Attention and Graph Matching Network for Retrieval-Based Dialogue System with Domain Knowledge

1st Xu Li
Computer Science and Technology
Heilongjiang University
Harbin, China
2181401@s.hlju.edu.cn

2nd Jinghua Zhu*
Computer Science and Technology
Heilongjiang University
Harbin, China
corresponding author: zhujinghua@hlju.edu.cn

***Abstract*—Building an effective and friendly human-machine dialogue system is one of the major challenges in Artificial Intelligence. This work proposes a new model named Graph and Attention Matching Network (AGMN) for response selection in retrieval-based dialogue system. AGMN model consists of two parts: cross attention mechanism and knowledge representation extractor. Specifically, the cross attention mechanism is exploited to obtain the dual representation from context and response words because these representations can provide the useful matching information for determining whether the next utterance is suitable response or not. Besides, the domain knowledge relationships which are extracted from Linux manuals are incorporated into the word representation by graph attention mechanism. Experimental results on Ubuntu Dialogue Corpus showed that both cross attention mechanism and domain knowledge can contribute to the performance of response selection and the AGMN model proposed in this paper outperforms the state-of-art approaches.**

***Index Terms*—dialogue system, domain knowledge, graph attention.**

I. Introduction

Recently, building an efficient dialogue system for human beings and machines is attracting more and more attention from industry and academia. According to how machine gives the response, the dialogue system can be divided into two categories. One is generating the response word by word freely which is called generative dialogue systems [1] [2] and the other is retrieving the response from a set of candidate responses named retrieval-based dialogue systems [3] [4]. Although generative dialogue systems can produce responses by imitating human beings, they suffer from shortness and generality of responses [5] [6]. By contrast, retrieval-based dialogue systems are superior to generative dialogue systems because they can generate coherent and syntactically responses and they have mature industry products such as social bot XiaoIce from Microsoft [7]. And the example is shown in Table 1. Therefore, in this work, we only focus on the retrieval-based dialogue systems.

In retrieval-based dialogue systems, a key problem is how to evaluate the matching degree between the conversation context which consists of a series of history utterances and candidate responses. Low et al. [3] performed neural networks called Dual Encoder (DE) for multi-turn response selection by encoding all history utterances and candidate responses with a Long Short-term Memory (LSTM) [8]. The DE evaluates the matching degree for each candidate response and the same context based on the context and response encodings. Recently, some advance models have been applied to retrieval-based dialogue system by encoding context and response with the general idea and utilized embedding approaches [9] - [11]. However, all the above methods fail to keep logical consistence in long context scenario for selecting a proper response. Recent research finds that incorporating the domain knowledge is beneficial for dialogue system in domain specific conversation scenario.

In this paper, we propose a new architecture of neural network for multi-turn response selection which is an extension architecture presented by Low et al. [3]. The contributions of our work are three-folds:

- Our AGMN model is effective for capturing the overall relationships including not only the semantic relationships but also the utterance relationships between context and response words.
- We propose a method to incorporate the domain knowledge relationships between domain words into the neural network for domain specific conversation.
- The empirical evaluations on public multi-turn dialogue corpus shows that our model outperforms state-of-art methods for multi-turn response selection.

II. Related Work

Building an intelligent dialogue system can be divided into two categories. Given the context including all history utterances, the first category applied encoder-decoder architecture to generate the response freely which named retrieval-based based systems [12] - [14]. And the other category selects a response from a set of candidate responses that called retrieval-based systems [15] [16]. To begin with, researchers assumes all input information as a single message [17] [18] . Next, approaches are adopted by researchers that utilize the history

TABLE I
AN EXAMPLE OF MULTI-TURN DIALOGUE WITH DOMAIN WORDS (UBUNTU OPERATING SYSTEM RELATED) IN ITALICS.

Context	
Utterance 1:	Hi, I need to install *php4-dev* and *php5-dev* to solve dependencies. I'm using 16.04 lts and I can't seem to find *php4-dev* any suggestions.
Utterance 2:	What about *php5?*
Utterance 3:	I found that.
Utterance 4:	Well *apt-cache* search did.
Utterance 5:	So you solved it.
Utterance 6:	Nope I need to install *php4-dev*. I can find php5 and php5-dev but not *php4-dev*.
Response	
	Better if you upgrade to *php5-dev* this is to solve dependencies of *pecl solr*.

information of the conversation. Dual LSTM model are applied to encode context and response respectively [3] . A multi-view matching model [9] improve the response selection for multi-turn conversation with an utterance view and word view. A deep neural network [10] that concatenate all utterances as queries and then match with responses. More recently, an and sequential matching network [4] and deep attention matching neural network are proposed by [19].

