A Survey on Dialogue Systems:
Recent Advances and New Frontiers

Hongshen Chen[†], Xiaorui Liu[‡], Dawei Yin[†], and Jiliang Tang[‡]
[†]Data Science Lab, JD.com
[‡]Data Science and Engineering Lab, Michigan State University
chenhongshen@jd.com, yindawei@acm.org,{xiaorui,tangjili}@msu.edu

ABSTRACT

Dialogue systems have attracted more and more attention. Recent advances on dialogue systems are overwhelmingly contributed by deep learning techniques, which have been employed to enhance a wide range of big data applications such as computer vision, natural language processing, and recommender systems. For dialogue systems, deep learning can leverage a massive amount of data to learn meaningful feature representations and response generation strategies, while requiring a minimum amount of hand-crafting. In this article, we give an overview to these recent advances on dialogue systems from various perspectives and discuss some possible research directions. In particular, we generally divide existing dialogue systems into task-oriented and non-task-oriented models, then detail how deep learning techniques help them with representative algorithms and finally discuss some appealing research directions that can bring the dialogue system research into a new frontier.

1. INTRODUCTION

To have a virtual assistant or a chat companion system with adequate intelligence has seemed illusive, and might only exist in Sci-Fi movies for a long time. Recently, human-computer conversation has attracted increasing attention due to its promising potentials and alluring commercial values. With the development of big data and deep learning techniques, the goal of creating an automatic human-computer conversation system, as our personal assistant or chat companion, is no longer an illusion. On the one hand, nowadays we can easily access "big data" for conversations on the Web and we might be able to learn how to respond and what to respond given (almost) any inputs, which greatly allows us to build data-driven, open-domain conversation systems between humans and computers. On the other hand, deep learning techniques have been proven to be effective in capturing complex patterns in big data and have powered numerous research fields such as computer vision, natural language processing and recommender systems. Hence, a large body of literature has emerged to leverage a massive amount of data via deep learning to advance dialogue systems in many perspectives.

According to the applications, dialogue systems can be roughly categorized into two groups – (1) task-oriented systems and (2) non-task-oriented systems (also known as chat bots). Task-oriented systems aim to assist the user to complete
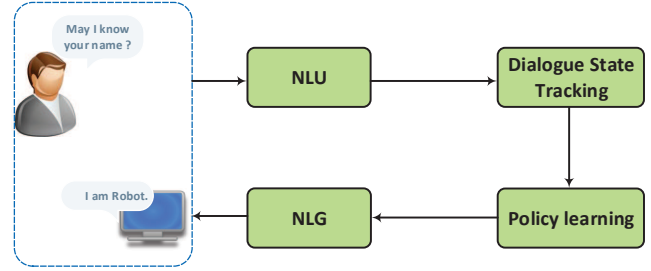

Figure 1: Traditional Pipeline for Task-oriented Systems.

certain tasks (e.g. finding products, and booking accommodations and restaurants). The widely applied approaches to task-oriented systems are to treat the dialogue response as a pipeline as shown in Figure 1. The systems first understand the message given by human, represent it as a internal state, then take some actions according to the policy with respect to the dialogue state, and finally the action is transformed to its surface form as a natural language. Though language understanding is processed by statistical models, most deployed dialogue systems still use manual features or hand crafted rules for the state and action space representations, intent detection, and slot filling. This not only makes it expensive and time-consuming to deploy a real dialogue system, but also limits its usage to other domains. Recently, many algorithms based on deep learning have been developed to alleviate those problems by learning feature representations in a high dimensional distributed fashion and achieve remarkable improvements in these aspects. In addition, there are attempts to build end-to-end task-oriented dialogue systems, which can expand the state space representation in the traditional pipeline systems and help generalize dialogues outside the annotated task-specific corpora.

Non-task-oriented systems interact with human to provide reasonable responses and entertainment. Typically, they focus on conversing with human on open domains. Though non-task-oriented systems seem to perform chit-chat, it dominates in many real word applications. As revealed in [111], nearly 80% utterances are chi-chat messages in the online shopping scenario and handling those queries is closely related to user experiences. In general, two major approaches have been developed for non-task-oriented systems – (1) generative methods such as sequence-to-sequence models, which generate proper responses during the conversation; and (2) retrieval-based methods, which learn to select responses from the current conversation from a repository.

Sentence	show	restaurant	at	New	York	tomorrow
Slots	O	O	O	B-desti	I-desti	B-date
Intent	Find Restaurant					
Domain	Order					

Table 1: An Illustrative Example of Natural Language Representation.

The recent development of big data and deep learning techniques has greatly advanced both task-oriented and non-oriented dialogue systems, which has encouraged a huge amount of deep learning based researches in dialogue systems. In this article, we aim to (1) give an overview about dialogue systems especially recent advances from deep learning; and (2) discuss possible research directions. The remaining of the article is organized as follows. We review task-oriented dialogue systems including pipeline and end-to-end methods in Section 2. In Section 3, we first introduce neural generative methods including popular models and hot research topics; and then detail the retrieval-based methods. In Section 4, we conclude the work with discussions on some research directions.

2. TASK-ORIENTED DIALOGUE SYSTEMS

Task-oriented dialogue systems have been an important branch of spoken dialogue systems. In this section, we will review pipeline and end-to-end methods for task-oriented dialogue systems.

2.1 Pipeline Methods

The typical structure of a pipeline based task-oriented dialogue system is demonstrated in Figure 1. It consists of four key components:

- Language understanding. It is known as natural language understanding (NLU), which parses the user utterance into predefined semantic slots.

- Dialogue state tracker. It manages the input of each turn along with the dialogue history and outputs the current dialogue state.

