“Chitty-Chitty-Chat Bot”: Deep Learning for Conversational AI

Rui Yan[†,‡]
[†]Institute of Computer Science and Technology, Peking University
[‡]Beijing Institute of Big Data Research, Beijing, China
ruiyan@pku.edu.cn

Abstract

Conversational AI is of growing importance since it enables easy interaction interface between humans and computers. Due to its promising potential and alluring commercial values to serve as *virtual assistants* and/or *social chatbots*, major AI, NLP, and Search & Mining conferences are explicitly calling-out for contributions from conversational studies. It is an active research area and of considerable interest.

To build a conversational system with moderate intelligence is challenging, and requires abundant dialogue data and interdisciplinary techniques. Along with the Web 2.0, the massive data available greatly facilitate data-driven methods such as deep learning for human-computer conversations. In general, conversational systems can be categorized into 1) task-oriented systems which aim to help users accomplish goals in vertical domains, and 2) social chat bots which can converse seamlessly and appropriately with humans, playing the role of a chat companion. In this paper, we focus on the survey of non-task-oriented chit-chat bots.

1 Introduction

Starting from ELIZA [Weizenbaum, 1966] in 1960s, non-task-oriented conversational systems (a.k.a., chatbots) have never been so popular as in recent years. Practical applications from the industry pioneered the way. Take Microsoft “Little Bing” (also known as *XiaoIce*) as an example. The chatbot, released by Microsoft to Chinese users in 2014, has now attracted more than 100 million users in China, Japan, U.S., India, and Indonesia. For years, companies (big names or start-ups) as well as academia have paid great attention to improve conversational AI for its functional, social, and entertainment roles in real-world applications.

Generally speaking, there are two types of conversational AI. One is designed for helping people to complete particular goals, ranging from train scheduling to restaurant reservation, which is known as task-oriented conversational AI [Li *et al.*, 2017; Yan *et al.*, 2017b]. The tasks are basically established in various vertical domains and the conversational systems are tailored to these domains. The other one, on the contrary, is non-task-oriented, and is usually used for social chit-chats as chatbots. Different from task-oriented systems, chatbots aim to engage users in human-computer conversations in the open domain for entertainments and/or emotional companionship. As a result, it is easier for chatbots to go viral among end users without any specific purposes than task-oriented conversational systems. For example, until 2017, users from the five countries have finished more than 20 billion turns of conversations with XiaoIce; and on average, each conversation lasts up to 20 turns. The promising user data indicate impressive popularity of the chatbot service.

There is a large volume of literature for each type of the conversational AI respectively, either task-oriented or non-task-oriented. In contrast to the prosperity of chatbots among end users, there are no systematic introductions to approaches about how to build the conversational engines behind chatbots in the research community. In this survey paper, we present a literature review for non-task-oriented conversational AI for chit-chats. Unlike the conventional conversation systems such as ELIZA which are built all but with hand-crafted rules, recently researchers begin to develop principled and data-driven approaches to build open domain conversational systems due to the benefits from the large scale social conversation data publicly available and the rapid progress of deep learning approaches. Therefore, we believe it is useful and valuable to summarize the survey about recent progress on deep learning approaches for conversational AI of chit-chats, i.e., chat bot engines in the open domain. The community would learn the insights behind chatbots to fulfill the gap between task-oriented conversational systems and non-task-oriented ones.

This survey paper is partially based on our continuous efforts on building conversational models with deep learning approaches for chatbots. We will summarize the problem formulation and data collection for chatbots, and give an overview of state-of-the-art methods for open domain chit-chats from several aspects.[1]

2 Formulation and Data Collection

People have developed a well-defined paradigm and deploy it into most of existing conversational systems. The user

[1] Due to strict format limits, a longer version will be on the arXiv.

inputs an utterance as the **query** into the computer, the system returns a **response**. Given a query q from the human, the conversational AI learns responding models either by organizing tokens and words to synthesize new responses (i.e., generation-based systems), or by finding existing responses in a pre-collected data repository which contains appropriate candidate utterances to reply (i.e., retrieval-based systems).