Another category researches on incorporating knowledge in conversation system is growing rapidly such as task-oriented dialogue systems [20] - [22] and open-domain dialogue systems [23] - [25]. At the beginning, Low et al. [3] incorporated unstructured domain knowledge into dialogue system. Then, Xu et al. [26] applied the loosely-structured domain knowledge into the neural network with a gating mechanism. Recently, the structured knowledge graph is adopted by Zhou [27] to generate knowledge-aware responses with graph attentions [28] . In this paper, we only focus on the retrieval-based method. Different from previous models, we use the graph attention mechanism to enhance the domain knowledge relationships between the words in context and response for multi-turn response selection.

III. MODEL

We use the dual encoder model as a basic structure of our model. Attention mechanism encoding at sentence level is extended in our model for capturing the overall relationship between context and response. Besides, in our model, we use graph attention network for incorporating the knowledge connection of domain words in every utterance. Both the two components are described detailed in the following sections. And the architecture of our model is given in Figure 1.

A. Attetion Encoding

As aforementioned in introduction, context and response are encoded separately with the same RNN network at word level in the dual encoder model. However, the relationship between utterances is not incorporated which determines whether the next utterance is proper or not for the current utterance. So, we apply cross attention mechanism and the Gated Recurrent Unit (GRU) architecture to encoding every utterance in the context and response into utterance vector representation for determining the response is the suitable or not.

Furthermore, we will describe the mechanism to construct the utterance vector representation formally. Denote the context c where all utterances are concatenated consisting of a long sequence of words as $c = (c_1, c_2, \ldots, c_m)$ where m is the word number of the context. And the concatenated response is denoted as $r = (r_1, r_2, \ldots, r_n)$ similarly. Given the sequence of the context or response, we use the word embedding matrix $e \in \mathbb{R}^{d\times|V|}$ to convert the c and r to vector sequences respectively:

$$c^e = (e(c_1), e(c_2), \ldots e(c_m)) \tag{1}$$

$$r^e = (e(r_1), e(r_2), \ldots e(r_n)) \tag{2}$$

where d is the dimension of the word embedding and $|V|$ is the vocabulary size. To construct the contextual meanings for each word, the GRU is used to encode the word embedding to get the sequence hidden states c^s and r^s:

$$c_i^s = \mathrm{GRU}(c_i^e, i) \tag{3}$$

$$r_j^s = \mathrm{GRU}\left(r_j^e, j\right) \tag{4}$$

where i is the i-th word in the context and j means the j-th word in the response similarly.

The word relevance in semantic representation between the context and response can provide the useful matching information for determining whether the next utterance is the suitable response or not. Therefore, we use the cross-attention mechanism to calculate the word relevant degree which is denoted as:

$$e_{ij} = (c_i^s)^T r_j^s \tag{5}$$

After that, the word representation including contextual meaning between context and response is computed by the word relevant degree. For a word in the context, its relevance is calculated with the response hidden status by attention mechanism:

$$\bar{c}_i = \sum_{j=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} r_j^s \tag{6}$$

$$\bar{r}_j = \sum_{i=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{kj})} c_i^s \tag{7}$$

To obtain the contextual representation of every word in response, the context hidden status is computed by attention mechanism similarly in equation 5. By reinforcing the semantic relevance in context and response, we model enhanced representations as follows:

$$c_i' = [c_i^s; \bar{c}_i; c_i^s - \bar{c}_i; c_i^s \odot \bar{c}_i] \tag{8}$$

$$r_j' = \left[r_j^s; \bar{r}_j; r_j^s - \bar{r}_j; r_j^s \odot \bar{r}_j\right] \tag{9}$$

where difference and element-wise operation is used between the hidden status and word relevant representation . The enhanced representation is performed by concatenating all above vectors.