- Dialogue policy learning. It learns the next action based on current dialogue state.

- Natural language generation (NLG). It maps the selected action to its surface and generates the response.

In the following subsections, we will give more details about each component with the state-of-the art algorithms.

2.1.1 Language Understanding

Given an utterance, natural language understanding maps it into semantic slots. The slots are pre-defined according to different scenarios. Table 1 illustrates an example of natural language representation, where "New York" is the location specified as slot values, and the domain and intent are also specified, respectively. Typically, there are two types of representations. One is the utterance level category, such as the user's intent and the utterance category. The other is the word-level information extraction such as named entity recognition and slot filling.
An intent detection is performed to detect the intent of a user. It classifies the utterance into one of the pre-defined intents. Deep learning techniques have been successfully

applied in intent detection [15; 84; 112]. In particular, [25] used convolutional neural networks (CNN) to extract query vector representations as features for query classification. The CNN-based classification framework also resembled [29] and [74]. Similar approaches are also utilized in category or domain classification.

Slot filling is another challenging problem for spoken language understanding. Unlike intent detection, slot filling is usually defined as a sequence labeling problem, where words in the sentence are assigned with semantic labels. The input is the sentence consisting of a sequence of words, and the output is a sequence of slot/concept IDs, one for each word. [17] and [15] used deep belief networks (DBNs), and achieved superior results compared to CRF baselines. [51; 115; 66; 113] applied RNN for slot filling. The semantic representation generated by NLU is further processed by the dialogue management component. A typical dialogue management component includes two stages – dialogue state tracking and policy learning.

2.1.2 Dialogue State Tracking

Tracking dialogue states is the core component to ensure a robust manner in dialog systems. It estimates the users goal at every turn of the dialogue. A dialogue state H_t denotes the representation of the dialogue session till time t. This classic state structure is commonly called slot filling or semantic frame. The traditional methods, which have been widely used in most commercial implementations, often adopt hand-crafted rules to select the most likely result [23]. However, these rule-based systems are prone to frequent errors as the most likely result is not always the desired one [101].
A statistical dialog system maintains a distribution over multiple hypotheses of the true dialog state, facing with noisy conditions and ambiguity [117]. In Dialog State Tracking Challenge (DSTC) [100; 99], the results are in the form of a probability distribution over each slot for each turn. A variety of statistical approaches, including robust sets of hand-crafted rules [93], conditional random fields [36; 35; 63], maximum entropy models [98] and web-style ranking [101] have emerged in Dialog State Tracking Challenge (DSTC) shared tasks.
Recently, [26] introduced deep learning in belief tracking. It used a sliding window to output a sequence of probability distributions over an arbitrary number of possible values. Though it was trained in one domain, it can be easily transferred to new domains. [58] developed multi-domain RNN dialog state tracking models. It first used all the data available to train a very general belief tracking model, and then specialized the general model for each domain to learn the domain-specific behavior. [59] proposed a neural belief tracker (NBT) to detect the slot-value pairs. It took the system dialogue acts preceding the user input, the user utterance itself, and a single candidate slot-value pair which it needs to make a decision about, as input, and then iterated over all candidate slot-value pairs to determine which ones have just been expressed by the user.

2.1.3 Policy learning

Conditioned on the state representation from the state tracker, the policy learning is to generate the next available system action. Either supervised learning or reinforcement learning can be used to optimize policy learning [14]. Typically,

a rule-based agent is employed to warm-start the system [111]. Then, supervised learning is conducted on the actions generated by the rules. In online shopping scenario, if the dialogue state is "Recommendation", then the "Recommendation" action is triggered, and the system will retrieve products from the product database. If the state is "Comparison", then the system will compare target products/brands[111]. The dialogue policy can be further trained end-to-end with reinforcement learning to lead the system making policies toward the final performance. [14] applied deep reinforcement learning on strategic conversation that simultaneously learned the feature representation and dialogue policy, the system outperformed several baselines including random, rule-based, and supervised-based methods.

2.1.4 Natural Language Generation

The natural language generation component converts an abstract dialogue action into natural language surface utterances. As noticed in [78], a good generator usually relies on several factors: adequacy, fluency, readability, and variation. Conventional approaches to NLG typically perform sentence planning. It maps input semantic symbols into the intermediary form representing the utterance such as tree-like or template structures, and then converts the intermediate structure into the final response through the surface realization [90; 79].

[94] and [95] introduced neural network-based (NN) approaches to NLG with a LSTM-based structure similar with RNNLM [52]. The dialogue act type and its slot-value pairs are transformed into a 1-hot control vector and is given as the additional input, which ensures that the generated utterance represents the intended meaning. [94] used a forward RNN generator together with a CNN reranker, and backward RNN reranker. All the sub-modules are jointly optimized to generate utterances conditioned by the required dialogue act. To address the slot information omitting and duplicating problems in surface realization, [95] used an additional control cell to gate the dialogue act. [83] extended this approach by gating the input token vector of LSTM with the dialogue act. It was then extended to the multi-domain setting by multiple adaptation steps [96]. [123] adopted an encoder-decoder LSTM-based structure to incorporate the question information, semantic slot values, and dialogue act type to generate correct answers. It used the attention mechanism to attend to the key information conditioned on the current decoding state of the decoder. Encoding the dialogue act type embedding, the neural network-based model is able to generate variant answers in response to different act types. [20] also presented a natural language generator based on the sequence-to-sequence approach that can be trained to produce natural language strings as well as deep syntax dependency trees from input dialogue acts. It was then extended with the preceding user utterance and responses [19]. It enabled the model entraining (adapting) to users ways of speaking, which provides contextually appropriate responses.