A response can be obtained solely based on the input query. However, a more real scenario is that a query has contexts since a conversation session generally lasts for multiple turns. The contexts denote the previous utterances within the current conversation session.

Data. In early days, people focused on conversational systems established by rules or templates [Walker *et al.*, 2001; Williams *et al.*, 2014]. The idea is simple and such methods require no data or few data for training, while instead require a great many human efforts to create enough handcraft rules or templates to run the system. To build rule-based systems is costly. Yet a conversation goes out of scope easily. People begin to pay more attention to data-driven methods.

From human-driven conversational systems to data-driven conversational systems, the need for a much bigger amount of data is substantially increasing. Nowadays, with the prosperity of social media (e.g., microblogs), forums and other Web resources, people have conversations with each other on the Internet publicly. Although the data are often quite noisy [Shang *et al.*, 2018], it is feasible to collect a very large amount of human-to-human conversation samples [Ritter *et al.*, 2011; Wang *et al.*, 2013]. On the Internet, a user can publish an utterance visible to the public, and then receive one or more replies or comments in response to the message. There are many flexible forms to “respond” to a given message, which is exactly the nature of real conversations: various responses are all possibly appropriate. It is not difficult to collect sufficient data from social media such as microblogs, forums, QAs, etc. Researchers split the conversational data as “query”-“response” (q-r) pairs for learning to respond. Each q-r pair indicates an atomic conversation.

3 To Retrieve vs. to Generate a Response

3.1 Retrieval-based Conversational AI

With massive data available, it is intuitive to build a retrieval-based conversational system as information retrieval techniques are developing fast. Given a user input utterance as the query, the system searches for candidate responses by matching metrics. The core of retrieval-based conversational systems is formulated as a matching problem between the query utterance and the candidate responses. A typical way for matching is to measure the inner-product of two representing feature vectors for queries and candidate responses in a transformed Hilbert space. The modeling effort boils down to finding the mapping from the original inputs to the feature vectors [Lu and Li, 2013], which is known as representation learning. Wang *et al.* [2013] proposed to use a two-step retrieval technique to find appropriate responses from the massive data repository. The retrieval process consists of a fast ranking by standard TF-IDF measurement and the re-ranking process using conversation-oriented features designed with human expertise. Leuski *et al.* [2009] built systems to select the most suitable response to the query from the question-answer pairs using a statistical language model as cross-lingual information retrieval. These methods are based on shallow representations, which basically utilizes one-hot representation of words.

Most strong retrieval systems learn representations with deep neural networks (DNNs). DNNs are highly automated learning machines; they can extract underlying abstract features of data automatically by exploring multiple layers of non-linear transformation. Prevailing DNNs for sentence-level modeling include convolutional neural networks (C-NNs) and recurrent neural networks (RNNs). Due to the gradient problem in vanilla RNNs, Long-Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] and Gated Recurrent Unit (GRU) [Chung *et al.*, 2014] were proposed to address the long-term dependency issue.

A series of matching methods can be applied to short-text conversations for retrieval-based systems. Basically, these methods model sentences using convolutional [Lu and Li, 2013; Hu *et al.*, 2014] or recurrent [Palangi *et al.*, 2015; Wan *et al.*, 2016a] networks to construct abstractive representations. A series of matching metrics have been proposed for retrieval using deep neural networks. Palangi *et al.* [2015] proposed sentence matching based on vector similarities. A recursive schema is later introduced for incremental sentence modeling [Wan *et al.*, 2016b; Liu *et al.*, 2016b].