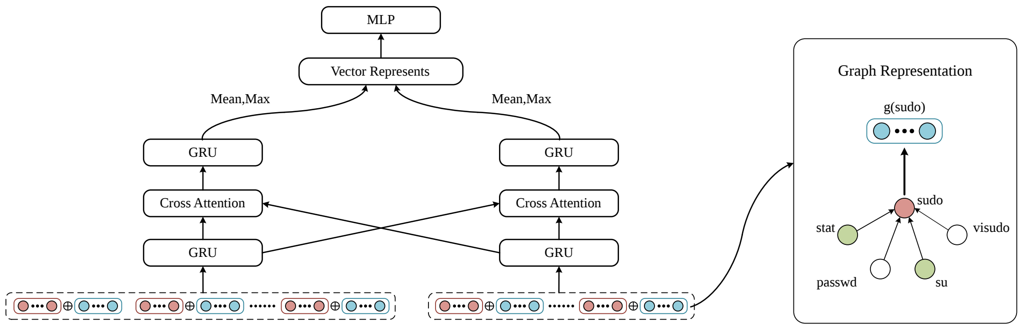

Fig. 1. The architecture of our model.

Since the utterance appears sequentially, we explore the utterance representation from the word relevant representation for building dependency relationships between each utterance. Another GRU is used for collecting all enhanced representations to generate the utterance representation as equation (8):

$$c_i^u = \mathrm{GRU}\left(c_i', i\right) \tag{10}$$

$$r_j^u = \mathrm{GRU}\left(r_j', j\right) \tag{11}$$

Though we use the same sequence structure to encode sentence information, the function of the GRU is most different from the contextual meaning encoding layer. Some identical word relevant vectors are learned for model to compute the matching degree in the utterance level. In this way, all hidden status vectors from the GRU is selected by mean and max operations and concatenated altogether to a dense vector for obtaining the overall relationship representation as follows:

$$m = [c_{\text{mean}}^u; c_{\text{max}}^u; r_{\text{mean}}^u; r_{\text{max}}^u] \tag{12}$$

The final vector representation is then fed into a multi-layer perceptron (MLP) classifier with softmax output. Finally, the MLP classifier generates a probability that indicates the overall matching degree between the next response and the current context.

B. Incorporating Domain Knowledge

The domain knowledge is essential for human beings to answer the professional problem. Analogously, domain knowledge can take the language understanding ability to the dialogue model which can find some words relationships between utterances even cross the utterances besides the word semantic relevance. To use the domain knowledge, we build the data including command triples from Linux Manual Pages. The command triple is denoted by $R = (u, r, v)$ where u is the command concept node, v is the neighbor command concept node and both nodes are connected by relation r. For a word in context or response utterance, if it appears in some command triples, we firstly retrieve knowledge triples. Then, the word is extended its meaning with the neighbor concept nodes by the graph network. Otherwise, if the word in the utterance is not in command triple, we only use its common meanings for constructing graph vector representations.

More formally, the concept representation in the domain knowledge is constructed by a series of triples, $G(x) = \{T_1, T_2, \ldots, T_n\}$ where T_i has the same concept node u but different neighbor concept v and the graph representation of the concept $g(x)$ can be calculated by graph attention mechanism as:

$$g(x) = \sum_{i=1}^{n} \alpha_{T_i} \left[u_i^e; v_i^e\right] \tag{13}$$

$$\alpha_{T_i} = \frac{\exp\left(\beta_{T_i}\right)}{\sum_{j=1}^{n} \exp\left(\beta_{T_i}\right)} \tag{14}$$

$$\beta_{T_i} = \mathrm{Relu}\left(\left[(u_i^e)^T W v_i^e\right]\right) \tag{15}$$

where $(u_i, r_i, v_i) = R_i \in G(x)$ is the i-th triple in the dataset. We also use word embedding method to convert the concept to vector representation $u_i^e = e\left(u_i\right)$, $r_i^e = e\left(r_i\right)$ and $v_i^e = e\left(v_i\right)$. Regarding the word not included in the command triples, we simply set its knowledge representation $g(x)$ to zero and only use common word embedding. After that, we add the word embedding and graph representation in context or response as follows:

$$e'\left(c_i\right) = e\left(c_i\right) + g\left(c_i\right) \tag{16}$$

$$e'\left(r_j\right) = e\left(r_j\right) + g\left(r_j\right) \tag{17}$$

In this scenario, the final word representation calculated by equation (1) and (2) in each utterance is updated as the following equations:

$$c^e = \left(e'\left(c_1\right), e'\left(c_2\right), \ldots e'\left(c_m\right)\right) \tag{18}$$

$$r^e = \left(e'\left(r_1\right), e'\left(r_2\right), \ldots e'\left(r_n\right)\right) \tag{19}$$

Intuitively, relationships between each concept which can cross all utterances including the current context utterance and response utterance are captured by graph attention network.

