Chat More: Deepening and Widening the Chatting Topic via A Deep Model

Wenjie Wang*
Shandong University
wenjiewang96@gmail.com

Minlie Huang
Tsinghua University
aihuang@tsinghua.edu.cn

Xin-Shun Xu
Shandong University
xuxinshun@sdu.edu.cn

Fumin Shen
University of Electronic Science and Technology of China
fumin.shen@gmail.com

Liqiang Nie
Shandong University
nieliqiang@gmail.com

ABSTRACT

The past decade has witnessed the boom of human-machine interactions, particularly via dialog systems. In this paper, we study the task of response generation in open-domain multi-turn dialog systems. Many research efforts have been dedicated to building intelligent dialog systems, yet few shed light on deepening or widening the chatting topics in a conversational session, which would attract users to talk more. To this end, this paper presents a novel deep scheme consisting of three channels, namely global, wide, and deep ones. The global channel encodes the complete historical information within the given context, the wide one employs an attention-based recurrent neural network model to predict the keywords that may not appear in the historical context, and the deep one trains a Multi-layer Perceptron model to select some keywords for an in-depth discussion. Thereafter, our scheme integrates the outputs of these three channels to generate desired responses. To justify our model, we conducted extensive experiments to compare our model with several state-of-the-art baselines on two datasets: one is constructed by ourselves and the other is a public benchmark dataset. Experimental results demonstrate that our model yields promising performance by widening or deepening the topics of interest.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Natural language generation**;

KEYWORDS

Multi-turn Dialog Systems; Response Generation; Deepening and Widening Topics; Multi-turn Dialog Dataset

ACM Reference Format:
Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat More: Deepening and Widening the Chatting Topic via A Deep Model. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210061

1 INTRODUCTION

Dialog systems, also known as conversational agents, have been widely used in a variety of applications, spanning from entertainment and knowledge sharing, to customer services. Roughly speaking, dialog systems can be divided into task-oriented and non-task-oriented categories. The former studies [19, 29] aim to accomplish tasks in vertical domains; whereas the latter studies [23, 32] target at chatting with people in open-domain topics. Dialog systems in these two categories can be implemented by either rule-, retrieval-, or generation-based methods. To be more specific, the heuristic templates defined by the rule-based methods somehow restrict the diversity of the desired dialog systems. As to the retrieval-based ones [31, 33, 34], they heavily depend on the archived repository. By contrast, generation-based methods are able to produce more flexible responses, usually by training a sequence-to-sequence network [2] which treats a post as input and the response as output. Despite their significance, the single-turn generation-based models [23, 32] neglect the contexts of the historical conversation session that play a pivotal role in the following chat. To alleviate such a problem, multi-turn dialog systems [20–22, 24] have been devised, whereby the context information is represented as a dense and continuous vector by numerous ways. For example, the hierarchical recurrent encoder-decoder model (HRED) [21] encodes the contexts hierarchically, using word-level and utterance-level recurrent neural networks. Despite the success of multi-turn dialog systems that leverage the context information in recent years, they still suffer from suboptimal performance due to the following problems: 1) According to our user study on 1,000 dialog sessions, no more than 45.2% of phrases in the contexts are directly helpful for response generation. Nevertheless, the majority of prior efforts consider all the phrases in the entire

*This work was partly done when the author was a summer intern at Tsinghua University.

* Corresponding author: Liqiang Nie (nieliqiang@gmail.com).

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210061

Table 1: An example of a multi-turn dialog that deepens and widens the chatting topic.

A dialog session	
person A:	There is a heavy rain today.
person B:	The umbrella is totally useless.
person A:	The rain is really heavy. (topic penetration) I got wet in the afternoon and caught a cold at night. (topic extension)
person B:	You should take some hot tea and get a good sleep.(topic extension)

context without elaborated distinction, which indeed incorporates noises and may thus hurt the desired performance. 2) Our findings show that, in a session, people often tend to deepen or widen the topic they are chatting about, leading the conversations to be more attractive and meaningful as shown in Table 1. However, few researchers thus far have addressed this issue. And even worse 3), current generation-based dialog systems frequently generate dull responses (such as "I don't know"), which are not informative or meaningless [21, 24]. In the light of this, a dialog system that is able to leverage the relevant context information and generate informative responses for offering deeper and wider conversations is highly desired.

However, it is non-trivial to tackle the aforementioned problems in multi-turn dialog systems, due to the following facts:

- The irrelevant phrases in the long contexts may overwhelm the relevant ones, which mislead the model and increase its computational burden. Therefore, how to identify the relevant words to effectively guide the response generation is an unsolved problem.
- Generating dull responses or talking about the same topic without going deeper or wider is boring, usually making people end the conversation quickly. Thereby, challenges exist in how we can avoid dull responses and generate responses that are not only relevant but also capable of deepening and widening the dialog topics.
- A large-scale dataset is crucial to ensure the robustness of the generation-based models. Yet, the released multi-turn dialog datasets are either in vertical domains or at a small scale [1, 9, 10, 16, 31].

To address the aforementioned problems, we develop a deep and wide neural network, named DAWnet, as shown in Figure 1, consisting of three parallel channels, namely global, deep, and wide channels. DAWnet is able to deepen and widen the chatting topics. More specifically, DAWnet segments the utterances and the global channel first transforms the given context to an embedding vector that encodes the complete historical information. DAWnet then extracts all the keywords from the context. On top of the collected keywords and the context embedding, the wide channel utilizes an attention-based recurrent neural network (RNN) to predict the wider keywords. The wider keywords means the keywords that may not appear in the given context and help to widen the topic. As to the deep channel, it trains a Multilayer Perceptron (MLP) model to select a few deeper keywords for an in-depth discussion, whereby the inputs are the context embedding vector and the collected keywords. The deeper keywords refer to the keywords in the context which support to deepen the current topic. The whole scheme ultimately inputs the outputs of the contextual encoder, the selected keywords in the deep channel, and the predicted keywords in the wide channel into a selector before decoding a meaningful response. The selector equipped with an attention mechanism judges the contribution of the three inputs.

