thus, researchers have begun gravitating towards this new model for neural dialogue generation.

Overall, there are four coarse grained research challenges that will guide this survey. First, how can researchers encourage the chatbot to generate more interesting replies grounded in external knowledge? We find this question essential because a key aspect of human conversation is the ability to refer to external knowledge. Others may refer to this as having some "common sense." Thus, truly intelligent chatbots must have some kind of knowledge grounding mechanism. Second, how can we give the chatbot a consistent personality? Chatbots with consistent personalities are less likely to contradict themselves and seem more human. Third, how can we teach the chatbot to generate more emotionally intelligent replies? Without emotional intelligence, the chatbot cannot empathize with the user, so engagement will plummet. Fourth, how can we develop more accurate and efficient evaluation metrics for chatbots? In order to make any progress in developing next generation neural dialogue systems, researchers must be able to quantify progress. However, accurately measuring the quality of generated dialogue is difficult as it involves several very abstract notions.

One may wonder how our approach is different from other recent surveys on dialogue systems [4], [5] released over the last couple of years. Compared to Chen et al., our paper

focuses more heavily on key research challenges rather than a overview of methods. Huang et al. is the most similar to our work, but keep in mind this paper is designed to be a short primer for newcomers. We encourage those looking for a fully comprehensive view of dialogue system challenges to consult their paper. Que

II. Knowledge Grounding

Knowledge grounding is the act of embedding external facts into the chatbot generated dialogue. For example, if the user is talking about their favorite restaurants, the chatbot should be able to reference similar restaurants in its reply. To address this research challenge, researchers must consider multiple research questions:

1) What kind of information should be fed to the chatbot?
2) How should this information be collected?
3) What architecture is best for this task?

A. RQ1: Knowledge Domain

The ideal knowledge base will consist of a large variety of diverse facts that are likely to show up in everyday conversation. The developer can either focus on general purpose knowledge or domain specific knowledge.

There are many examples of general purpose knowledge bases. The first one that comes to mind may be Wikipedia [6]. Another possible source is social media, such as Twitter [7], Facebook, or Reddit [8]. These sources help the chatbot learn general information such as pop culture or politics.

For deep knowledge integration of a specific topic, the developer must consult domain specific sources, such as restaurant reviews [9] for chatbots primarily designed for tourist attraction. Other sources include Stack Overflow, IMDb, etc. In fact, there seems to be a shortage of papers that focus on crawling online question and answer forums, even though these knowledge sources feature rich examples of people integrating domain specific knowledge into their conversations.

Future research directions: 1) Develop chatbots that can learn from both general purpose and domain specific facts. 2) Exploit data from question and answer forums to show specific examples of knowledge integration to your chatbots. 3) Develop metrics to measure the quality of a particular knowledge base, such as diversity of topics covered and quality of the information sources embedded in the dialogue.

B. RQ2: Dataset Creation

After an appropriate knowledge base is selected, the developers must create the datasets. Two datasets are usually created: 1) a dataset of external facts and 2) a dataset of conversational exchanges.

Creating a dataset of external facts could involve heavy preprocessing. For example, Xing et al. [7] extract topic words from Twitter by using an LDA model to cluster words by topic and then filter out universal words like *a*. Dinan et al. [6] use an information retrieval system to select the most relevant Wikipedia passages for a given conversation topic. Qin et al. [8] convert raw Wikipedia pages to html documents in order to retain just the text. Sometimes just the raw facts are fine. For instance, Ghazvininejad et al. [9] show that plain resturant reviews are adequate.

To learn how to actually generate responses, chatbots need to learn from a dataset of message response pairs. One approach is to collect dialogue samples of a domain expert responding to questions from a curious conversation partner [6]. Alternatively, to teach the chatbot examples of commonsense in every conversations, one could extract conversations from Reddit and filter them by commonsense knowledge relations [10]. To help the chatbot understand how people cite sources, Qin et al. [8] extract messages from Reddit that cite web documents. Recall that in some scenarios, the message response pairs may just be plain conversation and another dataset is used for teaching the chatbot about external facts. In this case, some authors just collect plain message response pairs from Twitter [7], [9].

Future research directions: 1) Develop robust preprocessing pipelines for filtering out noise such as offensive language from conversational exchange datasets. 2) Lower the amount of user effort and cost involved in collecting datasets.

C. RQ3: Chatbot Architecture

Once the datasets have been established, it is time to plan out the chatbot architecture. So far, two approaches have surfaced in the state of the art. One traditional approach is to use an encoder to embed external facts into the feature space of the neural network. More recent works build upon the traditional approach by suggesting that the networks should have some kind of memory.

