

A Knowledge Enhanced Generative Conversational Service Agent

Yinong Long¹, Jianan Wang², Zhen Xu³, Zongsheng Wang⁴, Baoxun Wang⁴, Zhuoran Wang⁴

¹Central South University, Changsha, China

²Shanghai Jiao Tong University, Shanghai, China

³Harbin Institute of Technology, Harbin, China

⁴Tricorn (Beijing) Technology Co., Ltd, Beijing, China

¹yinonglong@csu.edu.cn, ²wangjianan@sjtu.edu.cn

³zxu@insun.hit.edu.cn, ⁴{wangzongsheng, wangbaoxun, wangzhuoran}@trio.ai

Abstract

In this paper, we describe our attempt at generating natural and informative responses for customer service oriented dialog incorporating external knowledge. Our system captures external knowledge for a given dialog using a search engine. Then a knowledge enhanced sequence-to-sequence framework is designed to model multi-turn dialogs on external knowledge conditionally. The whole system is trained and evaluated on the DSCT6 official dataset obtained from Twitter, based on a list of predefined customer service accounts. The experimental results demonstrate that the proposed system is promising to generate informative responses.

Index Terms: response generation, conversational agent, external knowledge

1. Introduction

Dialog systems such as Siri¹, Cortana², and XiaoIce³ have been widely used to assist users to achieve specific goals or to meet users' communication requirements. The fundamental purpose of most dialog systems is to convey information to human beings by providing query-relevant, natural and informative responses. However, previous dialog systems built by end-to-end training of neural networks tend to suffer from the so-called "safe response" problem [1], i.e. these systems often generate non-informative responses, such as "I am sorry to hear that" and "I think so". Customer service oriented agents particularly suffer from this problem, since the responses from these agents are expected to be more relevant to users' queries and topics of the corresponding customer service [2].

Although a large amount of works had been conducted to model the relevance between queries and responses [2, 3, 4, 5, 6], only a few researches have been done exploring response generation conditioning on conversation related external knowledge [7, 8], which includes topic clues to the corresponding customer service. Since Dialog systems often benefit greatly from external knowledge, the DSTC6-Task2 [9] competition promotes research aiming at combining conversational dialogs with task-oriented knowledge using unstructured data. Figure 1 illustrates one dialog example of DSTC6-Task2.

To generate natural and informative response to a user input in a given dialog history for DSTC6-Task2, we propose a response generation system incorporating external knowledge. The proposed system obtains the dialog-related external knowledge using a knowledge extractor, after that it takes account of

this knowledge in the later response generation process.

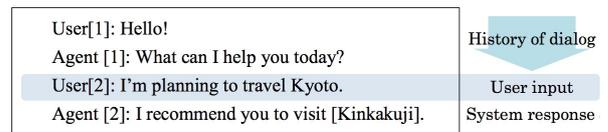


Figure 1: The Dialog Example of DSTC6-Task2⁴

2. System Description

2.1. System Overview

We named the proposed system Search Engine Enhanced Response Generator (SEARG), of which Figure 2 illustrates the overall framework. The whole system consists of a knowledge extractor, a knowledge encoder, and a knowledge enhanced sequence-to-sequence model including a query encoder and a decoder for response generation. The knowledge extractor is designed to obtain dialog-related knowledge from the web in a given dialog history including the latest query. Then, the Convolutional Neural Network (CNN) based knowledge encoder is introduced to extract features of the obtained knowledge. In the knowledge enhanced sequence-to-sequence model, the encoder adopts the Long Short-Term Memory (LSTM) to project the dialog history and the user input into a real-valued vector, followed by the decoder which generates response based on the dialog history vector and the knowledge features.

2.2. The Knowledge Extractor

The proposed system aims for combining dialog system with task-oriented external knowledge, therefore the first step of our system is to extract related knowledge of a given conversation. A knowledge extractor is designed to obtain such knowledge from the web using a search engine, and the whole process is divided into two phases: keyword extraction and knowledge aggregation. Concretely, the knowledge extractor firstly gains keywords of the dialog history by the keyword extraction component, and then captures related knowledge from the web conditioning on the keywords during the knowledge aggregation phase using a search engine. We explain each of the phases in detail as follows.