To train DAWnet and evaluate its performance on improving the coherence, informativeness, and diversity of generated responses, we built a dataset of multi-turn dialogs in the open domain, named Sina Weibo Conversation Corpus. It covers a rich range of topics in our daily conversations. The dialogs are collected from Sina weibo[1], one of the most popular social media sites in China and used by over 30% of Internet users. To thoroughly justify DAWnet, we also evaluate DAWnet on a benchmark dataset, DailyDialog [7]. We compared DAWnet with several state-of-the-art baselines on the two datasets. Experimental results demonstrate that DAWnet yields promising performance in multi-turn dialog systems.

The contributions of our work are threefold.

- DAWnet is able to separate the topic-related keywords from the irrelevant ones, and use the selected relevant keywords to generate meaningful responses. This is a key step to avoid dull responses.
- As far as we know, this is the first work on deepening and widening the chatting topic in the multi-turn dialog systems by a hybrid RNN and DNN model which encourages users to talk more.
- We constructed a dataset of multi-turn dialogs in the open domain. In addition, we released the data, code and involved parameters to facilitate other researchers in this field[2].

The rest of this paper is organized as follows. In Section 2, we briefly survey the related works. Section 3 presents the details of our proposed model. We conduct experiments and analyze the results in Section 4, followed by the conclusion and future work in Section 5.

2 RELATED WORK

Dialog systems have been advanced substantially due to the increase of available datasets and the fast development of deep neural network technologies. Conventional dialog systems generally depend on hand-built templates and rules [28], hindering the generalization ability towards other domains. In recent years, more data-driven dialog systems have been proposed. In the open domain, they roughly fall into two major categories: retrieval- and generation-based methods. The former methods [12–14, 27, 31] generally select a suitable response by ranking the response candidates with various matching algorithms. Compared to single-turn dialog systems [26, 27], multi-turn dialog systems [31, 33, 34] have been explored to to leverage dialog context in recent years. Although retrieval-based models can retrieve informative and diverse responses from the repository, they have to meet this precondition: the selected responses should pre-exist. Thereby, the performance is restricted by the scale and

[1] https://weibo.com/.
[2] https://sigirdawnet.wixsite.com/dawnet.

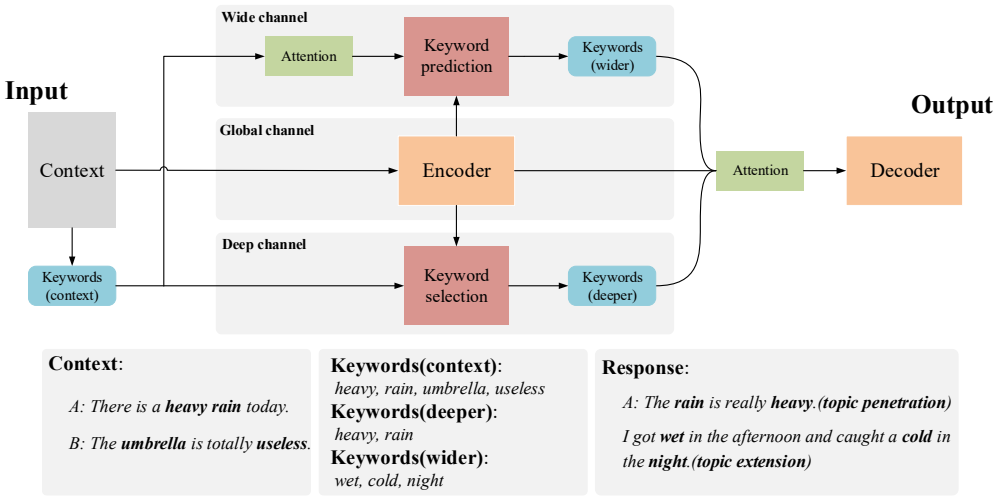

Figure 1: Schematic illustration of our proposed model, named DAWnet. DAWnet comprises three channels, namely global, wide and deep channels. The three channels respectively encode the context into an embedding vector, predict wider keywords, and select deeper keywords.

quality of the repository. The latter generation-based systems, inspired by statistical machine translation, model the mapping between a post and its response with data-driven methods. In the beginning, researchers leveraged the encoder-decoder framework [17, 23] to address the task of single-turn dialog systems. As the attention mechanism [2] in machine translation [25] becomes more popular, it has been more frequently incorporated into the encoder-decoder framework [32] to boost the performance. Later, more research efforts were devoted to modeling the conversational history in multi-turn dialog systems. Dynamic-context generative model [24] encodes the context and post into fixed-length vectors and feeds them into the Recurrent Neural Network Language Model for response generation. HRED models the hierarchy of contexts with two RNNs: one at the word-level and the other at the utterance level. Built upon HRED, VHRED [22] presents the latent stochastic variables into the model and MrRNNs [20] model multiple parallel sequences by factorizing the joint probability over the sequences. Interactive dialog context language model tracks the interactions between speakers in a dialog by using two parallel RNNs. Li et al. [6] simulated the dialog between two agents by integrating the strengths of neural SEQ2SEQ systems [2] and deep reinforcement learning. Mei et al. [11] proposed RNN-based dialog models equipped with a dynamic attention mechanism to increase the scope of attention on the conversational history.

A common problem shared by the generation-based models mentioned above is that they tend to generate dull responses due to the high frequency of the general words. To improve the quality of responses, various methods have been proposed. Yao et al. [35] added a RNN to model the dynamics of the intention process in a SEQ2SEQ model. Li et al. [5] used maximum mutual information as the objective function in neural models to alleviate dull responses. Topic-aware sequence-to-sequence model [32] considers topic words in response generation where the topic words are from a Twitter LDA model. However, obtaining such topic words using Twitter LDA is challenging because the posts are very short and the topics in open-domain conversations are sparse. According to our statistics on three million sessions from the Sina Weibo Conversation Corpus, there are over 25 million unique words and more than half of them appear less than five times. Considering the sparsity issue, it is hard to determine the related topic words based on a short post.