Simple encoder based approaches aim to translate abstract ideas into concrete vectorized representations, otherwise known as embeddings. The collection of all embeddings can be understood as a high dimensional feature space used to compare two different facts. Facts that are highly related are close to each other as measured by metrics like cosine distance. The encoder can be used to encode topics [7], commonsense knowledge relations [10], or world facts [9]. Note that this encoder design is simply an implementation of the classic seq2seq model.

The other approach is to draw upon some kind of memory component in the neural network to generate the dialogue. Note that memory networks have demonstrated better long term memory than standard long short term memory networks (LSTMS). Usually memory networks can be understood as a module of the neural network that can be read from and written to. Dinan et al. [6] construct a memory network by combining an information retrieval system and an encoder. Qin et al. [8] develop a new notion of memory that is distinct from memory networks. Essentially, they use a set of encoders and attention mechanisms to pool information coming from external facts and the conversation history. Attention mechanisms allow the chatbot to focus on the most important keywords in the conversation. For example, when asking someone to perform a task, the name of the person you are addressing is one of the most important words.

Future research directions. 1) Develop new attention mechanisms that allow the chatbot to pay special attention to the most relevant external facts for the conversation. 2) Consider training the chatbot on multiple tasks simultaneously. For example, instead of training the chatbot to generate plain dialogue and then training on the external facts, could we do this simultaneously? 3) Consider using Transformers instead of a seq2seq model.

III. Persona Consistency

Persona consistency can be understood as having a consistent personality. For example, if the chatbot were to say that it was a boy in one message but a girl in another message, there would be a clear inconsistency. In order to give chatbots a consistent persona, we face the following research questions:

1) How do we collect personality data?
2) How should we define personality?
3) How do we incorporate a personality module into the neural network?

A. RQ1: Persona Collection

The first task for researchers trying to build chatbots with a personality is collecting a dataset of personas. Naively training on traditional dialogue datasets will teach the chatbots to assume a personality averaged over many different users, which essentially yields a model with no personality at all.

One of the most popular datasets is Persona Chat [11]. The authors use Amazon Mechanical Turk to collect $1,155$ different personalities. To create the dataset, the authors ask two crowdworkers to act out a personality as described by a few sentences. The training examples are thus the conversational exchanges between these workers. Alternatively, Shuster et al. [12] ask two people to chat about an image given a single personality trait, such as *peaceful* or *absentminded*.

Even though the number of persona based datasets is growing, the dataset collection process still remains highly burdensome. As such, authors such as Zhang et al. [11] are only able to collect about $1,000$ different personas. Without a larger dataset, training chatbots to learn a personality would remain extremely difficult. To solve this problem, Mazaré et al. [13] propose training a million different personas. To do this, the authors scrape a Reddit dump for conversations and use hand crafted rules to extract personas. Madotto et al. [14] propose a meta-learning technique that can learn personas without any manual intervention.

Future research directions: 1) Develop a chatbot that can filter out toxic personas from its training set so the model does not end up learning offensive language. 2) Develop a chatbot that can extract personas from live conversations. Such an active learning approach would significantly lower the burden of collecting persona datasets.

B. RQ2: Persona Definition

Personality is a highly abstract concept. To capture this notion, researchers have relied on different metrics. Some of them may refer to personality as simply a collection of background facts about yourself such as where you live and how old you are, while others may define personality as specific behavioral traits such as whether you are absentminded or very talkative.

Most researchers seem to focus on background facts for now [11], [13]–[15]. In these works, the authors usually encode personality as either a set of simple sentences describing the persona or a set of key value pairs, each representing a specific fact about your background.

On the other hand, Shuster et al. [12] take a slightly different approach and encode personality as a behavioral trait. The chatbot is trained to discuss an image with a given style based on the assigned behavioral trait. For example, if the chatbot is given the *absentminded* personality trait, then it may respond frequently with clarifying questions.

Future research directions: 1) Develop a chatbot that can capture multiple dimensions of personality, such as one's background and one's behavior. 2) Develop new metrics to evaluate how well the chatbot can express its personality. For example, can the other person identify the persona of the chatbot? 3) Develop a chatbot that is not only able to answer personal questions related to its personality, but also change the style and tone of its writing to match its persona.

C. RQ3: Persona Training

Training the AI to internalize its assigned personality and adapt its replies to this personality is a challenging task in and of itself. The key challenge here is detecting what information, if any, from the given personality is relevant for a particular conversational context.