Usually, sentences are compared in a pairwise matching style via word-by-word matchings, known as sentence pair modeling [Hu *et al.*, 2014; Wan *et al.*, 2016a]. The chain-based matching is also demonstrated to be useful by mixing sentence information as a chain sequence [Li *et al.*, 2016a]. In chain-based matching, modeling the second sentence is not blind to the modeling of the first sentence. Although not all of these methods are originally designed for conversation, they are effective for short-text matching tasks and are included as strong baselines for retrieval-based conversational studies [Yan *et al.*, 2016b; 2016a; 2017a].

3.2 Generation-based Conversational AI

Another way to build a conversational system is to use language generation techniques. Higashinaka *et al.* [2014] proposed to combine language template generation with the search-based methods. Ritter *et al.* [2011] investigated the feasibility of conducting short text conversation by using statistical machine translation (SMT) techniques, learning from millions of naturally occurring conversation data in Twitter. In these approaches, a response is generated from a model, not retrieved from a repository, and thus it cannot be guaranteed to be a legitimate natural language text at all times [Yan *et al.*, 2016a].

With deep learning techniques applied, generation-based systems are greatly advanced. In general, the conversational system applies the sequence-to-sequence generation manner [Sutskever *et al.*, 2014]. A neural responding machine was proposed for single-turn conversations [Shang *et al.*, 2015]. The model was then extended to handle multi-turn conversations, trying different ways to use contexts [Sordoni *et al.*, 2015; Serban *et al.*, 2016; Tian *et al.*, 2017]. Based on the generative framework, researchers gradually introduce additional elements into generation, such as persona, knowledge and topic, etc. We will elaborate them later.

Encoding. In general, the framework consists of an encoder-decoder framework using the sequence-to-sequence model. The encoder converts a sequence of embedding inputs $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ to hidden representations $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n)$. Encoders can be implemented using LSTM or GRU units.

$$\mathbf{h}_t = \text{Encoder}(\mathbf{h}_{t-1}, \mathbf{x}_t) \tag{1}$$

Decoding. After the encoded information is obtained, the decoder takes as input a context vector $\mathbf{c}_t$ and the embedding of a previously decoded word $\mathbf{y}_{t-1}$ to update its state st using another LSTM/GRU sequence:

$$\mathbf{s}_t = \text{Decoder}(\mathbf{s}_{t-1}, [\mathbf{c}_t; \mathbf{y}_{t-1}]) \tag{2}$$

$[\mathbf{c}_t; \mathbf{y}_{t-1}]$ is the concatenation of the two vectors, serving as the input to decoder units. The context vector $\mathbf{c}_t$ is designed to dynamically attend on important information of the encoding sequence during the decoding process, namely attention mechanism [Bahdanau *et al.*, 2015]. Once the state vector $\mathbf{s}_t$ is obtained, the decoder generates a token by sampling from the output probability distribution $\mathbf{o}_t$ computed from the decoder's state $\mathbf{s}_t$ parameterized with $\mathbf{W}_o$:

$$\begin{aligned} y_t \sim \mathbf{o}_t &= p(y_t | y_1, y_2, \ldots, y_{t-1}, \mathbf{c}_t) \\ &= \text{softmax}(\mathbf{W}_o \mathbf{s}_t) \end{aligned} \tag{3}$$

3.3 System Ensemble

Retrieval-based systems represent the mainstream chatbot systems in industry for real-world applications since the responses are literally created by humans. The sentences are fluent and natural, and hence regarded to be reliable in practice. The database consisting of a large number of utterance pairs is a key to success [Leuski and Traum, 2011]. Researchers propose to augment the database with utterance pairs retrieved from plain texts [Nouri *et al.*, 2011]. However, the data repository is still the bottleneck for retrieval-based systems. Even though the retrieval repository can be large, if there is no appropriate responses for a given utterance, the system cannot "create" new appropriate candidates.

For generation-based systems, they are flexible enough to create unlimited responses given a relatively smaller vocabulary and a smaller training dataset. However, these systems have weaknesses, too: they are likely to over-generate or under-generate sentences which are not always guaranteed to be natural, fluent and legitimate [Yan *et al.*, 2016a].