2.2 End-to-End Methods

Despite a lot of domain-specific handcrafting in traditional task oriented dialogue systems, which are diffcult to adapt to new domains [7], [120] further noted that, the conventional pipeline of task-oriented dialogue systems has two main limitations. One is the credit assignment problem,

where the end user's feedback is hard to be propagated to each upstream module. The second issue is process interdependence. The input of a component is dependent on the output of another component. When adapting one component to new environment or retrained with new data, all the other components need to be adapted accordingly to ensure a global optimization. Slots and features might change accordingly. This process requires significant human efforts.

With the advance of end-to-end neural generative models in recent years, many attempts have been made to construct an end-to-end trainable framework for task-oriented dialogue systems. Note that more details about neural generative models will be discussed when we introduce the non-task-oriented systems. Instead of the traditional pipeline, the end-to-end model uses a single module and interacts with structured external databases. [97] and [7] introduced a network-based end-to-end trainable task-oriented dialogue system, which treated dialogue system learning as the problem of learning a mapping from dialogue histories to system responses, and applied an encoder-decoder model to train the whole system. However, the system is trained in a supervised fashion – not only does it require a lot of training data, but it may also fail to find a good policy robustly due to the lack of exploration of dialogue control in the training data. [120] first presented an end-to-end reinforcement learning approach to jointly train dialogue state tracking and policy learning in the dialogue management in order to optimize the system actions more robustly. In the conversation, the agent asks the user a series of Yes/No questions to find the correct answer. This approach was shown to be promising when applied to the task-oriented dialogue problem of guessing the famous people users think of. [45] trained the end-to-end system as a task completion neural dialogue system, where its final goal is to complete a task, such as movie-ticket booking.

Task-oriented systems usually need to query outside knowledge base. Previous systems achieved this by issuing a symbolic query to the knowledge base to retrieve entries based on their attributes, where semantic parsing on the input is performed to construct a symbolic query representing the beliefs of the agent about the user goal[97; 103; 45]. This approach has two drawbacks: (1) the retrieved results do not carry any information about uncertainty in semantic parsing, and (2) the retrieval operation is non differentiable, and hence the parser and dialog policy are trained separately. This makes online end-to-end learning from user feedback difficult once the system is deployed. [21] augmented existing recurrent network architectures with a differentiable attention-based key-value retrieval mechanism over the entries of a knowledge base, which is inspired by key-value memory networks[54]. [18] replaced symbolic queries with an induced "soft" posterior distribution over the knowledge base that indicates which entities the user is interested in. Integrating the soft retrieval process with a reinforcement learner. [102] combined an RNN with domain-specific knowledge encoded as software and system action templates.

3. NON-TASK-ORIENTED DIALOGUE SYSTEM

Unlike task-oriented dialogue systems, which aim to complete specific tasks for user, non-task-oriented dialogue systems (also known as chatbots) focus on conversing with hu-

Decoder

y_1 y_2 y_3 $y_{T`}$

Context
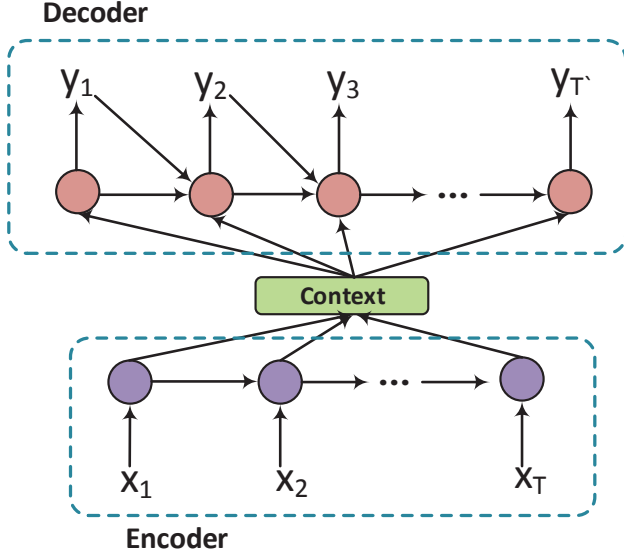
x_1 x_2 x_T

Encoder

Figure 2: An Illustration of the Encoder-Decoder Model.

man on open domains [64]. In general, chat bots are implemented either by generative methods or retrieval-based methods. Generative models are able to generate more proper responses that could have never appeared in the corpus, while retrieval-based models enjoy the advantage of informative and fluent responses[30], because they select a proper response for the current conversation from a repository with response selection algorithms. In the following sections, we will first dive into the neural generative models, one of the most popular research topics in recent years, and discuss their drawbacks and possible improvements. Then, we introduce recent advances of deep learning in retrieval based models.

3.1 Neural Generative Models

Nowadays, a large amount of conversational exchanges is available in social media websites such as Twitter and Reddit, which raise the prospect of building data-driven models. [64] proposed a generative probabilistic model, which is based on phrase-based Statistical Machine Translation [118], to model conversations on micro-blogging. It viewed the response generation problem as a translation problem, where a post needs to be translated into a response. However, generating responses was found to be considerably more difficult than translating between languages. It is likely due to the wide range of plausible responses and the lack of phrase alignment between the post and the response. The success of applying deep learning in machine translation, namely Neural Machine Translation, spurs the enthusiasm of researches in neural generative dialogue systems.

In the following subsections, we first introduce the sequence-to-sequence models, the foundation of neural generative models. Then, we discuss hot research topics in the direction including incorporating dialogue context, improving the response diversity, modeling topics and personalities, leveraging outside knowledge base, the interactive learning and evaluation.