One approach is to use a memory network. This is the path taken by Zhang et al [11]. The memory network stores an encoded dialogue history and encoded profile entries. A decoder can then access this memory, extract the appropriate personality related information, and produce the output text.

Another approach taken by Qian et al. [15] is to use a simple classifier to detect whether a profile fact needs to be extracted and a multi layer perception to determine what profile fact in particular should be chosen.

Shuster et al. [12] take a different approach that relies on Transformers. Three encoders are used for their model: an image encoder, a dialogue history encoder, and a style encoder. A Transformer decoder is then used to generate the actual response conditioned on these inputs.

Future research directions: 1) Develop an analysis that explains what features chatbots are learning in order to exhibit personalities. 2) Explore reinforcement learning techniques for personality generation, where the chatbot is given a reward for maintaining persona consistency for x number of turns. 3) Explore pretraining and fine tuning strategies. For example, is it possible to train a chatbot on a large corpus of a million personas quickly and then fine tune towards a specific personality later?

IV. Emotional Intelligence

Emotional intelligence is an important part of enjoyable human to human conversations. When the user is feeling down because of a bad day at work, the chatbot should be able to recognize this and provide an appropriately comforting message in reply. To address this research challenge, we propose the following research questions:

1) How should we define emotional intelligence?
2) How do we collect datasets labeled with emotions?
3) What neural network architecture is best for expressing emotion?

A. RQ1: Dimensions of Emotional Intelligence

There are various ways to characterize emotional quotient (EQ). Some researchers may define it as empathy, while others may consider it as affect or even politeness.

Rashkin et al. [16] take the empathy route. For example, given a message such as "I finally got promoted today at work," the chatbot should respond with an appropriate congratulatory message in order to demonstrate empathy. Niu et al. [17] discuss EQ in terms of politeness. For instance, if someone were to compliment the chatbot, the polite response would be to express gratitude. Ghosh et al. [18] define EQ as affect, which is a broad term that encompasses mood and personality. An important contribution of their work is that they are able to control the strength of the affective response generated by the chatbot. Note that while Ghosh et al. do not explcitly design a chatbot, their research is still highly relevant for imbuing conversational agents with EQ.

Future research directions: 1) Design a chatbot that can incorporate multiple dimensions of EQ, such as both empathy and politeness. 2) Conduct a user study examining what dimensions of EQ are most important for engaging human conversation, and how well existing chatbots measure up to these standards.

B. RQ2: Emotion Databases

Unfortunately, there are not many publicly available datasets to train emotionally intelligent chatbots with. As such, a major impediment to progress in this area is simply lack of data. However, over the past few years, researchers have been making some progress on this front.

A couple of researchers have curated their own datasets. For example, Rashkin et al. [16] describe one such effort: their dataset consists of conversational exchanges between two people, one dubbed a speaker and the other a listener. The speaker is supposed to describe a situation relating to a particular emotion while the listener is supposed to react to the speaker's explanation. No emotion labels are needed as just the dialogues are collected. Instead of using just text, Huber et al. [19] incorporate images to generate emotional responses as well. They extract a million images from Twitter in order to build their dataset.

Other researchers try to bootstrap their chatbot by using existing datasets. For example, Zhou et al. [20] train a classifier to detect emotion. They then use this classifier to automatically annotate a corpus of conversations. Ghosh et al. pretrain their model on telephone conversations first and then fine tune on other smaller datasets that feature emotions much more prominently. Niu et al. leverage parallel training of two different datasets: one dataset is designed to teach a classifier to detect politeness while another dataset is used to teach another model to generate plain dialogue.

Future research directions: 1) Develop a chatbot that can learn human emotions based on live interactions with users. 2) Develop a technique for the chatbot to automatically synthesize new emotions based on a mix of a few primary emotions. 3) Explore ways for the chatbot to leverage emotional knowledge from multiple datasets from different domains.

C. RQ3: Emotion Grounding Architecture

After collecting the necessary data, it is time to think about architectural considerations. The primary contribution will stem from how the authors embed emotion into a feature vector for the chatbot to use during training. Most architectures have a seq2seq backbone augmented with some novel modules.

Instead of just using the vanilla seq2seq model, researchers might opt for Transformers instead. This is the approach Rashkin et al. [16] take. They use a context encoder to convert the sentences representing the conversation history into a tensor and then feed this into a Transformer decoder.