Keyword Extraction: The keyword extraction part is designed to gain keywords for searching operation in the knowl-

¹<https://www.apple.com/ios/siri/>

²<https://www.microsoft.com/en-us/windows/cortana>

³<https://msxiaoice.com/>

⁴<http://workshop.colips.org/dstc6/proposals/End.To.End.Dialog-MERL.pdf>

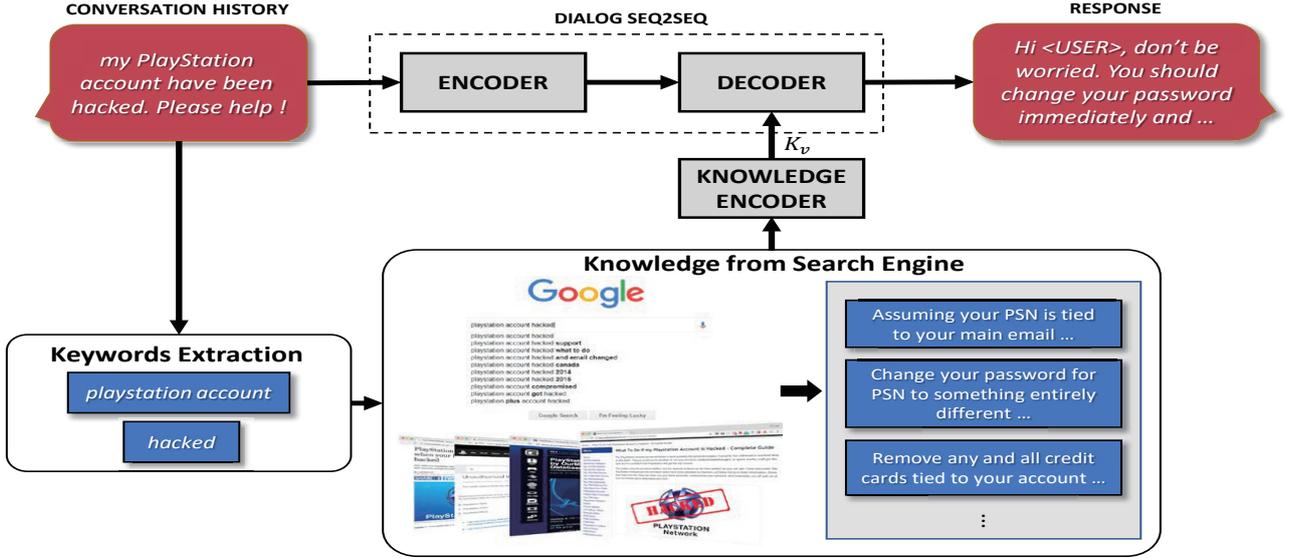


Figure 2: The Architecture of Our Dialog System.

edge aggregation process. There are two reasons of taking keywords rather than sentences in a given dialog as the input of the knowledge aggregation process. One of the reasons is that there are always more than one sentence in each given dialog history. To obtain knowledge covering the whole dialog history, the system should search each user utterance of the dialog, and then integrate all the search results. But this integration process is intractable and more complex than the original goal of our system that is to incorporate external knowledge into dialog systems. The other reason is that search engines always take keywords as their input, since keywords are more useful to gain valuable information according to the theory of information retrieval. Overall, the keyword extraction phase is of great importance in the process of knowledge extraction.

For the purpose of obtaining keywords automatically and rapidly, the proposed system adopts the Rapid Automatic Keywords Extraction (RAKE)[10] to extract keywords from the given dialog history. The RAKE algorithm used in our system can be summarized as follows: (1) The input text is first split into several pieces by delimiters and stop words, and then the RAKE takes each piece as a candidate keyword except delimiters and stop words; (2) The score of each word is estimated by the ratio of degree to frequency ($deg(w)/freq(w)$), where the $freq(w)$ is the word frequency and the $deg(w)$ is the word co-occurrence frequency with other words; (3) The score of each candidate keyword is defined as the sum of its member words' ratio of degree to frequency. Then we can gain keywords of the input text using these scores.