A feasible solution is to utilize the advantage of both systems. Given a query utterance, an optimized system will find candidate responses from both the retrieval-based and the generation-based system, merge the candidates from both systems together, and output the best response in the merged list, namely *ensemble* [Qiu *et al.*, 2017; Song *et al.*, 2018a].

4 Contexts Are Important

There are typically two setups for conversational AI studies: 1) single-turn and 2) multi-turns. Single-turn conversation indicates, perhaps, the simplest setting where the model only takes the query utterance into account to output the response. However, most real world dialogues generally consist of multiple turns. Previous utterances before the query utterance (referred to as "*context*") could also provide useful information about the conversation and are the key to success in multi-turn conversations.

For retrieval-based systems, single-turn conversations are conducted by matching responses with the query utterances [Wang *et al.*, 2013; Yan *et al.*, 2016b]. The matching metrics mentioned in Section 3.1 are all applicable in single-turn conversations. For generation-based systems, single-turn conversations are basically established using sequence-to-sequence with attention [Shang *et al.*, 2015].

Later, studies have revealed the importance of context, and researchers proposed several context-aware conversational systems. In general, context-aware methods all utilize utterances occurred before the query utterance. If the context is too long or the topic of the conversation has shifted, *context segmentation* is required to keep the relevance context information only [Song *et al.*, 2016]. We categorize different context modeling as follows. The simplest way to incorporate contextual information is to concatenate all utterances in the contexts from head to tail as a long sequence [Zhou *et al.*, 2016]. In this way, only words are modeled as atomic units on the same hierarchy for representation learning. It is useful to incorporate context in such a non-hierarchial way through pooling and concatenation [Sordoni *et al.*, 2015].

Li et al. [2015] proposed that words and sentences shall be modeled on different hierarchies for representation learning. Context learning is then extended to hierarchical representations by distinguishing word-level and utterance-level representations. The two levels of information is combined together for retrieval-based systems [Zhou *et al.*, 2016] or for generation-based systems [Serban *et al.*, 2016; 2017].

In contrast to the direct concatenation strategy (hierarchical or non-hierarchical), comprehensive combination strategies were also investigated. It is intuitive to propose a rank-and-rerank framework to rank responses with the query utterance and then to re-rank responses with contexts [Yan *et al.*, 2016b]. Yan *et al.* [2016a] proposed a reformulation framework which incorporates context utterances with the query utterance to reformulate a list of "pseudo" query utterances. The matched response with all pseudo utterances are deemed as appropriate given the query and the contexts. Insufficiently, order information is discarded by such a reformulation framework. A sequential matching network was introduced to integrate word and utterance representations from different hierarchies and match them sequentially [Wu *et al.*, 2017]. The framework is extended to matching sequences to preserve order information [Yan and Zhao, 2018a].

5 One-to-Many Diversity

Ideally, the intelligent conversational AI should be able to output grammatical, coherent responses that are diverse and interesting. There is a unique phenomenon for human conversations: given a particular query utterance, there can be multiple different responses. These responses can be totally dissimilar to each other in terms of language styles and contents, but they are all appropriate to respond the query utterance, from different aspects. Such a phenomenon is known as "one-to-many" diversity in conversations.

Modern conversational systems are driven by data, which face the challenge in learning patterns directly from the data. In contrast to the ideal situation, however, neural conversational models in practice learn to provide trivial or non-committal responses such as "I don't know", "Me too" or "I'm OK" [Sordoni *et al.*, 2015; Serban *et al.*, 2016]. The top candidates to respond are usually generic and universal. Since these responses are broad enough to respond many query utterances, they are actually meaningless for conversations. Li *et al.* [2016b] ascribed the reason for such responses to data distribution in the conversational corpus. In their study, 0.45% of the utterances in the conversational data are "I don't know", which takes up a relatively high proportion to respond many utterances. The observation explains the high-frequency along the lines of generic and meaningless responses, and the relative sparsity of more informative alternative candidates.