In addition to the seq2seq backbone, some authors add a memory network. Zhou et al. [20] use two distinct memory components: an internal memory and an external memory. Inspired by psychology studies, the internal memory helps the chatbot keep track of dynamic changes in emotion while generating the reply. The external memory helps the chatbot select the best word to include in the reply. For example, certain replies may not need overtly emotional responses, so the chatbot can choose generic words like *hello* instead of words like *happy*.

Other researchers mix different NLP models together or add other classifiers to improve dialogue generation. For example, if the chatbot is also designed to handle image data, then the neural network would include some CNN modules as well. For example, Huber et al. [19] deploy two CNNs and an SVM to detect important features in a photo that will be fed into a traditional seq2seq model for dialogue generation. As another example, Niu et al. [17] propose using a politeness classifier or language model to augment the seq2seq backbone of their chabots. A language model essentially learns to predict the next word when given previous words in a sentence.

Future research directions: 1) Design an architecture based on adversarial training. For example, could the chatbot have a generator create replies and have a discriminator distinguish whether these replies are truly emotional or faked? 2) Design new attention mechanisms that can focus on the most salient emotions in a conversation. This is important because humans often experience multiple emotions at the same time. 3) Create neural network ensembles to learn

different emotional categories and fuse all their replies together into a more robust emotional response.

V. Evaluation Methods

Evaluating chatbots remains a very challenging problem for researchers. Humans provide the most informative feedback when it comes to rating abstract notions such as engagingness and fluency. However, human evaluation is very time consuming and costly, especially when the researchers only seek to validate an early prototype. On the other hand, existing automatic evaluation metrics cannot always capture the quality of open domain chatbots very well. To guide future research in this area, we propose the following key questions:

1) How can we develop better human evaluation methods?
2) What are some common pitfalls of automatic evaluation methods?
3) How can we develop better automatic chatbot evaluation methods based on RQ2?

A. RQ1: Human Evaluation

Human evaluation is essential for capturing complex characteristics of human speech such as emotional intelligence, commonsense, personality, etc. The standard for conducting these studies is to use Amazon Mechanical Turk to assemble crowdworkers. These crowdworkers then rate the reply generated by the chatbot on a standard point scale based on a trait like engagingness.

Unfortunately, there is not much work evaluating how robust existing human evaluation schemes are for chatbots. Researchers take basic steps to eliminate threats to validity, such as calculating agreement scores among human evaluators, but sometimes these are very low. For example, the model developed by Qian et al. [15] was evaluated by humans, but Cohen's kappa score for naturalness is only 46%. The authors themselves state that this is because naturalness is difficult to judge.

Other problems also exist. For example, are there any kinds of bias in the human evaluators? If people have high expectations for an emotionally intelligent chatbot, the disappointment may skew the ratings in a negative direction. Additionally, most studies only recruit a small group of participants, which may be problematic for establishing a universal benchmark for the research community. For example, while Huber et al. [19] present an otherwise strong evaluation, they only use 10 independent crowdworkers to validate their chatbot.

Future research directions: 1) Develop a standard human evaluation scheme for chatbots that is robust against human and statistical bias. 2) Compare human evaluation schemes from multiple chatbot papers and study common pitfalls and notable strengths. 3) Develop ways to better integrate human and automatic evaluation, as humans can be very good judges of quality but cannot process large amounts of data as much as automatic metrics can.

B. RQ2: Common Pitfalls of Automatic Evaluation

Unfortunately, human evaluation remains very costly and time consuming to perform. Thus, researchers have been looking to automatic evaluation metrics. However, automatic evaluation is not very useful if the metrics used are not accurate indicators of performance themselves.

Indeed, Liu et al. [21] pointed out some fundamental flaws with some automatic metrics. They argue that standard NLP metrics actually do a very poor job of evaluating the performance of dialogue generation systems. Take bilingual evaluation understudy (BLEU) for example. BLEU is designed to compare the differences in text between the synthesized reply and the ground truth reply. Suppose that your conversational partner asks, "How are you?" The ground truth reply is, "Good! How are you?" However, if the chatbot responded with "Are you how? Good!" we would still obtain a perfect BLEU score because all of the words in the generated response can be found in the ground truth.

In the end, the authors of this paper quantitatively evaluate some popular automatic metrics by measuring their correlation to human judgements. Their analysis concludes that these metrics demonstrate very low human correlation on a Twitter dataset and almost no correlation in a techinical Ubuntu support dataset.

Future research directions: 1) Develop a standardized procedure to test the validity of an automatic evaluation metric that goes beyond simple human correlation calculations. For example, could you calculate agreement with other similar automatic metrics? 2) Consider the security implications of certain evaluation metrics. Could an adversary fool the chatbot into producing toxic replies while maintaining the same performance based on automatic evaluation?