Knowledge Aggregation: After gaining keywords from the last phase, a knowledge aggregation procedure is designed to extract external knowledge using a search engine. In the knowledge aggregation procedure, a search operation taking keywords as inputs is firstly conducted to capture the dialog-related information by a search engine. Since thousands of results would be recalled by the search engine and most of them are unstructured text, taking all these information as external knowledge of the given dialog is unwise. Consequently, our system take the top-K most relevant results (ranked by the search engine) as the dialog-related information, and then sub-

sequently distills knowledge from these information by extracting the abstract of the each search result provided by the search engine. we assume that ideally the top-N results cover most related knowledge of keywords, and the abstract of the search result summarizes the keyword-related information well. After that, several paragraphs including knowledge are obtained as shown in Figure 2, and will be taken as the inputs of the knowledge encoder. For further discussions, we introduce $\{p^1, p^2, \dots, p^n\}$, ($1 \leq n \leq N$) to denote the knowledge paragraphs of one dialog obtained in knowledge aggregation phase. In this paper, the N is set to 10, which means the knowledge paragraph number of each dialog ups to 10.

2.3. The Knowledge Encoder

To incorporate external knowledge obtained in subsection 2.2 into the procedure of response generation, we design a knowledge encoder to project these knowledge information into a real-value vector. The structure of the knowledge encoder is illustrated in Figure 3. The knowledge encoder firstly adopts CNN to encode knowledge-related paragraphs into dense vectors respectively. Since the numbers of knowledge-related vectors are different in each dialog as mentioned in subsection 2.2, a max-pooling operation is used on these paragraph vectors to encapsulates the knowledge information of each dialog into condensed vectors K_v of same size. In this case, this knowledge vector can be used as the input of the decoder to deliver external knowledge into the response generation procedure. The process of the knowledge encoder can be formulated as follows:

$$p_v^i = Conv(p^i), (1 \leq i \leq n) \quad (1)$$

$$K_v = max_pool(\{p_v^1, p_v^2, \dots, p_v^n\}) \quad (2)$$

2.4. The Knowledge Enhanced Sequence-to-Sequence Model

The final goal of our system is to generate informative response incorporating external knowledge. For this purpose, our system extends the simple sequence-to-sequence (Seq2Seq) model [11, 12] by augmenting the input with the knowledge vector

Table 1: Automatic Evaluation Based on Word-overlap Metrics

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
BaseLine	0.5905	0.3695	0.2435	0.1619	0.2041	0.3598	0.0825
2LDTM	0.5982	0.3729	0.2441	0.1602	0.2016	0.3606	0.0782
2LSTM	0.5988	0.3858	0.2607	0.1779	0.2085	0.3829	0.0978
GWGM	0.6054	0.3850	0.2578	0.1741	0.2024	0.3703	0.0994
SEARG	0.5478	0.3026	0.1768	0.1092	0.1731	0.3201	0.0702

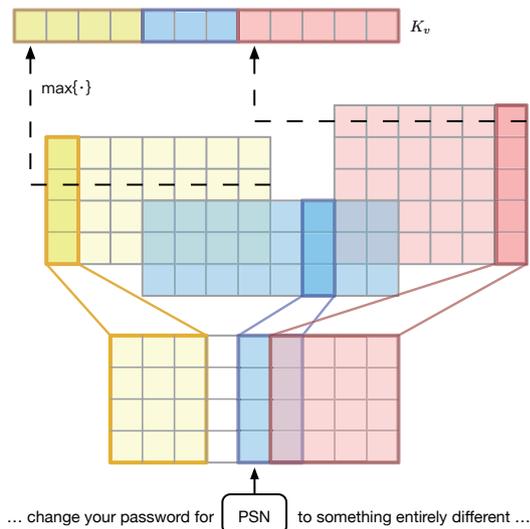


Figure 3: The Architecture of Our Knowledge Encoder.

K_v , so as to take account of the knowledge in the procedure of response generation, into the decoder of the Seq2Seq model. In our proposed knowledge enhanced Seq2Seq model, both the encoder and the decoder are composed of LSTM. The dialog (including the dialog history and the user input) and the latest response are notated as $q = \{u_1, s_1, u_2, s_2, \dots, u_k\}$ and $r = \{w_{r,1}, \dots, w_{r,i}, \dots, w_{r,m}\}$ respectively. For a given dialog (q, r) , q is firstly encoded into a vector representation q_v through the hierarchical encoder-decoder process [13], which is actually the last hidden state of u_k . Then, by taking q_v as the hidden initialization of the decoder, the probability of each word in r can be estimated. The decoding procedure can be formulated as follows:

$$x_i = [E(w_{r,i}), K_v] \quad (3)$$

$$h_i = LSTM(x_i, h_{i-1}) \quad (4)$$

$$p(w_{r,i+1}) = f(h_i) \quad (5)$$

where E denotes the word embedding, and f denotes the word projection layer mapping the hidden state to word distribution.