TABLE II
EVALUATION RESULTS OF OUR MODELS AND OTHER APPROACHES ON UBUNTU DIALOGUE CORPUS.

Model	R_2@1	R_{10}@1	R_{10}@2	R_{10}@5
DE-RNN (Kadlec et al. 2015)	0.768	0.403	0.547	0.819
DE-CNN (Kadlec et al. 2016)	0.848	0.549	0.684	0.896
DE-LSTM (Kadlec et al. 2015)	0.901	0.638	0.784	0.949
DE-BiLSTM (Kadlec et al. 2015)	0.895	0.630	0.780	0.944
Multi-View (Zhou et al. 2016)	0.908	0.662	0.801	0.951
DL2R (Yan et al. 2016)	0.899	0.626	0.783	0.944
r-LSTM (Xu et al. 2016)	0.889	0.649	0.857	0.932
MV-LSTM (Wan et al. 2016)	0.906	0.653	0.804	0.946
Match-LSTM (Wang et al. 2016)	0.904	0.653	0.799	0.944
QA-LSTM (Tan eta l. 2016)	0.903	0.633	0.789	0.943
SMN (Wu et al. 2017)	0.926	0.726	0.847	0.961
DUA (Zhang et al. 2018)	-	0.757	0.868	0.961
DAM (Zhou et al. 2018)	0.938	0.767	0.874	0.969
AGMN (Ours)	**0.944**	**0.783**	**0.883**	**0.973**

IV. EXPERIMENTS

A. Datasets

The Ubuntu Dialogue Corpus (UDC) introduced by Lowe et al. [3] is the most used domain-specific and multi-turn dialogue dataset. The conversation in the Freenode Internet Relay Chat (IRC) network about the Ubuntu topic specific chat rooms are extracted. In general, one user proposes a problem and a potential solution is given by experienced users. The conversation among these users often stops when the problem has been addressed. At some time, the conversation may continue to be conducted but the content is not related problem.

Based on the Ubuntu Dialogue Corpus, Wu et al. [4] further processed the corpus and provided all needed vocabulary. All numbers, URLs and system paths were replaced with special holders in the processing work. Besides, for obtaining the knowledge relationships among the domain specific concept, we resort to the Linux manual pages. Every page corresponds one command concept and contains the different items about the concept such as name, synopsis, descriptions, see also and so on. For our experiments, we extract these items from Linux manual pages into domain concept triples. This is additional processing performed by us. The Ubuntu Dialogue Corpus datasets consists of 1000,000 training triples, 500,000 validation triples and 500,000 test triples. The triple is composed of context, response and the label. The triple with label $y = 1$ is the positive sample if the response is suitable for the context and with label $y = 0$ is the negative sample if the response doesn t match the context. For the training dataset, one half samples are positive samples and the other half samples are negative. On the contrary, for every sample in the validation and test dataset, only one ground-truth response fits the context and nine negative response are not suitable for the same context. Therefore, the ratio between the positive and negative samples is 1:9 in the validation datasets and test datasets that makes us evaluate the model with Recall@k metrics.

B. Experimental Setting

In our experiments, we use binary cross-entropy loss between the golden label and the predicted output to train the model. Instead of initializing the word embedding metric with a normal distribution, we use the fastText [29] to pre-train the word embedding as done by Wu et al. [4]. Meanwhile, the dimension of the pre-training word embedding is set to 300. For discarding the information far from the last context, we set the maximum length of the concatenated sentences to 300. The maximum length of the response is set to 150 similarly. Owing to the limit of model parameters and the GPU memory, we had to choose the batch size of 32. The Adam [30] with the initial learning rate 0.0001 is used to optimize the model parameters. At the same time, we use the dropout method with the rate 0.3 after the GAT layer and GRU layer. We set the maximum of the training epochs to 15 because it is enough for our model to achieve the best performance. The training process will be stopped if the recall metric in the validation dataset does not increase. Finally, the model is evaluated in the test dataset with the best validation recall.