Our work is different from the aforementioned studies in the following aspects.

- We studied the context-aware dialog systems because a dialog is continuous, and thus response generation has to consider the history of a conversation.
- Since it is challenging to assign a suitable topic to a compact post and choose the meaningful words from a topic, we predict the wider keywords that may not appear in the context by the wide channel and select the more specific keywords using a deep channel, which helps the model to deepen and widen the chatting topic.

3 MODEL

To deepen and widen the chatting topics, we present a scheme to explore the keywords in a dialog as shown in Figure 1. The scheme first segments the utterances and extracts the keywords from the context. After that, the model inputs the context and its keywords into three parallel channels, namely, global, wide and deep channels. These channels respectively encode the context into an embedding vector, predict wider keywords, and select deeper keywords based on the context and its keywords. Ultimately, the model adopts an attention mechanism to weigh the context and keywords before feeding them into the RNN decoder that is used to generate a response. In this section, we will detail each component of this scheme.

3.1 Keyword Extraction

We apply term frequency-inverse document frequency (well-known as TF-IDF) [4, 18, 30], to extract keywords from context. In areas such as information retrieval, text mining, and user modeling, TF-IDF is largely applied to measure the importance of a word in the document. It assumes that the importance of a word is directly proportional to the times it appears in the document and is often offset by the frequency of the word in the entire corpus. We removed the stop words and only retained nouns, verbs, and adjectives in DailyDialog [7] and Sina Weibo Conversation Corpus. We then considered a session as a document and a word as a term to calculate the TF-IDF value of each word. We finally chose top 20 keywords from each session.

3.2 Global Channel

In this channel, we leverage a RNN equipped with gated recurrent units (GRUs) to encode the given context into a vector. Given the context comprising several utterances, we consider it as a sequence of tokens. The RNN encoder then calculates the context vector as follows,

$$\begin{cases} C = \{w_1 ... w_t, ..., w_T\}, \\ \mathbf{h_t} = f(\mathbf{h_{t-1}}, \mathbf{e_{w_t}}), \end{cases} \tag{1}$$

where the context C denotes a sequence of T tokens, $\mathbf{e_{w_t}}$ refers to the embedding vector of the t-th token, $\mathbf{h_t}$ is the hidden state of the RNN at time t, and f is a non-linear function. In our work, the GRU network is parameterized as

$$\begin{cases} \mathbf{z} = \sigma_g(\mathbf{W_z x_t} + \mathbf{U_z h_{t-1}}), \\ \mathbf{r} = \sigma_g(\mathbf{W_r x_t} + \mathbf{U_r h_{t-1}}), \\ \mathbf{s} = \sigma_h(\mathbf{W_s x_t} + \mathbf{U_s}(\mathbf{h_{t-1}} \circ \mathbf{r})), \\ \mathbf{h_t} = (1 - \mathbf{z}) \circ \mathbf{s} + \mathbf{z} \circ \mathbf{h_{t-1}}, \end{cases} \tag{2}$$

where $\mathbf{x_t}$ is the input vector, $\mathbf{h_t}$ is the output vector, $\mathbf{z}$ is the update gate vector, $\mathbf{r}$ is the reset gate vector, $\mathbf{W_z}$, $\mathbf{W_r}$, $\mathbf{W_s}$, $\mathbf{U_z}$, $\mathbf{U_r}$ and $\mathbf{U_s}$ are parameter matrices, $\circ$ denotes element-wise multiplication, σ_g and σ_h are sigmoid and tanh activation functions, respectively.

3.3 Wide Channel

In this channel, we train an attention-based RNN model to predict keywords to extend topics. Given the vector $\mathbf{c}$, the RNN for keyword prediction is initialized by the last hidden state of the encoder and updated via

$$\mathbf{s_t} = f(\mathbf{s_{t-1}}, [\mathbf{e}_{\mathbf{k}^p_{t-1}}, \mathbf{c_t}]), \tag{3}$$

where $\mathbf{e}_{\mathbf{k}^p_{t-1}}$ is the embedding vector of the keyword at time t-1 in the sequence of the predicted keywords, $\mathbf{c_t}$ is the vector at time t acquired from the attention mechanism, $[\mathbf{e}_{\mathbf{k}^p_{t-1}}, \mathbf{c_t}]$ is the concatenation of the two vectors, $\mathbf{s_t}$ is the hidden state of the RNN at time t. The vector $\mathbf{c_t}$ at time t is calculated by

$$\begin{cases} \mathbf{m_i} = \mathbf{W_t} \mathbf{e}_{\mathbf{k}^c_i}, \\ \mathbf{c_t} = \sum_{i=1}^{T} \alpha_{ti} \mathbf{h_i} + \sum_{i=T+1}^{T+M} \alpha_{ti} \mathbf{m_i}, \end{cases} \tag{4}$$

where $\mathbf{e}_{\mathbf{k}^c_i} \in \mathbb{R}^{d_e}$ is the embedding vector of the i-th contextual keyword, $\mathbf{W_t} \in \mathbb{R}^{d_h \times d_e}$ is a transform matrix projecting the

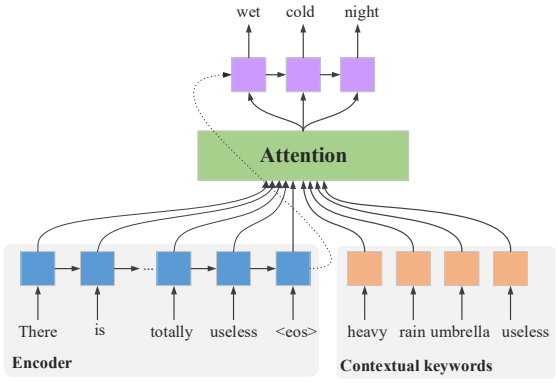

Figure 2: Keyword prediction for topic extension. The decoder for keyword prediction is initialized by the last hidden state of the encoder, and generates keywords by attending to the hidden states of the encoder and contextual keywords.