It is necessary to optimize for the likelihood of candidates by lowering the probability and weights of generic responses. Penalizing universal responses intrinsically brings diverse outputs. Li et al. [2016b] propose to capture this intuition by changing the original optimization target $\log p(r|q)$ to $\log p(r|q) - \log p(r)$, which penalizes universal responses with high frequency in the corpus. The idea is straightforward yet quite effective. Another intuitive way to incorporate diversity is to re-rank the candidate responses by selecting as many diverse responses as possible into the top ranked list. It is useful to diversify the candidate response list using Maximum Marginal Relevance (MMR) [Song *et al.*, 2017] or Determinantal Point Processes (DPP) [Song *et al.*, 2018b].

For retrieval-based systems, diversity is achieved by selecting diverse candidates while for generation-based systems, it is straightforward to incorporate diversity directly into the decoding process. A diverse decoding method is proposed by diversifying the standard beam search process to generate diverse N-best lists [Li *et al.*, 2016c]. The model adds an intra-sibling ranking term to the standard beam search algorithm, favoring choosing hypotheses from diverse parents. Song *et al.* [2018b] proposed to augment the Determinantal Point Processes (DPPs) with a Diversity Net so that the decoder selects items with good diversity and quality in balance. Tao *et al.* [2018a] modeled diversity with multi-head attention.

These methods model diversity in an explicit way. It is also possible to incorporate diversity implicitly during generation. Zhou *et al.* [2017] proposed to extend *encoder-decoder* as *encoder-diverter-decoder*. The diverter indicates a latent responding mechanism to characterize content semantics or intentions. Mechanisms are learned automatically and different responses associated with different mechanisms are naturally diverse. Conditional Variational Auto-Encoder (CVAE) can capture the discourse-level diversity during encoding [Zhao *et al.*, 2017]. CVAE learns latent variables to depict a distribution over potential conversational intents and generates diverse responses accordingly [Zhao *et al.*, 2017; Serban *et al.*, 2017]. However, implicit diversity modeling lacks the interpretability of how diversity is formulated.

6 Proactive Conversational AI

A standard chatbot system presumes that only humans will take the initiative role in chit-chats, and computers need only to "respond" to the best of its capability [Li *et al.*, 2016d]. Such a process is regarded as "passive". In human-human conversations, both participants can be initiative. To this end, conversational AI should also be proactive and can introduce new content when it is necessary to be initiative in chats.

• **Proactive Suggestions in Response Retrieval.** Conversations may go stalemate when the speaker does not know what to say or how to say. The problem of initiative stalemate-breaking is raised. Existing mixed-initiative systems are typically designed for vertical domains, such as train scheduling or ticket booking with certain slots are required to be filled. Such design philosophy hardly applies to non-task-oriented, chat-style conversations. In the open domain, a variety of responses are all plausible.

Li et al. [2016d] proposed a proactive conversational system named StalemateBreaker. The system detects *when* there occurs a stalemate, and determines *what* "intriguing" contents to introduce by re-ranking candidate responses.

Given a human utterance as the input query, the mainstream conversational systems would return a response. *query suggestion* has been shown to be helpful to bring information outside of a particular user's scope in traditional information retrieval. Inspired by this intuition, a proactive conversation mode with "response ranking" as well as "next utterance suggestion" is demonstrated to improve conversation experiences in practice [Yan *et al.*, 2017a; Yan and Zhao, 2018b]. The query, responses and suggestions are jointly learned.

• **Controllable Response Generation.** Identifying appropriate responses to introduce new contents in the pre-collected data repository is quite suitable for retrieval-based conversational systems. For generation-based conversational systems, new methods are in need.