V. RESULTS AND DISCUSSION

As aforementioned in section 4.1, we use the information retrieval metric Recall@k denoted as R_n@k. The metric R_n@k in our experiments is the fraction of examples for the correct response which is under the k best result of n candidate responses. And these candidate responses are ranked by predicted distributions of the model. Specifically, R_{10}@1, R_{10}@2, R_{10}@5 and R_2@1 are used in our experiments.

A. Results

We refer to our model as Attention and Graph Matching Network(AGMN), which is compared with the previous models tested on the Ubuntu Dialogue Corpus: the DE models originally researched by Low et al. [3] with various encoder architectures such as recurrent neural network (DE-RNN), convolutional neural network (DE-CNN), LSTM (DE-LSTM) and bi-directional LSTM (DE-biLSTM); hierarchy based architectures for matching context and response named DL2R [10] and Multi-View [9]; sequence-based models MV-LSTM

[31]and Match-LSTM [32]; architectures for processing utterances in the context respectively named SMN [4] and DUA [33]; and we also choose the recent model DAM [34]improved with the Transformer architecture [35]as a baseline of our model.

The recall results of AGMN and baselines on the UDC datasets are reported in Table 2. In order to achieve such results, the model parameters were fixed after being trained 23 hours. The AGMN outperforms all other models used as baselines. Specifically, comparing the best baseline model DAM, our model achieves a relative improvement 1.6% and 0.9% corresponding to absolute improvement 0.02 and 0.01 with respect to R_{10}@1 and R_{10}@2 metrics. We also observe the modest improvements of 0.6% (0.006) and 0.4% (0.004) for R_2@1 and R_{10}@5 metrics. Comparing the best baseline model (DAM), our results are identical better with p-value $<$ 0.01 for these four metrics.

B. Ablations

In this work, we focus on the effectiveness of different components in the AGMN model. All ablation test results are presented in Table 3. The knowledge graph representation component is removed firstly and metrics of R_{10}@1 and R_{10}@2 are degraded to 0.775 and 0.880. Furthermore, if we continue to discard the components of utterance representation, R_{10}@1 and R_{10}@2 metrics are decreased to 0.771 and 0.877. Comparing the reduction without these two components, we can observe that 0.8 of removing the knowledge graph representation is greater than 0.4 of dropout the utterance representation. Similarly, in metric R_{10}@2, the reduction 0.4 is also worse than 0.2 with respect to the knowledge graph representation and utterance representation. Therefore, it is reasonable that the knowledge graph representation is more important than the utterance representation for our model AGMN. Meanwhile, we found that the reduction of utterance representation is equal to the knowledge graph representation reduction for R_{10}@5 metric which may result from more candidate responses for model to select, it also demonstrates that both two components contribute to the performance of the AGMN model.

TABLE III
ABLATION TEST WITH REMOVING THE KNOWLEDGE GRAPH REPRESENTATION AND THE UTTERANCE REPRESENTATION COMPONENTS OF THE AGMN MODEL.

Model	R_{10}@1	R_{10}@2	R_{10}@5
AGMN	0.783	0.883	0.973
- knowledge graph	0.775	0.879	0.972
- utterance representation	0.771	0.877	0.971

C. Visualization

In this section, we focus on what are learned between the context and response by our model. Thus, the visualization of cross attention weight computed by Equation 5 is shown in Figure 2. We select an example from the test dataset that is consist of right response and wrong response with respect to the same context. In Figure 2, the relevance only calculated by word embedding before cross attention layer is shown in the subgraph top right. Then, the relevance computed after cross attention layer is presented in the subgraph at the top left. Meanwhile, the relevance of the same context but wrong response is shown in two bottom subgraphs similarly. We can see that the word “gutsy” in the last utterance which is a history ubuntu release selected some relevant words “feisty”, “package” and “gutsy” in the response to derive matching features of context and response. By contrast, comparing to the weight before cross attention layer, matching features of these three words is approximate to its neighbor words that are difficult for model to extract the key information between the context and response. For two subgraphs at the bottom, we observed that the attention weight between the context and wrong response is more ambiguous than two top subgraphs. Consequently, the visualization demonstrates that