embedding vector of keywords to a high dimensional space with d_h dimension, $\mathbf{m_i}$ is the vector acquired by projection, d_h is the size of the hidden state, T is the number of contextual tokens, and M is the number of contextual keywords. The weight coefficient α_{ti} is defined as

$$\begin{cases} \alpha_{ti} = \dfrac{exp(e_{ti})}{\sum_{j=1}^{T+M} exp(e_{tj})}, \\ e_{ti} = \eta(\mathbf{s_{t-1}}, \mathbf{h_i}) \quad i = 1, ..., T, \\ e_{ti} = \eta(\mathbf{s_{t-1}}, \mathbf{m_i}) \quad i = T+1, ..., M, \end{cases} \tag{5}$$

where η is implemented by a MLP model with *tanh* as an activation function. The RNN decoder calculates the probability of the predicted keyword at each step by a projection layer,

$$p(k_t^p | C, k_1^c, ..., k_M^c, k_1^p, ..., k_{t-1}^p) = \mathbf{o_t^k} \cdot \sigma_s(\mathbf{W_k s_t} + \mathbf{b_k}),$$

whereby d_v^k is the size of the keyword vocabulary, $\mathbf{W_k} \in \mathbb{R}^{d_v^k \times d_h}$ and $\mathbf{b_k} \in \mathbb{R}^{d_v^k}$ are the parameters of the projection layer, σ_s means the softmax function, k_t^p is the i-th predicted keyword and $\mathbf{o_t^k}$ is the one-hot vector of k_t^p. The inputs of the projection layer are the hidden states of the decoder and it outputs the probability distribution of all keywords. Formally, the probability of generating all predicted keywords $p(k_1^p, ..., k_N^p | C, k_1^c, ..., k_M^c)$ is formulated as

$$p(k_1^p, ..., k_N^p | C, k_1^c, ..., k_M^c) = p(k_1^p | C, k_1^c, ..., k_M^c) \prod_{t=2}^{N} p(k_t^p | C, , k_1^c, ..., k_M^c, k_1^p, ..., k_{t-1}^p).$$

In this channel, the decoder essentially predicts a sequence of wider keywords which will be fed into the decoder for response generation.

3.4 Deep Channel

The objective of the deep channel is to choose the useful keywords from the context to deepen the topic of interest. A MLP model with RELU as an activation function is employed to calculate the weights of the contextual keywords. The inputs are the last hidden

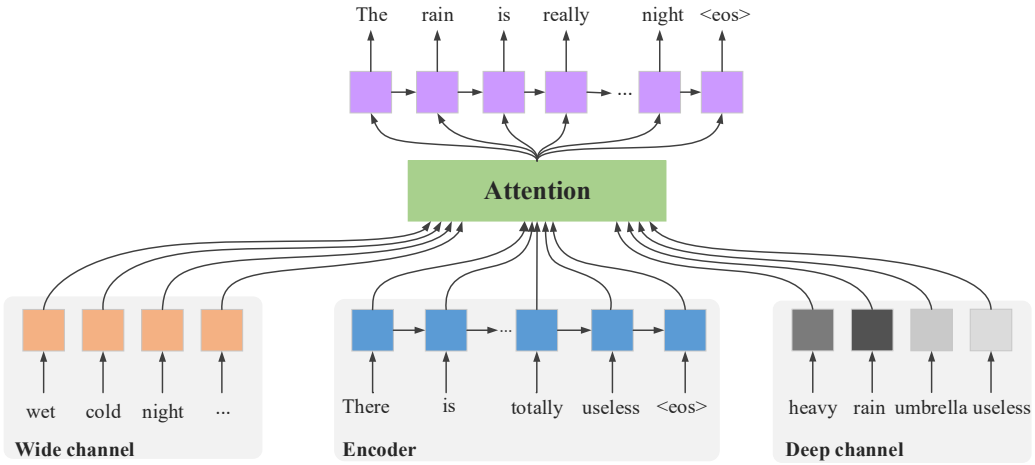

Figure 3: The decoder for response generation. The decoder, initialized by the last hidden state of the encoder, generates tokens by attending to the hidden states of the encoder, the selected keywords from the deep channel, and the predicted keywords from the wide channel.

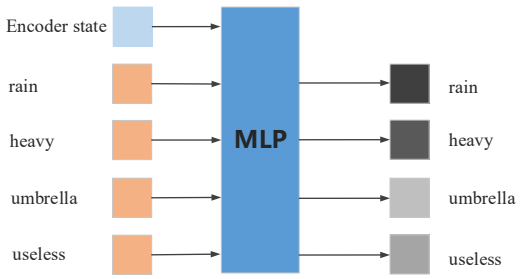

Figure 4: Deep keyword selection procedure. The MLP model takes as input the hidden state of the encoder and contextual keywords, and outputs the weights of contextual keywords.

state of the encoder and the embedding vectors of contextual keywords. The output is given by

$$\begin{cases} \mathbf{l_0} = [\mathbf{h_T}, \mathbf{e_{k_1^c}}, \mathbf{e_{k_2^c}}, ..., \mathbf{e_{k_M^c}}], \\ \mathbf{q} = \mathbf{MLP}(\mathbf{l_0}), \end{cases} \tag{6}$$

where $\mathbf{l_0}$ is calculated by concatenating the last hidden state of the encoder and the embedding vectors of M contextual keywords, $\mathbf{q} \in \mathbb{R}^M$ represents the weights of contextual keywords and the MLP model is implemented by four layers of neurons which has relu as the activation function in the first three layers and sigmoid in the output layer. The vectors of selected keywords are updated by

$$\mathbf{m_i} = q_i \mathbf{W_t} \mathbf{e_{k_i^c}}. \tag{7}$$

The vectors of selected keywords in this channel will be used in the decoder for response generation.