3.1.1 Sequence-to-Sequence Models

Given a source sequence (*message*) $X = (x_1, x_2, ..., x_T)$ consisting of T words and a target sequence (*response*) $Y = (y_1, y_2, ..., y_{T'})$ of length T', the model maximizes the generation probability of Y conditioned on X: $p(y_1, ..., y_{T'} | x_1, ..., x_T)$. Specifically, a sequence-to-sequence model (or Seq2Seq) is in an encoder-decoder structure. Figure 2 is a general illustration of such structure. The encoder reads X word by word and represents it as a context vector c through a recurrent neural network (RNN), and then the decoder estimates the generation probability of Y with c as the input. The encoder RNN calculates the context vector $\mathbf{c}$ by

$$\mathbf{h}_t = f(x_t, \mathbf{h}_{t-1}),$$

where $\mathbf{h}_t$ is the hidden state at time step t, f is a non-linear function such as long-short term memory unit (LSTM) [27] and gated recurrent unit (GRU) [12], and c is the hidden state corresponding to the last word $\mathbf{h}_T$. The decoder is a standard RNN language model with an additional conditional context vector $\mathbf{c}$. The probability distribution $\mathbf{p}_t$ of candidate words at every time t is calculated as

$$\mathbf{s}_t = f(y_{t-1}, \mathbf{s}_{t-1}, \mathbf{c}),$$
$$\mathbf{p}_t = softmax(\mathbf{s}_t, y_{t-1}),$$

where $\mathbf{s}_t$ is the hidden state of the decoder RNN at time t and y_{t1} is the word at time $t1$ in the response sequence. The objective function of Seq2Seq is defined as:

$$p((y_1, ..., y_{T'} | x_1, ..., x_T) = p(y_1|\mathbf{c}) \prod_{t=2}^{T'} p(y_t | \mathbf{c}, y_1, ..., y_{t-1}).$$

[5] then improved the performance by the attention mechanism, where each word in Y is conditioned on different context vector $\mathbf{c}$, with the observation that each word in Y may relate to different parts in x. In particular, y_i corresponds to a context vector $\mathbf{c}_i$, and $\mathbf{c}_i$ is a weighted average of the encoder hidden states $\mathbf{h}_1, ..., \mathbf{h}_T$:

$$\mathbf{c}_i = \Sigma_{j=1}^{T} \alpha_{ij} \mathbf{h}_j,$$

where $\alpha_{i,j}$ is computed by:

$$\alpha = \frac{exp(e_{ij})}{\Sigma_{k=1}^{T} exp(e_{ik})},$$
$$e_{ij} = g(\mathbf{s}_{t-1}, \mathbf{h}_j),$$

where g is a multilayer perceptron.

[71] applied the recurrent neural network encoder-decoder framework [12] to generate responses on Twitter-style micro-blogging websites, while [87] utilized a similar model described in [50]. In general, these models utilize neural networks to represent dialogue histories and to generate appropriate responses. Such models are able to leverage a large amount of data in order to learn meaningful natural language representations and generation strategies, while requiring a minimum amount of domain knowledge and hand-crafting.

3.1.2 Dialogue Context

The ability to take into account previous utterances is key to building dialog systems that can keep conversations active and engaging. [77] addressed the challenge of the context sensitive response generation by representing the whole dialogue history (including the current message) with continuous representations or embeddings of words and phrases.

The response is generated as RNN language model [52], the same as the decoder in [12]. [68] used hierarchical models, first capturing the meaning of individual utterances and then integrating them as discourses. [109] extended the hierarchical structure with the attention mechanism [5] to attend to important parts within and among utterances with word level attention and utterance level attention, respectively. [82] conducted a systematic comparison among existing methods (including non-hierarchical and hierarchical models) and proposed a variant that weights the context with respect to context-query relevance. It found that (1) hierarchical RNNs generally outperform non-hierarchical ones, and (2) with context information, neural networks tend to generate longer, more meaningful and diverse replies.

3.1.3   Response Diversity

A challenging problem in current sequence-to-sequence dialogue systems is that they tend to generate trivial or non-committal, universally relevant responses with little meaning, which are often involving high frequency phrases along the lines of *I dont know* or *Im OK* [77; 87; 68].

This behavior can be ascribed to the relative high frequency of generic responses like *I dont know* in conversational datasets, in contrast with the relative sparsity of more informative alternative responses. One promising approach to alleviate such challenge is to find a better objective function. [38] pointed out that neural models assign high probability to "safe responses when optimizing the likelihood of outputs given inputs. They used a Maximum Mutual Information (MMI), which was first introduced in speech recognition [6; 9], as an optimization objective. It measured the mutual dependence between inputs and outputs, where it took into consideration the inverse dependency of responses on messages. [114] incorporated inverse document frequency (IDF) [65] into the training process to measure the response diversity.

Some researches realized that the decoding process is another source of redundant candidate responses. [86],[72] and [42] recognized that the beam-search, an approximate inference algorithm to decode output sequences for neural sequence models, lacks diversity when generating candidates in the beam. [86] augmented the beam-search objective with a dissimilarity term that measured the diversity between candidate sequences. [72] introduced a stochastic beam-search procedure, while [42] added an additional term for beam search scoring to penalize the siblings–expansions of the same parent node in the search. [38; 77; 72] further performed a re-ranking step with global features to avoid generating dull or generic responses. [57] conjectured that not only the problem lies in the objective of decoding and response frequency, but also the message itself may lack sufficient information for the replay. It proposed to use pointwise mutual information (PMI) to predict a noun as a keyword, reflecting the main gist of the reply, and then generates a reply containing the given keyword.