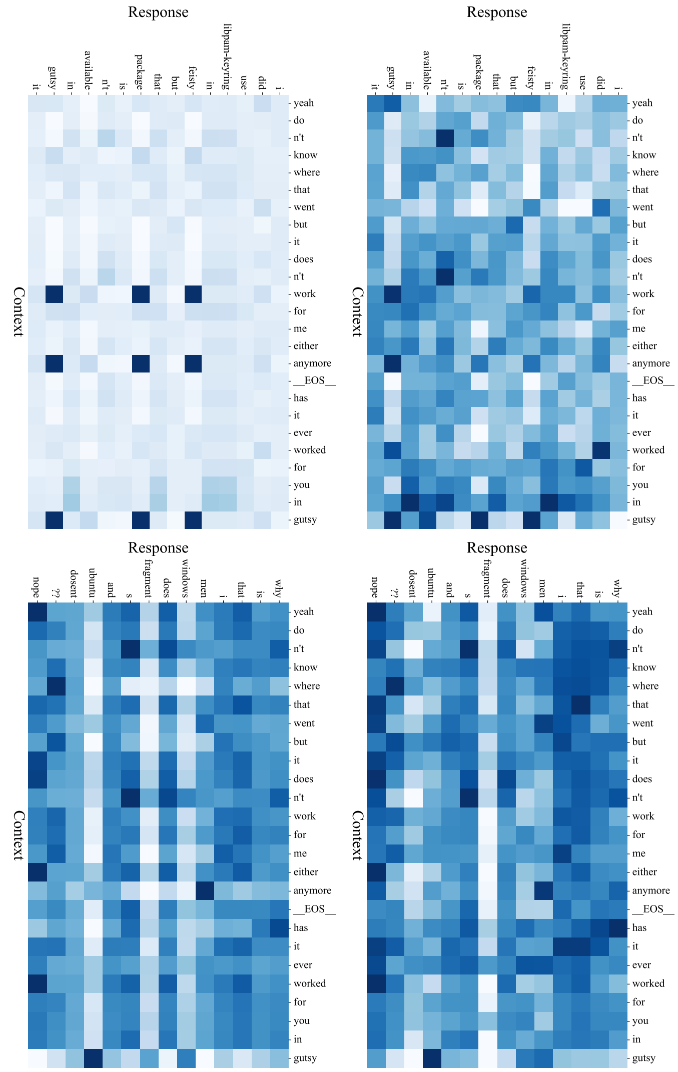

Fig. 2. Cross attention visualization between the context and response. Top and bottom subgraphs represent the right and wrong response corresponding to the same context. Each colored unit represents the relevance between two words. The deeper color grids correspond to the higher words relevance.

cross attention layer in our model is effective for the AGMN model to select semantic relevant words between context and response.

VI. Conclusions

In this paper, a new model named Attention and Graph Matching Network (AGMN) is proposed for multi-turn response selection. We demonstrated that incorporating the knowledge and cross attention mechanism are contribute to the performance of our model. Domain knowledge further strengthen associations between domain field words but not all words viewed as equal in other methods. And cross attention mechanism used in our model for capturing identical features between the context and response for the last matching result. Experimental results suggested that AGMN model derived from the dual encoder architecture outperformed all previous methods on the Ubuntu Dialogue Corpus. In the future, we will research our model by utilizing more knowledge associations in domain field not only domain words in the multi-turn response selection problem.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61602159, 61100048). Jinghua Zhu is the corresponding author.

References

[1] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y.F. Ji, M. Mitchell, J.Y. Nie, J.F. 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*, Denver, Colorado, USA, pp.196-205, 2015.

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

[3] R. Lowe, N. Pow, I. Serban, 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*, Prague, Czech Republic, pp.285-294, 2015.

[4] Y. Wu, W. Wu, C. Xing, M. Zhou and Z.J. Li, "Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots", Vancouver, Canada, pp.496-505, 2017.

[5] J. Li, M. Galley, C. Brockett, J.F. 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*, San Diego California, USA, pp.110-119, 2016.

[6] J.F. Gao, M. Galley and L.H. Li, "Neural Approaches to Conversational AI", in *Proceedings of the 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Ann Arbor, MI, USA, pp.1371-1374, 2018.

[7] H.Y. Shum, X.D. He, D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots", Frontiers of Information Technology & Electronic Engineering, Vol.19, No.1, pp.10-26, 2018.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neural Computation, Vol.9, No.8, pp.1735-1780, 1997.

[9] X.Y. Zhou, D.X. Dong, H. Wu, S.Q. Zhao, D.H 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*, Austin, Texas, USA, pp.372-381, 2016.