C. RQ3: Designing New Automatic Metrics

Ultimately, researchers need to take the lessons learned from RQ2 and develop improved automatic evaluation metrics from them. At the very least, these metrics should show good correlation with human judgements.

One approach could be to train an evaluator. Li et al. [22] propose a technique to evaluate chatbots based on how well they can fool an adversary into thinking the generated response is actually a human response. A semantic evaluator works as well [23]. The authors design this evaluator to predict how a human would rate the appropriateness of a generated response by calculating semantic similarity with the ground truth.

Yet another approach is to combine different metrics together. Ruber is an example of such an approach [24]. In this paper, one metric computes the semantic similarity between the generated response and the ground truth while the other metric is a neural network that simply predicts how relevant the generated response is to the conversation. The first metric is good at rewarding high quality responses, but neglects the fact that there could be multiple correct answers. The second metric addresses this. Additionally, Hashimoto et al. [25] propose a technique to combine both human and automatic evaluation metrics. Human metrics are very good at capturing

quality, but not diversity. Automatic evaluation metrics can capture diversity very well.

Future research directions: 1) Explore reinforcement learning approaches for evaluating chatbots. 2) Design metrics that integrate multiple notions of response quality together, such as emotional intelligence, personality, knowledge, etc.

VI. Conclusion

This survey has focused on four key research challenges for open domain generative chatbots. The first challenge concerns finding a good way to ground conversational agents in external knowledge so they can generate much more interesting responses. Many of the works in this area attempt to condition the chatbot on some specific external knowledge such as Wikipedia, but very few papers try to develop a unified model of knowledge that incorporates external facts from multiple domains. The second challenge is primarily concerned with preventing chatbots from contradicting themselves. Most papers focus on predefined character traits such as the chatbot's name, gender, etc. Future research can explore how chatbots can maintain persona consistency in other dimensions, such as maintaining a coherent opinion on a popular topic. The third challenge is focused on emotional intelligence, which is a mandatory trait for any chatbot that seeks to emulate real life human conversations. Newcomers should feel free to investigate the many unexplored dimensions of emotional intelligence such as sense of humor. The final challenge concerns evaluation methods, and is a major impediment to future advances in chatbot research. Future research can focus on integrating multiple notions of response quality such as EQ and informativeness into one unified metric for chatbot performance. Solving any of these open questions in chatbot development could pave the way for next generation chatbots that can truly act like intelligent agents.

VII. Acknowledgement

The authors would like to thank all the reviewers for their helpful comments. This project was supported in part by US National Science Foundation grant CNS-1816399. This work was also supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation and workforce development. For more information about CCI, visit cyberinitiative.org.

References

[1] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li, "Building task-oriented dialogue systems for online shopping," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017.

[4] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *Acm Sigkdd Explorations Newsletter*, 2017.

[5] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, 2020.

[6] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *arXiv preprint arXiv:1811.01241*, 2018.

[7] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[8] L. Qin, M. Galley, C. Brockett, X. Liu, X. Gao, B. Dolan, Y. Choi, and J. Gao, "Conversing by reading: Contentful neural conversation with on-demand machine reading," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019.

[9] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W. tau Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[11] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "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)*, Jul. 2018.

[12] K. Shuster, S. Humeau, A. Bordes, and J. Weston, "Image-chat: Engaging grounded conversations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020.

[13] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Oct.-Nov. 2018.

[14] A. Madotto, Z. Lin, C.-S. Wu, and P. Fung, "Personalizing dialogue agents via meta-learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5454–5459.

[15] Q. Qian, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Assigning personality/profile to a chatting machine for coherent conversation generation." in *IJCAI*, 2018, pp. 4279–4285.

[16] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019.

[17] T. Niu and M. Bansal, "Polite dialogue generation without parallel data," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 373–389, 2018.

[18] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-lm: A neural language model for customizable affective text generation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 634–642.

[19] B. Huber12, D. McDuff, C. Brockett, M. Galley, and B. Dolan, "Emotional dialogue generation using image-grounded language models," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[20] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Nov. 2016.

[22] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Sep. 2017.

[23] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic Turing test: Learning to evaluate dialogue responses," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jul. 2017.

[24] C. Tao, L. Mou, D. Zhao, and R. Yan, "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] T. Hashimoto, H. Zhang, and P. Liang, "Unifying human and statistical evaluation for natural language generation," in *ACL 2019*.