3. Experimental Setup

3.1. Dataset

The training, test and development datasets of DSTC6-Task2 are built using the official scripts⁵, we downloaded and formatted Twitter⁶ data based on the predefined customer service accounts. We finally collected 887,419, 107,434 and 2000 dialog

samples for training, development and test, respectively. Note that since the necessary preprocessing steps have been done using official tools, these data can be directly used to train our models.

3.2. Baselines and Training Strategies

In addition to the official baseline, this paper also adopts three models (2LDTM, 2LSTM and GWGM) using attention augmented Seq2Seq as baselines. The difference between these baseline models is the training strategy, and we named them these names just for abbreviation. GWGM was only trained on the official training dataset, while 2LDTM was tuned by diversified data obtained by removing some non-informative responses in the training dataset and 2LSTM was tuned by trial dataset. Note that all the baselines are trained without external knowledge.

According to the response distribution analysis of the training dataset, about 25%-30% responses start with “We’re or I’m sorry to hear that”, which provide little information to the user query. The statistical models trained by such dataset tend to produce non-informative responses. To achieve our goal to generate natural and informative responses, we tuned our model using an additional dataset built by removing dialogs whose responses in the high-frequency response list from official training dataset.

Same as that in the official baseline model, the vocabulary size used in our model is set to 20,000, since it covers the majority of frequent words in the training dialogs. The word embedding and LSTM hidden sizes are set to 100 and 256 respectively. We use a two layers bidirectional LSTM as the encoder of our system to obtain better representations of dialog histories and user inputs. The knowledge encoder is trained jointly with the encoder-decoder architecture. In the generation procedure, we adopt the sentence length penalty into the beam search algorithm to alleviate non-informative short response generation problem.

4. Results and Discussions

4.1. Automatic Evaluation

The proposed dialog systems are first automatically evaluated by several word-overlap metrics, including BLEU, METEOR, ROUGE and CIDEr [14]. Then, metrics based on word embedding representing the semantics of words are introduced to further evaluate our systems.

4.1.1. Word-overlap Metrics

From Table 1, it can be seen that in word-overlap metrics, our non-external knowledge models (including 2LDTM, 2LSTM

⁵<https://github.com/dialogtekgreek/DSTC6-End-to-End-Conversation-Modeling>

⁶<https://twitter.com/>

Table 2: Responses Generated by SEARG and the Baseline

<p>U: purchased refrigerator 4 yrs ago & already not working . error code rf bad fan motor . poor product and poor customer svc exp .</p> <p>REF: we hate to see this . what is the model number , please ?</p> <p>SEARG: we are the experts in fixing household appliances . have a look ! <URL></p> <p>Baseline: we 're sorry to hear this . please follow & dm contact info along with model / serial # if you are in need of assistance .</p>
<p>U: hey @theqarena i forgot my password and security questions for my account is there someone i can email to get my password</p> <p>REF: saw your tweet ! is this for <URL> login or for @cavsteamshop login ? thanks !</p> <p>SEARG: hi there . just to confirm , did you try resetting your password w / the steps on this page : <URL> ? keep us posted .</p> <p>Baseline: we 'll be in touch via . dm shortly if that 's okay . please follow us if you haven 't already</p>
<p>U: i know our neighbors have power and we don 't and i have 2 sick kids with asthma and this heat isnt helping .</p> <p>REF: sometimes individual residences left in dark when others have power . occurs when on different circuit or transformers are out</p> <p>SEARG: hi <USER> , we are aware of the outage and are working to restore service . we are working to restore power as quickly as possible .</p> <p>Baseline: we apologize for the inconvenience . please dm your address and we will be happy to look into this for you .</p>
<p>U: hello team , when do we get ota naugat update for moto x play ? moto declered it almost two months ago .</p> <p>REF: we don 't provide release dates but we 're working to roll it out to all eligible devices globally as quickly as possible .</p> <p>SEARG: hi <USER> , we will be launching our products globally in the coming months (may-june) . stay tuned to our social media pages for updates .</p> <p>Baseline: hi <USER> , i 'm sorry to hear that . please dm us your email address so we can look into this for you . thanks !</p>