For most of the encoder-decoder model in sequence-to-sequence learning, once the model is learned, the sentences are generated autonomously without any human interference. If new contents are going to be expressed, certain controlling mechanisms should be devised. Given the designated content to introduce, a big challenge is how to inject the words to explicitly occur in the generated utterance. Mou *et al.* [2015] proposed a *backward*-and-*forward* (B/F) language modeling algorithm, which starts the sentence generation from the designated content word(s). In this way, the generated sentence is guaranteed to contain the content word. This method is applied to conversations as a sequence-to-B/F (Seq2BF) model [Mou *et al.*, 2016]. After encoding the input utterance, the neural generator decodes from the introduced content word(s) to respond the input utterance.

There are two ways to introduce contents during response generation: 1) the words *explicitly* exist in the utterance or 2) the semantics of the words are *implicitly* included in the utterance. Seq2BF belongs to the explicit way while Wen *et al.* [2015b; 2015a] proposed a semantically controlled neural generator for conversational systems by incorporating particular semantic information named as dialogue acts. The generator can be trained on unaligned data by jointly optimizing the sentence planning and surface realization components using a simple cross entropy criterion without any heuristics or handcrafting. Yao *et al.* [2017] proposed an implicit content-introducing conversational model by combining a standard decoder, a designated content decoder and a fusion decoder fused together to generate responses. The model decodes an utterance with the semantics of designated word(s) incorporated when appropriate. Compared with the "hard" way to introduce contents explicitly during generation, the implicit way is softer, more relaxed, with better flexibility.

7 Evaluation for Conversational AI

Automatic evaluation is crucial for language generation tasks, while most existing metrics evaluate generated sentences by measuring word overlap, referring to ground truth sentence(s). For example, BLEU [Papineni *et al.*, 2002] is a standard evaluation metric for the machine translation task which computes geometric mean of the precision for n-gram (n=1, 2, 3, 4). METEOR [Banerjee and Lavie, 2005] considers precision as well as recall for more comprehensive matching. For the summarization task, recall-oriented metrics like ROUGE [Lin, 2004] and pyramid [Nenkova and Passonneau, 2004] methods were proposed to evaluate the quality of summary contents.

For retrieval-based conversational systems, traditional information retrieval evaluation metrics such as precision@n, mean average precision (MAP) and normalized Discounted Cumulative Gain (nDCG) are applicable. For generation-based conversational systems, since there is no specific evaluation measurement for dialogues, metrics for machine translation (BLEU) and/or summarization (ROUGE) are "borrowed" in the majority of conversational studies to evaluate the quality of responses.

Unfortunately, conversational models are not similar to translation or summarization models. With the unique characteristics of context information and one-to-many diversity, it is assumed that simply measuring the word overlap between a candidate and a ground truth reference is insufficient. The assumption concurs with the observation in [Liu *et al.*, 2016a]. Researchers conduct extensive empirical experiments and show weak correlation of existing metrics (e.g., BLEU, ROUGE, and METEOR) with human judgements for conversational systems.

Automatic evaluation metrics can be divided into non-learnable and learnable approaches. Non-learnable metrics (e.g., BLEU and ROUGE), which typically measure the quality of generated sentences by heuristics, are already demonstrated to be insufficient to evaluate conversations [Liu *et al.*, 2016a]. Another feasible solution is to devise learnable approaches. Very recently, Lowe *et al.* [2017] proposed a neural network-based learning metric for conversation evaluations. The model learns to predict a score of a response given the query utterance (previous user-issued utterance) as well as the ground-truth response. The model requires human annotated scores to supervise model training. Later, a referenced metric and unreferenced metric blended evaluation routine for open-domain conversational systems was proposed and this model requires no human scoring [Tao *et al.*, 2018b]. The metric consists of 1) a referenced part to measure the overlap between the system response and the ground truth, and 2) an unreferenced part to measure the correlation between the system response and the query utterance. A good response can either resemble the ground truth well or be closely related to the query utterance. The learnable evaluation metric is next extended for multilingual adaption [Tong *et al.*, 2018].