3.5 Decoder

As shown in Figure 3, the RNN decoder for response generation is similar to that for keyword prediction but the vector $\mathbf{c}$ is different,

$$\begin{cases} \mathbf{n_i} = \mathbf{W_t} \mathbf{e_{k_i^p}}, \\ \mathbf{c_t} = \sum_{i=1}^{T} \alpha_{ti} \mathbf{h_i} + \sum_{i=T+1}^{T+M} \alpha_{ti} \mathbf{m_i} + \sum_{i=T+M+1}^{T+M+N} \alpha_{ti} \mathbf{n_i}, \end{cases} \tag{8}$$

where N is the number of the predicted keywords, $\mathbf{e_{k_i^p}}$ is the embedding vector of the i-th predicted keyword, $\mathbf{W_t} \in \mathbb{R}^{d_h \times d_e}$ is a transform matrix, the same as that of the wide channel, and $\mathbf{n_i}$ is acquired by projecting $\mathbf{e_{k_i^p}}$. Given the last hidden state of the encoder and vector $\mathbf{c}$, the decoder RNN predicts the target response token-by-token,

$$p(y_t|C, k_1^p, ..., k_N^p, k_1^s, ..., k_M^s, y_1, ..., y_{t-1}) = \mathbf{o_t^y} \cdot \sigma_s(\mathbf{W_y s_t} + \mathbf{b_y}),$$

where $\mathbf{W_y} \in \mathbb{R}^{d_v^y \times d_h}$ and $\mathbf{b_y} \in \mathbb{R}^{d_v^y}$ are the parameters of the projection layer that projects the hidden state to the probability distribution of all words in the vocabulary, d_v^y is the size of the vocabulary, k_i^s represents the i-th selected keyword, y_t represents the t-th token in the response, σ_s is the softmax function, and $\mathbf{o_t^y}$ is the one-hot vector of y_t. The probability of generating the response $p(y_1, ..., y_L|C, k_1^p ... k_N^p, k_1^s ... k_M^s)$ is calculated by

$$\begin{cases} \mathcal{U} = \{C, k_1^p, ..., k_N^p, k_1^s, ..., k_M^s\}, \\ p(y_1, ..., y_L|\mathcal{U}) = p(y_1|\mathcal{U}) \prod_{t=2}^{L} p(y_t|\mathcal{U}, y_1, ..., y_{t-1}), \end{cases} \tag{9}$$

where L is the number of tokens in a response.

3.6 Loss Function

Formally, let us denote Θ as the parameter set of the whole model, and we estimate Θ from $\mathcal{D} = \{(C_i, \mathcal{K}_i^c, \mathcal{K}_i^p, \mathcal{K}_i^s, \mathcal{R}_i)\}_{i=1}^{i=I}$, where $C_i, \mathcal{K}_i^c, \mathcal{K}_i^p, \mathcal{K}_i^s$ and $\mathcal{R}_i$ represent the context, contextual keywords, predicted keywords, selected keywords and responses,

respectively. We optimize Θ by minimizing the following objective function ℓ,

$$\begin{cases} \ell_0 = -\frac{1}{I}\sum_{i=1}^{I} \log p(y_1, ..., y_L|\mathcal{U}), \\ \ell_1 = -\frac{1}{I}\sum_{i=1}^{I} \log p(k_1^p, ..., k_N^p|C, k_1^c, ..., k_M^c), \\ \ell_2 = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{M}(u_j \log q_j + (1-u_j)\log(1-q_j)), \\ \ell = \ell_0 + \beta_1\ell_1 + \beta_2\ell_2, \end{cases} \tag{10}$$

where β_1 and β_2 are two parameters in the objective function, ℓ_0, ℓ_1 and ℓ_2 correspond to the objective function of the response decoder, wide channel and deep channel, respectively. Meanwhile, we have to mention that $u \in \{0, 1\}$ and $u_i = 1$, if and only if $u_i \in \mathcal{K}_i^s$.

4 EXPERIMENTS

4.1 Dataset

To validate the performance of DAWnet, we trained and evaluated DAWnet one two open-domain multi-turn dialog datasets, DailyDialog[3] [7] and Sina Weibo Conversation Corpus.

DailyDialog is a human-written dataset and covers various topics in our daily life such as finance, politics to tourism. The dataset contains 13,118 multi-turn dialog sessions. The average turns per dialog, tokens per dialog, and tokens per utterance are 7.9, 114.7 and 14.6, respectively. We extracted keywords and finally constructed 13,118 samples, including 11,118 for training, 1,000 for validation, and 1,000 for test.

In addition to the English dataset of DailyDialog, we constructed a representative Chinese dataset, namely Sina Weibo Conversation Corpus. In particular, we crawled massive conversations between two people from Sina Weibo, one of the most popular social media sites in China, used by over 30% of Internet users, covering rich real-world topics in our daily life. The raw data comprises about 20 million sessions and each session contains many post-response pairs between two people. Thereafter, we selected the sessions which satisfy the following rules: 1) The turns in the session are more than three. 2) The response is meaningful and has two keywords at least. The keywords refer to those with the TF-IDF value no less than a given threshold. Following the selection, we pre-processed the sessions by removing the noisy words and converting the emojis into the corresponding words. Ultimately, we had a dataset consisting of 1,587,119 sessions in total. The average turns and tokens per dialog are 3.71 and 42.17, respectively. In Sina Weibo Conversational Corpus, we performed Chinese word segmentation with the help of a public tool[4], extracted keywords, and then randomly chose 1,407,119 samples for training, 9,000 for validation and 9,000 for test.

In the two datasets, the last utterance was used as the response and the remaining ones were treated as the context. The keywords in a response were divided into two categories: the predicted keywords and the selected ones. The selected keywords directly relate to the current topic, and appear in the context; whereas the predicted ones are relevant to the topic but not necessarily occur in the context.