and GWGM) outperform the official baseline, while the knowledge enhanced model SEARG gains relative low performance. This phenomenon is attributed to the following reasons: (a) 2LDTM, 2LSTM and GWGM are attached with attention mechanism, which enhances the models' capacity to fit the training data. (b) In SEARG, the external knowledge stimulates the decoder to generate informative responses, which deviate from the references in the official dataset. In this case, SEARG obtains lower scores in the word-overlap metrics.

Table 3: Automatic Evaluation Based on Embedding Metrics

Method	Skip	Average	Extrema	Greedy
BaseLine	0.6380	0.9132	0.6073	0.7590
2LDTM	0.6474	0.9074	0.6031	0.7567
2LSTM	0.6259	0.9201	0.6106	0.7683
GWGM	0.6348	0.8985	0.6000	0.7573
SEARG	0.6132	0.8977	0.5870	0.7420

4.1.2. Embedding Based Metrics

Furthermore, we adopt embedding-based metrics Skip Thought (Skip), Embedding Average (Average), Vector Extrema (Extrema) and Greedy Matching (Greedy) to make more feasible evaluation of our systems [15]. These embedding-based metrics firstly approximate response embeddings based on word embedding, and then calculate cosine similarity between the candidate and target embeddings [16]. The evaluation results of our models and the official baseline model in embedding-based metrics are illustrated in Table 3.

It can be observed that scores of all the models have less discrepancy from the embedding metric. This observation demonstrates that all the models even SEARG is able to generate

query-relevant responses. It is because that comparing to word-overlap metrics, embedding metrics tend to capture semantics beyond word-overlap. Despite of the robust of embedding metrics, our knowledge enhanced model(SEARG) still gains relative low score. We ascribe this observation to the infusion of external knowledge in the generation process tends to bias the decoder to generate some informative words, which is consistent with our observation in subsection 4.2.

4.1.3. Human Evaluation

Table 4: Human Evaluation Result

Method	HumanRating
BaseLine	3.3638
2LDTM	3.4504
2LSTM	3.5239
GWGM	3.5082
SEARG	3.3919

The human rating scores in Table 4 represent the average score for every system. From official organizer's announcement, the 10 different humans rated all systems responses in the range of 1 (very poor) to 5 (very good) where system responses were displayed in random order with the dialog context for the human subjects.

In the view of human evaluation, our own models (2LDTM, 2LSTM, GWGM and SEARG) all perform better than the official baseline model. Noticing that although the SEARG didn't match the performance of official baseline model on word-overlap metrics and embedding based metrics, it still beat official baseline model by human evaluation. This consequence and some selected good cases (see Table 2) indicate that SEARG could be a promising model for generating informative re-

Table 5: Responses Generated by SEARG and 2LSTM

<p>U: at what places do i get discounts S: are you referring to our purchase payback reward program or some other type of discount or reward ? thanks . U: the rewards program REF: cashback offers for this program can be found in the rewards center online . let me know if you need help locating this . SEARG: hi <USER> , we do not have any price codes available at this time . please keep an eye on our social media pages for updates .</p>
<p>U: a year ago today i passed my driving test REF: time flies alice ! what are you driving now ? SEARG: congratulations ! welcome to the audi family . we hope you have a great rest of your day !</p>

sponses in some way and taking use of external knowledges. However, it still has a little distance to the our other models (2LDTM, 2LSTM and GWGM) without considering external knowledge. The reasons for the lower human-evaluation score of the proposed model will be discussed in detail in Subsection 4.2. Especially, our proposed baselines (2LDTM, 2LSTM and GWGM) have indeed obtain satisfying results on both automatic and human evaluations, which indicates that the complicated memory architecture still has potential for response generation, meanwhile, since the models without any external knowledge have got higher score in the perspective of human evaluation, it should be further discussed whether the released training set contains enough external knowledge requirement or not.