[10] R. Yan, Y.P. 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*, Pisa, Italy, pp.55-64, 2016.

[11] G.Z. An, M. Shafiee and D. Shamsi, "Improving Retrieval Modeling Using Cross Convolution Networks And Multi Frequency Word Embedding", *arXiv preprint*, arXiv: 1802.05373, 2018.

[12] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models ", in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, pp.3376-3384, 2016.

[13] J.W. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley and J.F. Gao, "Deep Reinforcement Learning for Dialogue Generation", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, pp.1192-1202, 2016.

[14] N. Dziri, E. Kamalloo, K.W. Mathewson and O.R. Zaiane, "Topic aware neural response generation", in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, pp.3351-3357, 2017.

[15] Z.C. Ji, Z.D. Lu and H. Li, "An Information Retrieval Approach to Short Text Conversation", *arXiv preprint*, arXiv: 1408.6988, 2014.

[16] M.X. Wang, Z.D. Lu, H. Li and Q. Liu, "Syntax-Based Deep Matching of Short Texts", in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, pp. 1354-1361, 2015.

[17] H. Wang, Z.D. Lu, H. Li and E.H. Chen, "A Dataset for Research on Short-Text Conversations", in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp.935-945, 2013.

[18] B.T. Hu, Z.D. Lu, H. Li and Q.C. Chen, "Convolutional Neural Network Architectures for Matching Natural Language Sentences", in *Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, pp.2042-2050, 2014.

[19] H. Zhou, M.L. Huang, T.Y. Zhang, X.Y. Zhu and B. Li, "Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory ", in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, pp.730-739, 2018.

[20] A. Madotto, C.S. Wu and P. Fung, "Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, pp.1468-1748, 2018.

[21] C.S. Wu, R. Socher, C.M. Xiong, "Global-to-local Memory Pointer Networks for Task-Oriented Dialogue", in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA, 2019.

[22] J.Q. He, B. Wang, M.M. Fu, T.Q. Yang and X.M Zhao, "Hierarchical Attention and Knowledge Matching Networks With Information Enhancement for End-to-End Task-Oriented Dialog Systems", IEEE Access, Vol.7, pp.18871-18883, 2019.

[23] S.D. Han, J. Bang, S.H. Ryu, and G.G. Lee, "Exploiting knowledge base to generate responses for natural language dialog listening agents", in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic, pp.129-133, 2015.

[24] N. Moghe, S. Arora, S. Banerjee and M.M. Khapra, "Towards Exploiting Background Knowledge for Building Conversation Systems", in *Proceedingsof the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp.2322-2332, 2018.

[25] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli and J. Weston, "Wizard of Wikipedia: Knowledge-Powered Conversational Agents", in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA, 2019.

[26] Z. Xu, B.Q. Liu, B.X. Wang, C.J. Sun and X.L. Wang, "Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM", in *Proceedings of the 2017 International Joint Conference on Neural Networks*, Anchorage, AK, USA, pp.3506-3513, 2017.

[27] H. Zhou, M.L. Huang, T.Y. Zhang, X.Y. Zhu and B. Liu, "Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory ", in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, pp.730-739, 2018.

[28] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, "Graph Attention Networks", *arXiv preprint*, arXiv: 1710.10903, 2017.

[29] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching Word Vectors with Subword Information", *arXiv preprint*, arXiv:1607.04606, 2016.

[30] D.P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, in *Proceedings of 3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015.

[31] S.X. Wan, Y.Y. Lan, J. Xu, J.F. Guo, L. Pang and X.Q. Cheng, “Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN”, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, NY, USA, pp.2922-2928, 2016.

[32] S.H. Wang and J. Jiang, “Learning Natural Language Inference with LSTM”, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego California, USA, pp.1442-1451, 2016.

[33] Z.S. Zhang, J.T Li, P.F. Zhu, H. Zhao and G.S. Liu, “Modeling Multi-turn Conversation with Deep Utterance Aggregation”, in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp.3740-3752, 2018.

[34] X.Y. Zhou, L. Li, D.X. Dong, Y. Liu, Y. Chen, W.X. Zhao, D.H. Yu and H. Wu, “ Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network”, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, pp.1118-1127, 2018.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, “Attention is All you Need”, in *Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, Long Beach, CA, pp.5998-6008, 2017.