Another series of works have focused on generating more diverse outputs by introducing a stochastic latent variable. They demonstrated that natural dialogue is not deterministic – replies for a same message may vary from person to person. However, current response is sampled from a deterministic encoder-decoder model. By incorporating a latent variable, these models have the advantage that, at the generation time, they can sample a response from the distribution

by first sampling an assignment of the latent variables, and then decoding deterministically. [11] presented a latent variable model for one-shot dialogue response. The model contained a stochastic component z in the decoder $P(Y|z, X)$, where z is computed following the variational auto-encoder framework[34; 33; 75]. [69] introduced latent variables to the hierarchical dialogue modeling framework [68]. The latent variable is designed to make high-level decisions like topic or sentiment. [73] conditioned the latent variable on explicit attributes to make the latent variable more interpretable. These attributes can be either manually assigned or automatically detected such topics, and personality.

3.1.4   Topic and Personality

Learning the inherent attributes of dialogues explicitly is another way to improve the diversity of dialogues and ensures the consistency. Among different attributes, topic and personality are widely explored.

[108] noticed that people often associate their dialogues with topically related concepts and create their responses according to these concepts. They used Twitter LDA model to get the topic of the input, fed topic information and input representation into a joint attention module and generated a topic-related response. A small improvement in decoder had achieved a better result in [107]. [13] made a more thorough generalization of the problem. They classified each utterance in the dialogue into one domain, and generated the domain and content of next utterance accordingly.

[67] jointly modeled the high-level coarse tokens sequence and the dialogue generation explicitly, where the coarse tokens sequence aims to exploit high-level semantics. They exploited nouns and activity-entity for the coarse sequence representation.

[122] added emotion embedding into a generative model and achieved good performance in perplexity. [3] enhanced the model of producing emotionally rich responses from three aspects: incorporating cognitive engineered affective word embeddings, augmenting the loss objective with an affect-constrained objective function, and injecting affective dissimilarity in diverse beam-search inference procedure[86]. [61] gave the system an identity with profile so that the system can answer personalized question consistently. [39] further took the information of addressee into consideration to create a more realistic chatbot.

Since the training data comes from different speakers with inconsistency, [119] proposed a two-phase training approach which initialized the model using large scale data and then fine-tuned the model to generate personalized response. [55] used transfer reinforcement learning to eliminate inconsistencies.

3.1.5   Outside Knowledge Base

An important distinction between human conversation and dialogue system is whether it is combined with reality. Incorporating an outside Knowledge Base (KB) is a promising approach to bridge the gap of background knowledge between a dialogue system and human.

Memory network is a classic method dealing with question answering tasks with knowledge base. Thus, it is quite straightforward to apply it in dialogue generation. [22] made attempts on top of this and has achieved good performance in open-domain conversations. [88] also worked on open-domain conversations with background knowledge by cou-

pling CNN embedding and RNN embedding into multimodal space and made progress in perplexity. A similar task is to generate an answer for a question according to outside knowledge. Unlike the general method of tuple retrieval in knowledge base, [116] used words from knowledge base together with common words in generation process. Empirical studies demonstrated that the proposed model was capable of generating natural and right answers to the questions by referring to the facts in the knowledge base.

3.1.6 Interactive Dialogue learning

Learning through interaction is one of the ultimate goals of dialogue systems. [43] simulated dialogues between two virtual agents. They defined simple heuristic approximations to rewards that characterize good conversations: good conversations are forward-looking [1] or interactive (a turn suggests a following turn), informative, and coherent. The parameters of an encoder-decoder RNN defined a policy over an infinite action space consisting of all possible utterances. The agent learned a policy by optimizing the long-term developer-defined reward from ongoing dialogue simulations using policy gradient methods [104], rather than the MLE objective defined in the standard SEQ2SEQ models. [40] further attempted to improve the bot's ability to learn from interaction. By using policy learning and forward prediction on both textual and numerical feedback, the model can improve itself by interacting with human in a (semi-)online way. Instead of using hand-crafted reward functions for online reinforcement learning, [4] performed online human in-the-loop active learning by repeatedly letting human select one of the K responses, generated by an offline supervised pretrained dialogue agent, as the 'best' response, and then respond to the selected response. The network is also trained with the joint corss-entropy loss function.

As most human respondents may ask for clarification or hints when not confident about the answer, it is natural to make the bot owning such a capability. [41] defined three situations where the bot has problems in answering a question. Compared the experimental results of not using a asking-question way, this method made great improvement in some scenarios. [37] explored the task on negotiation dialogues. As conventional sequence-to-sequence models simulate human dialogues but fail to optimize a specific goal, this work took a goal-oriented training and decoding approach and demonstrated a worthwhile perspective.

3.1.7 Evaluation

Evaluating the quality of the generated response is an important aspect of dialogue response generation systems[46]. Task-oriented dialogue system can be evaluated based on human-generated supervised signals, such as a task completion test or a user satisfaction score[89; 56; 31], however, automatically evaluating the quality of generated responses for non-task-oriented dialogue systems remains an open question due to the high response diversity [2]. Despite the fact that word overlap metrics such as BLEU, METEOR, and ROUGE have been widely used to evaluate the generated responses, [46] found that those metrics, as well as word embedding metrics derived from word embedding models such as Word2Vec[53] have either weak or no correlation with human judgements, although word embedding metrics are able to significantly distinguish between baselines and state-of-the-art models across multiple datasets.

[80] proposed to use two neural network models to evaluate a sequence of turn-level features to rate the success of a dialogue. [47] encoded the context, the true response and the candidate response into vector representations using RNN, and then computed a score using a dot-product between the vectors in a linearly transformed space. [81] combined referenced and unreferenced metrics, where the former measured the similarity between reply and the groundtruth through word embedding, and the latter scored the correlation between reply and query trained with a max-margin objective function, where the negative reply is randomly sampled.