8 Conclusion Remarks

We have witnessed a rapid surge of conversational studies recently, especially the chit-chat research in the open domain. As to our research team, we are continuously working on the exciting and challenging parts of conversational systems.

Conversational AI is catching on fire: academic conferences especially add new research tracks for conversational studies and attract unexpected growth in the number of submissions to these tracks; companies from industry are making great efforts to develop conversational products. We are entering the AI era in which large-scale big data become more easily available and learning techniques become more powerful. We may stand at the entrance of future success in more advanced conversational systems (social chatbots and/or virtual assistants). Although we still face bottlenecks and obstacles to improve human-computer conversations, it is optimistic about the future of conversational AI when more efforts are devoted into this research area.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61672058). Rui Yan was sponsored by CCF-Tencent Open Research Fund and Microsoft Research Asia (MSRA) Collaborative Research Program.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR'15*, 2015.

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

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[Higashinaka and others, 2014] Ryuichiro Higashinaka et al. Towards an open domain conversational system fully based on natural language processing. In *COLING'14*, pages 928–939, 2014.

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

[Hu *et al.*, 2014] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *NIPS'14*, pages 2042–2050, 2014.

[Leuski and Traum, 2011] Anton Leuski and David Traum. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56, 2011.

[Leuski *et al.*, 2009] Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. Building effective question answering characters. In *SIGDIAL'09*, pages 18–27, 2009.

[Li *et al.*, 2015] Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *ACL'15*, pages 1106–1115, 2015.

[Li *et al.*, 2016a] Chaozhuo Li, Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. Detecting context dependent messages in a conversational environment. In *COLING'16*, pages 1990–1999, 2016.

[Li *et al.*, 2016b] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL'16*, pages 110–119, 2016.

[Li *et al.*, 2016c] Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*, 2016.

[Li *et al.*, 2016d] Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *IJCAI'16*, pages 2845–2851, 2016.

[Li *et al.*, 2017] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-end task-completion neural dialogue systems. In *IJCNLP'17*, pages 733–743, 2017.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[Liu *et al.*, 2016a] Chia-Wei Liu, Ryan Lowe, et al. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP'16*, pages 2122–2132, 2016.

[Liu *et al.*, 2016b] Pengfei Liu, Xipeng Qiu, Jifan Chen, and Xuanjing Huang. Deep fusion lstms for text semantic matching. In *ACL'16*, pages 1034–1043, 2016.

[Lowe *et al.*, 2017] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL'17*, pages 1116–1126, 2017.

[Lu and Li, 2013] Zhengdong Lu and Hang Li. A deep architecture for matching short texts. In *NIPS'13*, pages 1367–1375, 2013.

[Mou *et al.*, 2015] Lili Mou, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Backward and forward language modeling for constrained sentence generation. *arXiv preprint arXiv:1512.06612*, 2015.

[Mou *et al.*, 2016] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING'16*, pages 3349–3358, 2016.

[Nenkova and Passonneau, 2004] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL'04*, 2004.

[Nouri *et al.*, 2011] Elnaz Nouri, Ron Artstein, Anton Leuski, and David R Traum. Augmenting conversational characters with generated question-answer pairs. In *AAAI Fall Symposium: Question Generation*, 2011.

[Palangi *et al.*, 2015] Hamid Palangi, Li Deng, et al. Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval. *arXiv:1502.06922*, 2015.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL'02*, pages 311–318, 2002.

[Qiu *et al.*, 2017] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In *ACL'17*, pages 498–503, 2017.

[Ritter *et al.*, 2011] Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In *EMNLP'11*, pages 583–593, 2011.

[Serban *et al.*, 2016] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI'16*, pages 3776–3783, 2016.

[Serban *et al.*, 2017] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI'17*, pages 3295–3301, 2017.

[Shang *et al.*, 2015] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *ACL-IJCNLP'15*, pages 1577–1586, 2015.

[Shang *et al.*, 2018] Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. Learning to converse with noisy data: Generation with calibration. In *IJCAI'18*, 2018.