[3]The dataset is available on http://yanran.li/dailydialog.
[4]https://github.com/fxsjy/jieba

4.2 Experimental Settings

4.2.1 Hyper parameters. In our experiments, we extracted at least five keywords from the context and two keywords from the response of each dialog session. The dimension of word embedding was set to 100 and the embedding matrix was randomly initialized. The vocabulary size is 20,000 in DailyDialog, and 40,000 in Sina Weibo Conversational Corpus. All words out of vocabulary were mapped to a special token UNK. The RNNs have 4-layer GRU structures with 1024 hidden cells for each layer. The MLP model in the wide channel comprises 1,024, 512, 128 nodes in the first three layers, respectively. We used Adam [3] to optimize the objective function and the learning rate was initialized as 0.001, which changed dynamically in the training. In the objective function ℓ, we finally chose $\{\beta_1, \beta_2\} = \{0.5, 0.5\}$ out of $\{\beta_1, \beta_2\} \in \{\{0.25, 0.25\}, \{0.5, 0.5\}, \{0.75, 0.75\}, \{1, 1\}\}$ based on the perplexity of responses. In the training, we used the validation set for early stop.

4.2.2 Evaluation Metrics. To measure the performance of DAWnet, we followed existing studies and adopted several standard metrics: perplexity (PPL), BLEU [15], and diversity-based Distinct-1 [5]. We compared DAWnet with the baselines in terms of these metrics. In particular, PPL describes how well a probability model predicts the target samples. BLEU quantifies n-gram overlaps between the generated response and the reference response. In some way, Distinct-1 reflects the diversity of the responses.

PPL is widely used in probability models to quantify their performance. Formally, in a language model, given a utterance $\mathcal{S} = \{w_1, w_2, ..., w_N\}$, PPL is defined as

$$PPL = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 q(w_i|w_1, w_2, ..., w_{i-1})}, \tag{11}$$

where $q(w_i|w_1, w_2, ..., w_{i-1})$ means the probability of generating the word w_i in the language model. When PPL is smaller, the model performs better.

BLEU is a metric for machine translation. Formally, BLEU-N score is calculated by

$$score = \exp(min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n), \tag{12}$$

where r and c respectively denote the lengths of the reference response and candidate one, p_n presents the modified n-gram precision [15], N means using n-grams up to length N and $w_n = \frac{1}{N}$. Higher the BLEU value is, more similar the reference response and candidate response are.

To validate the diversity of responses, we adopted the Distinct-1 metric designed by Li et al [5]. Distinct-1 is calculated as the number of distinct unigrams in the generated responses scaled by the total number of generated tokens. The higher Distinct-1 value somehow means that the responses are more diverse.

Table 2: Performance comparison between DAWnet and the baselines on the DailyDialog dataset and Sina Weibo Conversation Corpus.

Dataset	Model	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Distinct-1
DailyDialog Dataset	SEQ2SEQ	41.60	0.1614	0.0897	0.0228	0.0061	0.0281
	HRED	41.38	0.0640	0.0213	0.0098	0.0044	0.0442
	VHRED	64.66	0.1227	0.0346	0.0126	0.0053	0.0525
	DAWnet	**39.36**	**0.1690**	**0.1004**	**0.0424**	**0.0198**	**0.0778**
Sina Weibo Conversation Corpus	SEQ2SEQ	39.82	**0.1117**	0.0318	0.0103	**0.0003**	0.0061
	HRED	38.32	0.0020	0.0007	0.0004	0.0001	0.0184
	VHRED	42.76	0.0110	0.0026	0.0009	0.0001	0.0288
	DAWnet	**36.91**	0.0913	**0.0320**	**0.0145**	0.0002	**0.0760**

Table 3: The results of subjective evaluation.

Opponent	Win	Loss	Tie	Kappa
DAWnet vs. SEQ2SEQ	27.7%	21.5%	50.8%	0.35
DAWnet vs. HRED	28.4%	26.2%	45.4%	0.31
DAWnet vs. VHRED	29.7%	21.2%	49.1%	0.33

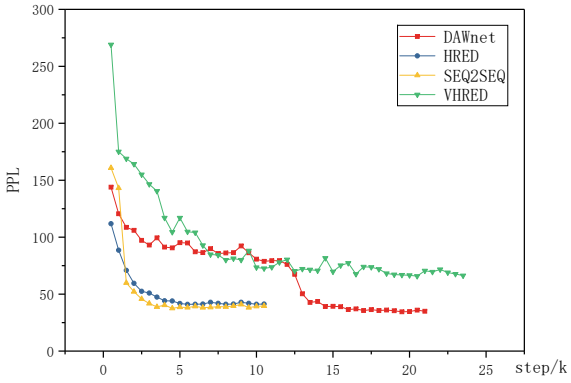

Figure 5: The curve of PPL along the training steps on the DailyDialog dataset.

4.2.3 Baselines.

- SEQ2SEQ+Attention: Attention-based SEQ2SEQ [2] has shown its promising performance in many NLP tasks and is widely used as a baseline in generation-based dialog systems. It is denoted as SEQ2SEQ hereafter.
- HRED: HRED [21] is capable of capturing the useful information in a long context by modeling the context in hierarchical RNNs. It has demonstrated its effectiveness in multi-turn dialog systems.
- VHRED: Built on HRED, VHRED is also a neural network-based generative model but with latent stochastic variables. In previous work [22], VHRED performs better and facilitates the generation of long responses.