One promising approach comes from the idea of Turing test[85]–employing an evaluator to distinguish machine-generated texts from human-generated ones. [32] and [10] explored adversarial evaluation model [8] which assigns a score based on how easy it is to distinguish the generated responses from human responses, while [44] directly applied adversarial learning[24; 16] into dialogue generation.

3.2 Retrieval-based Methods

Retrieval-based methods choose a response from candidate responses. The key to retrieval-based methods is message-response matching. Matching algorithms have to overcome semantic gaps between messages and responses [28].

3.2.1 Single-turn Response Matching

Early studies of retrieval-based chatbots mainly focus on response selection for single-turn conversation[91], where only the message is used to select a proper response. Typically, the context and the candidate response are encoded as a vector respectively, and then the matching score is computed based on those two vectors. Suppose $\mathbf{x}$ is the vector representation of a message and $\mathbf{y}$ corresponds to the response vector representation, the matching function between $\mathbf{x}$ and $\mathbf{y}$ can be as simply as bilinear matching:

$$match(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{y},$$

where $\mathbf{A}$ is a pre-determined matrix, or more complicated ones. [49] proposed a DNN-based matching model for short texts response selection and combined the localness and hierarchy intrinsic in the structure. [28] improved the model by utilizing a deep convolutional neural network architecture to learn the representation of message and response, or directly learn the interacted representation of two sentences, followed by a multi-layer perceptron to compute the matching score. [92] extracted dependency tree matching patterns and used them as sparse one-hot inputs of a deep feed-forward neural network for context-response matching. [105] incorporated the topic vector generated by Twitter LDA model [121] into the CNN based structure to boost responses with rich content.

3.2.2 Multi-turn Response Matching

In recent years, multi-turn retrieval-based conversation draws more and more attention. In multi-turn response selection, current message and previous utterances are taken as input. The model selects a response that is natural and relevant to the whole context. It is important to identify important information in previous utterances and properly model the utterances relationships to ensure conversation consistency. [48] encoded the context (a concatenation of all previous utterances and current message) and candidate response into a context vector and a response vector through a RNN/LSTM

based structure, respectively, and then computed the matching degree score based on those two vectors. [110] selected the previous utterances in different strategies and combined them with current messages to form a reformulated context. [124] performed context-response matching on not only the general word level context vector but also the utterance level context vector. [106] further improved the leveraging of utterances relationship and contextual information by matching a response with each utterance in the context on multiple levels of granularity with a convolutional neural network, and then accumulated the vectors in a chronological order through a recurrent neural network to model relationships among utterances.

3.3 Hybrid Methods

Combing neural generative and retrieval based models can have significant effects on performance.[76] and [62] attempted to combine both methods. Retrieval-based systems often give precise but blunt answers, while generation-based systems tend to give fluent but meaningless responses. In an ensemble model, the retrieved candidate, along with the original message, are fed to an RNN-based response generator. The final response is given by a post-reranker. This approach combined the advantages of retrieval and generation based models, which was appealing in performance. [70] integrated natural language generation and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence neural network and latent variable neural network models and applied reinforcement learning to crowdsourced data and real-world user interactions to select an appropriate response from the models in its ensemble.

4. DISCUSSION AND CONCLUSION

Deep learning has become a basic technique in dialogue systems. Researchers investigated on applying neural networks to the different components of a traditional task-oriented dialogue system, including natural language understanding, natural language generation, dialogue state tracking. Recent years, end-to-end frameworks become popular in not only the non-task-oriented chit-chat dialogue systems, but also the task-oriented ones. Deep learning is capable of leveraging large amount of data and is promising to build up a unified intelligent dialogue system. It is blurring the boundaries between the task-oriented dialogue systems and non-task-oriented systems. In particular, the chit-chat dialogues are modeled by the sequence-to-sequence model directly. The task completion models are also moving towards an end-to-end trainable style with reinforcement learning representing the state-action space and combing the whole pipelines.

It is worth noting that current end-to-end models are still far from perfect. Despite the aforementioned achievements, the problems remain challenging. Next, we discuss some possible research directions:

- Swift Warm-Up. Although end-to-end models have drawn most of the recent research attention, we still need to rely on traditional pipelines in practical dialogue engineering, especially in a new domain warm-up stage. The daily conversation data is quite "big", however, the dialogue data for a specific domain is quite limited. In particular, domain specific dialogue data collection and dialogue system construction are laborsome. Neural network based models are better

at leveraging large amount of data. We need new way to bridge over the warm-up stage. It is promising that the dialogue agent has the ability to learn by itself from the interactions with human.

- Deep Understanding. Current neural network based dialogue systems heavily rely on the huge amount of different types of annotated data, and structured knowledge base and conversation data. They learn to speak by imitating a response again and again, just like an infant, and the responses are still lack of diversity and sometimes are not meaningful. Hence, the dialogue agent should be able to learn more effectively with a deep understanding of the language and the real world. Specifically, it remains much potential if a dialogue agent can learn from human instruction to get rid of repeatedly training. Since a great quantity of knowledge is available on the Internet, a dialogue agent can be smarter if it is capable of utilizing such unstructured knowledge resource to make comprehension. Last but not least, a dialogue agent should be able to make reasonable inference, find something new, share its knowledge across domains, instead of repeating the words like a parrot.

- Privacy Protection. Widely applied dialogue system serves a large number of people. It is quite necessary to notice the fact that we are using the same dialogue assistant. With the ability of learning through interactions, comprehension and inference, a dialogue assistant can inadvertently and implicitly store some of sensitive information [60]. Hence, it is important to protect users' privacy while building better dialogue systems.