[Song *et al.*, 2016] Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. Dialogue session segmentation by embedding-enhanced texttiling. *INTERSPEECH'16*, pages 2706–2710, 2016.

[Song *et al.*, 2017] Yiping Song, Zhiliang Tian, Dongyan Zhao, Ming Zhang, and Rui Yan. Diversifying neural conversation model with maximal marginal relevance. In *IJCNLP'17*, pages 169–174, 2017.

[Song *et al.*, 2018a] Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *IJCAI'18*, 2018.

[Song *et al.*, 2018b] Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. Towards a neural conversation model with diversity net using determinantal point processes. In *AAAI'18*, 2018.

[Sordoni *et al.*, 2015] Alessandro Sordoni, Michel Galley, Michael Auli, et al. A neural network approach to context-sensitive generation of conversational responses. In *NAACL'15*, pages 196–205, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *NIPS'14*, pages 3104–3112, 2014.

[Tao *et al.*, 2018a] Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI'18*, 2018.

[Tao *et al.*, 2018b] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI'18*, 2018.

[Tian *et al.*, 2017] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. How to make contexts more useful? an empirical study to context-aware neural conversation models. In *ACL'17*, pages 231–236, 2017.

[Tong *et al.*, 2018] Xiaowei Tong, Zhenxin Fu, Mingyue Shang, Dongyan Zhao, and Rui Yan. One "ruler" for all languages: Multi-lingual dialogue evaluation with adversarial multi-task learning. In *IJCAI'18*, 2018.

[Walker *et al.*, 2001] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *ACL'01*, pages 515–522, 2001.

[Wan *et al.*, 2016a] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *AAAI'16*, pages 2835–2841, 2016.

[Wan *et al.*, 2016b] Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. Match-srnn: Modeling the recursive matching structure with spatial rnn. In *IJCAI'16*, pages 2922–2928, 2016.

[Wang *et al.*, 2013] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. A dataset for research on short-text conversations. In *EMNLP'13*, pages 935–945, 2013.

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

[Wen *et al.*, 2015a] Tsung-Hsien Wen, Milica Gašic, Dongho Kim, Nikola Mrkšic, Steve Young, et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *SIGDIAL'15*, page 275, 2015.

[Wen *et al.*, 2015b] Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP'15*, pages 1711–1721, 2015.

[Williams *et al.*, 2014] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *SIGDIAL'13*, pages 404–413, 2014.

[Wu *et al.*, 2017] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *ACL'17*, 2017.

[Yan and Zhao, 2018a] Rui Yan and Dongyan Zhao. Coupled context modeling for deep chit-chat: Towards conversations between human and computer. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[Yan and Zhao, 2018b] Rui Yan and Dongyan Zhao. Smarter response with proactive suggestion: A new generative neural conversation paradigm. In *IJCAI'18*, 2018.

[Yan *et al.*, 2016a] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR'16*, pages 55–64, 2016.

[Yan *et al.*, 2016b] Rui Yan, Yiping Song, Xiangyang Zhou, and Hua Wu. Shall i be your chat companion?: Towards an online human-computer conversation system. In *CIKM'16*, pages 649–658, 2016.

[Yan *et al.*, 2017a] Rui Yan, Dongyan Zhao, and Weinan E. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *SIGIR'17*, pages 685–694, 2017.

[Yan *et al.*, 2017b] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. Building task-oriented dialogue systems for online shopping. In *AAAI'17*, pages 4618–4626, 2017.

[Yao *et al.*, 2017] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP'17*, pages 2190–2199, 2017.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL'17*, pages 654–664, 2017.

[Zhou *et al.*, 2016] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view response selection for human-computer conversation. In *EMNLP'16*, pages 372–381, 2016.

[Zhou *et al.*, 2017] Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. Mechanism-aware neural machine for dialogue response generation. In *AAAI'17*, pages 3400–3407, 2017.