4.3 Overall Performance

We compared DAWnet with all the baselines with respect to several standard metrics. Figure 5 plots the convergence of all the models. The results are summarized in Table 2. We also conducted the t-test between DAWnet and each of the baselines in terms of perplexity. We observed that all the p-vlaues are much smaller than 0.05, indicating that our model is statistically significant. From Table 2, we can observe the following points:

- DAWnet achieves the lowest perplexity on the two datasets. The advantageous performance verifies that the predicted and selected keywords actually provide more information to benefit the model.
- The higher BLEU values demonstrate the responses are closer to the ground truth. On Sina Weibo Conversation Corpus, VHRED, DAWnet and particularly HRED perform badly in terms of BLEUs. In fact, Liu et al. [8] have reported the weak correlation between BLEU and human evaluation. We also found that many responses of HRED, VHRED and DAWnet are reasonable but quite different from the ground truth.
- The Distinct-1 value of DAWnet is much higher than all the baselines, which demonstrates that DAWnet can generate more different words and the responses are hence more diverse. The rich keywords in wide and deep channels ensure the response diversity.

4.4 Subjective Evaluation

In addition to the objective evaluation, we carried out a subjective one. In particular, we randomly selected 500 samples from the two datasets, respectively. For each sample, we used DAWnet and the three baselines to generate the responses. Accordingly, we obtained 3,000 triplets (sample, response$_1$, response$_2$) whereby one response is generated by DAWnet and the other is generated by a baseline. We then invited three undergraduate students to annotate each triplet by following these rules: 1) The system identifier from which a response is generated is masked to the annotators; 2) Each annotator is required to independently rate among win, loss, and tie (win: response$_1$ is better; loss: response$_2$ is better; tie: they are equally good or bad); 3) Before labeling, the annotators were trained with a few samples and they were required to comprehensively consider four factors when rating: relevance, logical consistency, fluency and informativeness; And 4) the strategy of majority voting was employed to judge which one is better. Notably, if three annotators rate three different options, we counted this triplet as "tie". Table 3 summarizes the results of subjective evaluation. The kappa scores indicate that the annotators came to a fair agreement in the judgment.

Table 4: Case study. The test samples of DAWnet and the baselines. The reference means the ground-truth response in the datasets.

Context	Response	The predicted keywords by DAWnet
A: Are you busy tomorrow morning? B: I'm free. What's up? A: Someone has to pick up the boss at the airport.	DAWnet: **where** is the **airport**? SEQ2SEQ: I am a truth. HRED: I'm sorry , but I don't have any time to do. VHRED: Do you have any plans for me? Reference: Oh , I just remembered I've got a report to write.	**where airport**
A: What's wrong with you, young man? B: Doctor, I have a bad cough and a headache.	DAWnet: Do you have a **fever**? SEQ2SEQ: It's Gone I think to have a instructions. HRED: You can take a car. VHRED: You are kidding! I have to be able to get you to the hospital. You are really busy with me. Reference: Do you have a fever?	**fever**
A: I know I am late. I'm really sorry. I lost my bag. B: I'll call the lost and found office.	DAWnet: Thank you. SEQ2SEQ: I'm questioning it. HRED: what? VHRED: Are you kidding? It is a nice day . Reference: I didn't think of it. Thank you.	**help** you thank
A: Excuse me. I have an appointment with Mr Li at nine. May I come in? B: Yes, come in please. I am Mr Li. You must be Mr Liu, right? A: Yes, I am Mr Liu. Thanks.	DAWnet: **News** that you have a noncommercial **company**. SEQ2SEQ: Yes, I am to accommodate the requisition of the office. HRED: Good! VHRED: Let me see. Reference: I'd like to start this interview with some questions.	**company** speak **News**

From Table 3, we have the following observations:

- DAWnet outperforms the baselines. Among the three baselines, SEQ2SEQ performs the worst and HRED the best. We analyzed the bad cases of the baselines. Responses generated by SEQ2SEQ are not fluent or logical. HRED tends to generated general and less informative responses, for example "Ok." and "Sure.". Responses generated by VHRED are usually long and informative but most of them are irrelevant to the context.
- Over 45% of the triplets were labeled as "tie". By checking these samples, we found most of them are illogical or irrelevant.

4.5 Discussions

4.5.1 Case Study. Table 4 lists some responses generated by DAWnet and the baselines. From Table 4, we can observe a few findings:

- In the first sample, HRED's response is reasonable but conflicts with the context "I'm free.". The responses of SEQ2SEQ and VHRED are fluent but they are not relevant to the context. DAWnet predicts the keywords "where" and "airport", and generates "where is the airport?" as a response, which is reasonable.
- The response of VHRED in the second sample explains our prior conclusion that VHRED tends to generate long and informative responses but irrelevant to the context. In this case, DAWnet predicts the related keyword "fever" and this benefits good response generation. It demonstrates that the wide channel works well and thus DAWnet is able to widen the topic.
- DAWnet in the third sample predicts the keyword "help" and by nature generates "Thank you." as a response. This means it predicts that "I'll call the lost and found office." is a kind of help and the wide channel plays a pivotal role.
- The fourth case is a typical and controversial example in the test samples of DAWnet. The response of DAWnet seems to transfer the topic but it actually has some problems in the coherence and grammar. This is because that it is probably much influenced by the predicted keywords and they dominate the decoder for response generation.

4.5.2 Model Ablation. To examine the effectiveness of the wide and deep channels, we eliminated one of the channels each time, and verified the performance. We conducted two individual experiments by removing the wide or deep channel respectively. From the results presented in Table 5, we observed that: 1) The performance dropped when we removed the wide or deep channel,

Table 5: The results of ablation test on the DailyDialog dataset.

Model	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Distinct-1
No wide channel	43.72	0.1164	0.0706	0.0249	0.0108	0.0663
No deep channel	43.35	0.1550	0.0923	0.0372	0.0173	**0.0810**
DAWnet	**39.36**	**0.1690**	**0.1004**	**0.0424**	**0.0198**	0.0778

Table 6: Four examples which explain the four kinds of errors. The ungrammatical, irrelevant, illogical and universal response are displayed from top to bottom.