5. REFERENCES

[1] J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26, 1992.

[2] R. Artstein, S. Gandhe, J. Gerten, A. Leuski, and D. Traum. Semi-formal evaluation of conversational characters. pages 22–35, 2009.

[3] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou. Affective neural response generation. *arXiv preprint arXiv:1709.03968*, 2017.

[4] N. Asghar, P. Poupart, X. Jiang, and H. Li. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 78–83, 2017.

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] L. Bahl, P. Brown, P. De Souza, and R. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 49–52. IEEE, 1986.

[7] A. Bordes, Y. L. Boureau, and J. Weston. Learning end-to-end goal-oriented dialog. In *ICLR*, 2017.

[8] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning.*

[9] P. F. Brown. The acoustic-modeling problem in automatic speech recognition. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1987.

[10] E. Bruni and R. Fernández. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288, 2017.

[11] K. Cao and S. Clark. Latent variable dialogue models and their diversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 182–187, Valencia, Spain, April 2017. Association for Computational Linguistics.

[12] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[13] S. Choudhary, P. Srivastava, L. Ungar, and J. Sedoc. Domain aware neural dialog system. *arXiv preprint arXiv:1708.00897*, 2017.

[14] H. Cuayhuitl, S. Keizer, and O. Lemon. Strategic dialogue management via deep reinforcement learning. *arxiv.org*, 2015.

[15] L. Deng, G. Tur, X. He, and D. Hakkani-Tur. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 210–215. IEEE, 2012.

[16] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[17] A. Deoras and R. Sarikaya. Deep belief network based semantic taggers for spoken language understanding. In *Interspeech*, pages 2713–2717, 2013.

[18] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[19] O. Dušek and F. Jurcicek. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190, Los Angeles, September 2016. Association for Computational Linguistics.

[20] O. Dušek and F. Jurcicek. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany, August 2016. Association for Computational Linguistics.

[21] M. Eric and C. D. Manning. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*, 2017.

[22] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*, 2017.

[23] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. A form-based dialogue manager for spoken language applications. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 701–704. IEEE, 1996.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[25] H. B. Hashemi, A. Asiaee, and R. Kraft. Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*, 2016.

[26] M. Henderson, B. Thomson, and S. Young. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471, 2013.

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

[28] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

[29] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM, 2013.

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

[31] C. Kamm. User interfaces for voice applications. *Proceedings of the National Academy of Sciences*, 92(22):10031–10037, 1995.

[32] A. Kannan and O. Vinyals. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198*, 2017.

[33] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 4:3581–3589, 2014.

[34] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[35] S. Lee. Structured discriminative model for dialog state tracking. In *SIGDIAL Conference*, pages 442–451, 2013.

[36] S. Lee and M. Eskenazi. Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In *SIGDIAL Conference*, pages 414–422, 2013.

[37] M. Lewis, D. Yarats, Y. Dauphin, D. Parikh, and D. Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2433–2443, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[38] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[39] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.

[40] J. Li, A. H. Miller, S. Chopra, M. Ranzato, and J. Weston. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*, 2016.

[41] J. Li, A. H. Miller, S. Chopra, M. Ranzato, and J. Weston. Learning through dialogue interactions by asking questions. *arXiv preprint*, 2017.

[42] J. Li, W. Monroe, and J. Dan. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*, 2016.

[43] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas, November 2016. Association for Computational Linguistics.

[44] 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*, pages 2147–2159, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[45] X. Li, Y.-N. Chen, L. Li, and J. Gao. End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv:1703.01008*, 2017.

[46] 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*, pages 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics.

[47] 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)*, pages 1116–1126, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[48] R. Lowe, N. Pow, I. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[49] Z. Lu and H. Li. A deep architecture for matching short texts. In *International Conference on Neural Information Processing Systems*, pages 1367–1375, 2013.

[50] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July 2015. Association for Computational Linguistics.

[51] G. Mesnil, X. He, L. Deng, and Y. Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. *Interspeech*, 2013.

[52] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

[53] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[54] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas, November 2016. Association for Computational Linguistics.

[55] K. Mo, S. Li, Y. Zhang, J. Li, and Q. Yang. Personalizing a dialogue system with transfer reinforcement learning. *arXiv preprint*, 2016.

[56] S. Möller, R. Englert, K. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake, and N. Reithinger. Memo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In *Ninth International Conference on Spoken Language Processing*, 2006.

[57] L. Mou, Y. Song, R. Yan, G. Li, L. Zhang, and Z. Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[58] N. Mrkšić, D. Ó Séaghdha, B. Thomson, M. Gasic, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 794–799, Beijing, China, July 2015. Association for Computational Linguistics.

[59] N. Mrkšić, D. Ó Séaghdha, T.-H. Wen, B. Thomson, and S. Young. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[60] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *ICLR*, 2017.

[61] Q. Qian, M. Huang, H. Zhao, J. Xu, and X. Zhu. Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861*, 2017.

[62] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 498–503, 2017.

[63] H. Ren, W. Xu, Y. Zhang, and Y. Yan. Dialog state tracking using conditional random fields. In *SIGDIAL Conference*, pages 457–461, 2013.

[64] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Conference on Empirical Methods in Natural Language Processing*, pages 583–593, 2011.

[65] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[66] R. Sarikaya, G. E. Hinton, and B. Ramabhadran. Deep belief nets for natural language call-routing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5680–5683, 2011.

[67] I. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI Conference on Artificial Intelligence*, 2017.

[68] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence*, 2016.

[69] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI Conference on Artificial Intelligence*, 2017.

[70] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, et al. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*, 2017.

[71] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China, July 2015. Association for Computational Linguistics.