4.2. Case Study

Table 2 shows the generated responses by our model and official baseline model, and it demonstrates the ability of our knowledge enhanced model to generate informative responses intuitively. It can be seen that the responses generated by SEARG contain informative content words, and also provide valuable information to the given users' input, with the help of external knowledge in the generation process. In contrast, most responses generated by the official baseline as shown in Table 2 are generic and non-informative, as many of these responses are likely to start with "we are(re) sorry to" or "i 'm sorry to" and end with "please dm us" or "please follow". Therefore, our knowledge enhanced method is promising to generate informative responses using external knowledge.

Nevertheless, the human evaluation score of the proposed SEARG does not fit our observation, the reason of which can be indicated by the cases given in Table 5. According to this table, responses generated by the SEARG model indeed contain the external knowledge from the search engine, such as about "price codes" in the first case and "audi family" in the second case, and these knowledges are apparently relevant to the topic of the given dialog, which can not be generated by any complicated models without external knowledge. However, such keywords are not strictly correct for composing the responses in the corresponding dialog situations, and we infer this should be the reason for our model not obtaining good human evaluation results, since the evaluation concerns more about the correctness of the generated replies. This problem is attributed to that knowledges used in this paper are abstracts obtained from open domain sources using the search engine, and the search engine is a weak knowledge source in fact, since even human beings have to distinguish the useful information from the noise when using search engines. Thus, according to these two bad cases generated by SEARG (see Table 5) and our own experi-

ment process, we found that there are many noises in extracted knowledges, such as the timeliness and the query matching issues, which make generated responses unsuitable to the corresponding queries, even they are topic relevant with more information contained.

So as to make our proposed architecture perform in the expected way to generate both correct and informative results, it is necessary to explore and adopt more accurate knowledge sources, e.g., the ones from the restricted domains, and our future work will be carried out along this direction.

5. Conclusions

In this paper, we proposed a dialog system which incorporates external knowledge into the procedure of response generation in custom service dialogs. The experimental results show that our system tend to generate more informative responses. However, the proposed system may encounter a little problem, that is the knowledge obtained using search engine may mix with noise information sometime. In this case, the external knowledge plays a negative role in the procedure of response generation. In the future, we will explore how to obtain credible knowledge of a given dialog history and approaches to improve the usage of these knowledge in response generation.

6. References

- [1] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2016, pp. 110–119.
- [2] X. Chen, W. Wei, W. Yu, L. Jie, H. Yalou, Z. Ming, and M. Wei-Ying, "Topic aware neural response generation," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 3351–3357.
- [3] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," in *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2015, pp. 196–205.
- [4] B. Wu, B. Wang, and H. Xue, "Ranking responses oriented to conversational relevance in chat-bots," in *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, 2016, pp. 652–662.
- [5] L. Ryan, Thomas, P. Nissan, S. Iulian, Vlad, C. Laurent, L. Chia-Wei, and P. Joelle, "Training end-to-end dialogue systems with the ubuntu dialogue corpus," *Dialogue & Discourse*, vol. 8, no. 1, pp. 31–65, 2017.

- [6] 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 (EMNLP)*, 2016, pp. 372–381.
- [7] M. Ghazvininejad, C. Brockett, M. W. Chang, B. Dolan, J. Gao, W. T. Yih, and M. Galley, "A knowledge-grounded neural conversation model," 2017.
- [8] T. Young, E. Cambria, I. Chaturvedi, M. Huang, H. Zhou, and S. Biswas, "Augmenting end-to-end dialog systems with commonsense knowledge," 2017.
- [9] C. Hori and T. Hori, "End-to-end conversation modeling track in dstc6," *arXiv:1706.07440*, 2017.
- [10] S. J. Rose, D. W. Engel, N. O. Cramer, and W. E. Cowley, "Automatic keyword extraction from individual documents," *Text Mining: Applications and Theory*, 2010.
- [11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 3104–3112.
- [12] K. Cho, B. van Merriënboer, 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)*, 2014, pp. 1724–1734.
- [13] S. Iulian, Vlad, S. Alessandro, B. Yoshua, C. C. Aaron, and P. Joelle, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3776–3784.
- [14] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [15] C.-W. Liu, R. Lowe, I. V. 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 (EMNLP)*, 2016, pp. 2122–2132.
- [16] S. Sharma, L. El Asri, H. Schulz, and J. Zumer, "Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation," *CoRR*, vol. abs/1706.09799, 2017. [Online]. Available: <http://arxiv.org/abs/1706.09799>