Context	Response
A: Is it expensive?	Yes. I gotten for 250. (an ungrammatical response)
A: I need to buy some flowers for my wife. B: Perhaps you'd be interested in red roses.	How is the counter-offer? (an irrelevant response)
A: How much is it? B: It's free of charge.	Fairly expensive. (an illogical response)
A: Can I have the roll of film developed here?	I'm not sure. (a universal response)

demonstrating that both the wide channel and the deep channel are indispensable to improve the performance; 2) The Distinct-1 value increases when we removed the deep channel. The model with the only wide channel tends to generate some relevant but different words. Impacted by these words, the model is inclined to transfer the topic, which could explain the phenomenon. However, if the wide channel dominates the decoder for response generation, the coherence and relevance may become poor during the whole session. The forth case in Table 4 shows an example for the problem. Therefore, the balance between the wide and deep channels is crucial.

4.5.3 Error analysis. To further improve the performance of DAWnet, we chose the bad samples judged during the subjective evaluation and analyzed why they were worse in the comparison. The bad cases fall into four categories: the ungrammatical response, the irrelevant response, the illogical response, and the universal response. The ungrammatical response refers to the response that is not fluent and has grammatical errors. If the response is fluent but irrelevant to the context, we will call it as the irrelevant response. The illogical response means that it is fluent and relevant, but it conflicts with itself or the given context in logic. The universal response is less informative and general, for example "I don't know." or "Ok.". The four kinds of bad cases are explained with examples in Table 6. As for the specific quantitative analysis, ungrammatical, irrelevant, illogical and universal responses occupy 25.8%, 27.9%, 37.6%, and 8.6%, respectively. The results imply that:

- DAWnet generates the least universal responses in four kinds of bad cases.
- Although the context, the predicted and selected keywords provide abundant information of the conversational history, the model still has the difficulty in generating fluent and relevant responses perfectly.
- The logic issue is a particularly serious problem in the bad cases. In fact, the logic problem is one of the most challenging problems in neural language generation models.

The analysis enlightens us to further improve the model in the future work.

5 CONCLUSION AND FUTURE WORK

In this work, we study the task of response generation in open-domain, multi-turn dialog systems. We present a novel deep scheme to deepen and widen topics in conversation. The scheme first segments the utterances and extracts the keywords from the context. Subsequently, the context and the keywords are fed into three parallel channels, namely, global, wide, and deep channels. The global channel encodes the context into an embedding vector, the wide channel predicts the wider keywords and the deep channel selects the deeper keywords from the contextual keywords. Ultimately, the model adopts an attention mechanism to weigh the context and keywords before feeding them into the RNN decoder that is used to generate a response. Extensive experiments were conducted on two datasets. By analyzing the results, we can draw the following conclusions: 1) The wide and deep channels encourage the diversity and informativeness of the generated responses; 2) The wide channel is able to transfer the topic but the relevance to the context may drop if the wider keywords dominate the decoder for response generation. Thereby, the deep channel is essential in the scheme.

As future work, we will shed light on the logical and semantic consistency between the responses and the historical contexts.

ACKNOWLEDGMENTS

This work is supported by the National Basic Research Program of China (973 Program), No.: 2015CB352502; National Natural Science Foundation of China, No.: 61772310, No.: 61702300, and No.: 61702302; and the Project of Thousand Youth Talents 2016.

REFERENCES

[1] James F. Allen, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. 1996. A Robust System for Natural Spoken Dialogue. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*. ACL, 62–70.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473* (2014).
[3] Jimmy Lei Ba. Diederik P. Kingma. 2015. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2015).
[4] Warren R. Greiff. 1998. A Theory of Term Weighting Based on Exploratory Data Analysis. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 11–19.
[5] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*. ACL, 110–119.
[6] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1192–1202.
[7] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings*

of the International Joint Conference on Natural Language Processing. ACL, 986–995.

[8] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2122–2132.

[9] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards Micro-video Understanding by Joint Sequential-Sparse Modeling. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 970–978.

[10] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL, 285–294.

[11] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2017. Coherent Dialogue with Attention-Based Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 3252–3258.

[12] L. Nie, M. Wang, Y. Gao, Z. J. Zha, and T. S. Chua. 2013. Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information. *IEEE Transactions on Multimedia* 15, 2 (2013), 426–441.

[13] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua. 2015. Disease Inference from Health-Related Questions via Sparse Deep Learning. *IEEE Transactions on Knowledge and Data Engineering* 27, 8 (2015), 2107–2119.

[14] Liqiang Nie, Yi-Liang Zhao, Xiangyu Wang, Jialie Shen, and Tat-Seng Chua. 2014. Learning to Recommend Descriptive Tags for Questions in Social Forums. *ACM Trans. Inf. Syst.* 32, 1 (2014), 5:1–5:23.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*. ACL, 311–318.

[16] Volha Petukhova, Martin Gropp, Dietrich Klakow, Anna Schmidt, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, and John Dines. 2014. The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. In *Proceedings of International Conference on Language Resources and Evaluation*. ELRA, 252–258.

[17] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven Response Generation in Social Media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 583–593.

[18] Thomas Roelleke. 2003. A Frequency-based and a Poisson-based Definition of the Probability of Being Informative. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM, 227–234.

[19] Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. ACL, 438–449.

[20] Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 3288–3294.

[21] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 3776–3784.

[22] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 3295–3301.

[23] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Natural Language Processing*. ACL, 1577–1586.

[24] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*. ACL, 196–205.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the Neural Information Processing Systems Conference on Neural Information Processing Systems*. MIT Press, 3104–3112.

[26] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 935–945.

[27] Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-Based Deep Matching of Short Texts. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 1354–1361.

[28] Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Blac. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL Conference on Discourse and Dialogue*. SIGDIAL, 404–413.

[29] Jason D. Williams and Geoffrey Zweig. 2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269* (2016).

[30] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting TF-IDF Term Weights As Making Relevance Decisions. *ACM Transactions on Information System* 26, 3 (2008), 13:1–13:37.

[31] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 496–505.

[32] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 3351–3357.

[33] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 55–64.

[34] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 685–694.

[35] Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with Intention for a Neural Network Conversation Model. *arXiv preprint arXiv:1510.08565* (2015).