[72] L. Shao, S. Gouws, D. Britz, A. Goldie, and B. Strope. Generating long and diverse responses with neural conversation models. *arXiv preprint arXiv:1701.03185*, 2017.

[73] X. Shen, H. Su, Y. Li, W. Li, S. Niu, Y. Zhao, A. Aizawa, and G. Long. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[74] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM, 2014.

[75] K. Sohn, X. Yan, and H. Lee. Learning structured output representation using deep conditional generative models. In *International Conference on Neural Information Processing Systems*, pages 3483–3491, 2015.

[76] Y. Song, R. Yan, X. Li, D. Zhao, and M. Zhang. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*, 2016.

[77] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[78] A. Stent, M. Marge, and M. Singhai. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 341–351, 2005.

[79] A. Stent, R. Prasad, and M. Walker. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 79. Association for Computational Linguistics, 2004.

[80] P.-H. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T.-H. Wen, and S. Young. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. *arXiv preprint arXiv:1508.03386*, 2015.

[81] C. Tao, L. Mou, D. Zhao, and R. Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *arXiv preprint arXiv:1701.03079*, 2017.

[82] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao. How to make context more useful? an empirical study on context-aware neural conversational models. In *Meeting of the Association for Computational Linguistics*, pages 231–236, 2017.

[83] V. K. Tran and L. M. Nguyen. Semantic refinement gru-based neural language generation for spoken dialogue systems. In *PACLING*, 2017.

[84] G. Tur, L. Deng, D. Hakkani-Tür, and X. He. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5045–5048. IEEE, 2012.

[85] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[86] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[87] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[88] P. Vougiouklis, J. Hare, and E. Simperl. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3370–3380, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[89] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics, 1997.

[90] M. A. Walker, O. C. Rambow, and M. Rogati. Training a sentence planner for spoken dialogue using boosting. *Computer Speech & Language*, 16(3):409–433, 2002.

[91] H. Wang, Z. Lu, H. Li, and E. Chen. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[92] M. Wang, Z. Lu, H. Li, and Q. Liu. Syntax-based deep matching of short texts. In *IJCAI*, 03 2015.

[93] Z. Wang and O. Lemon. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *SIGDIAL Conference*, pages 423–432, 2013.

[94] T.-H. Wen, M. Gasic, D. Kim, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[95] T.-H. Wen, M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[96] T.-H. Wen, M. Gašić, N. Mrkšić, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, and S. Young. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–129, San Diego, California, June 2016. Association for Computational Linguistics.

[97] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gasic, L. M. Rojas Barahona, P.-H. Su, S. Ultes, and S. Young. A

network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics.

[98] J. Williams. Multi-domain learning and generalization in dialog state tracking. In *SIGDIAL Conference*, pages 433–441, 2013.

[99] J. Williams, A. Raux, D. Ramachandran, and A. Black. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, 2013.

[100] J. D. Williams. A belief tracking challenge task for spoken dialog systems. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 23–24, 2012.

[101] J. D. Williams. Web-style ranking and slu combination for dialog state tracking. In *SIGDIAL Conference*, pages 282–291, 2014.

[102] J. D. Williams, K. Asadi, and G. Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–677, Vancouver, Canada, July 2017. Association for Computational Linguistics.

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

[104] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[105] Y. Wu, W. Wu, Z. Li, and M. Zhou. Topic augmented neural network for short text conversation. *CoRR*, 2016.

[106] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[107] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W. Y. Ma. Topic augmented neural response generation with a joint attention mechanism. *arXiv preprint arXiv:1606.08340*, 2016.

[108] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma. Topic aware neural response generation. In *AAAI Conference on Artificial Intelligence*, 2017.

[109] C. Xing, W. Wu, Y. Wu, M. Zhou, Y. Huang, and W. Y. Ma. Hierarchical recurrent attention network for response generation. *arXiv preprint arXiv:1701.07149*, 2017.

[110] R. Yan, Y. Song, and H. Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 55–64, New York, NY, USA, 2016. ACM.

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

[112] D. Yann, G. Tur, D. Hakkani-Tur, and L. Heck. Zero-shot learning and clustering for semantic utterance classification using deep learning. 2014.

[113] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi. Spoken language understanding using long short-term memory neural networks. In *IEEE Institute of Electrical & Electronics Engineers*, pages 189 – 194, 2014.

[114] K. Yao, B. Peng, G. Zweig, and K. F. Wong. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292*, 2016.

[115] K. Yao, G. Zweig, M. Y. Hwang, Y. Shi, and D. Yu. Recurrent neural networks for language understanding. In *Interspeech*, 2013.

[116] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li. Neural generative question answering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2972–2978. AAAI Press, 2016.

[117] S. Young, M. Gai, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. 24(2):150–174, 2010.

[118] R. Zens, F. J. Och, and H. Ney. Phrase-based statistical machine translation. In *German Conference on Ai: Advances in Artificial Intelligence*, pages 18–32, 2002.

[119] W. Zhang, T. Liu, Y. Wang, and Q. Zhu. Neural personalized response generation as domain adaptation. *arXiv preprint arXiv:1701.02073*, 2017.

[120] T. Zhao and M. Eskenazi. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Los Angeles, September 2016. Association for Computational Linguistics.

[121] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *European Conference on Advances in Information Retrieval*, pages 338–349, 2011.

[122] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*, 2017.

[123] H. Zhou, M. Huang, and X. Zhu. Context-aware natural language generation for spoken dialogue systems. In *COLING*, pages 2032–2041, 2016.

[124] X. Zhou, D. Dong, H. Wu, S. Zhao, D. Yu, H. Tian, X. Liu, and R. Yan. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas, November 2016. Association for Computational Linguistics.