Enhancing Multi-turn Dialogue Modeling with Intent Information for E-commerce Customer Service

Ruixue Liu*[1], Meng Chen(✉)*[1], Hang Liu[1], Lei Shen[2], Yang Song[1], Xiaodong He[1]

[1] JD AI, Beijing, China
[2] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences
liuruixue@jd.com, chenmeng20@jd.com, liuhang55@jd.com,
shenlei17z@ict.ac.cn, songyang23@jd.com, xiaodong.he@jd.com

Abstract. Nowadays, it is a heated topic for many industries to build intelligent conversational bots for customer service. A critical solution to these dialogue systems is to understand the diverse and changing intents of customers accurately. However, few studies have focused on the intent information due to the lack of large-scale dialogue corpus with intent labelled. In this paper, we propose to leverage intent information to enhance multi-turn dialogue modeling. First, we construct a large-scale Chinese multi-turn E-commerce conversation corpus with intent labelled, namely **E-IntentConv**, which covers 289 fine-grained intents in after-sales domain. Specifically, we utilize the attention mechanism to extract Intent Description Words (IDW) for representing each intent explicitly. Then, based on E-IntentConv, we propose to integrate intent information for both retrieval-based model and generation-based model to verify its effectiveness for multi-turn dialogue modeling. Experimental results show that extra intent information is useful for improving both response selection and generation tasks.

Keywords: Multi-turn Dialogue Modeling · Large-scale Dialogue Corpus · Intent Information.

1 Introduction

With the rapid development of artificial intelligence, many conversational bots have been built for the purpose of customer service, especially in E-commerce. Building a human-like dialogue agent has lots of benefits for the E-commerce customer service industry. It can not only improve the working efficiency for the professional customer service staffs, but also save amount of labor costs for the E-commerce company.

Existing approaches to building end-to-end conversational systems mainly concentrate on retrieval-based models [19, 25, 26], generative-based models [5, 6, 13] and hybrid models [14, 21]. Impressive progress has been made on modeling the context [19, 24], leveraging external knowledge [6], and promoting the language diversity of response [5, 7]. However, previous works did not pay enough attention on the user intent in conversations. There are two major issues: 1) existing dialogue datasets are deficient

*Equal Contribution
(✉) Corresponding Author

for intent-augmented dialogue modeling. There is basically no large-scale multi-turn dialogue corpus with intent labelled. 2) Most existing neural conversation models do not explicitly model user intent in conversations. More research needs to be made to understand the user intent and to develop intent-augmented response selection and generation models, which is exactly the target of this paper.

To tackle above two obstacles, we firstly construct a large-scale multi-turn dialogue corpus with intent labelled, namely **E-IntentConv**, which consists of more than 1 million conversations about after-sales topics between users and customer service staffs in E-commerce scenario. Nearly three hundreds fine-grained intents are summarized and provided based on real business of E-commerce customer service for understanding the user intents accurately. To represent the user intents explicitly, we also extract tens of words (Intent Description Words, denoted as IDW) to depict each intent with attention mechanism. Then, we propose a novel intent-aware response ranking method and an intent-augmented neural response generator to leverage the extra intent information for dialogue modeling. For the response ranking model, an ensemble modeling paradigm with three intent-aware features is well-designed. For the neural dialogue generator, an extra intent classifier is integrated into the decoder to enhance the intent expression. Experimental results validate that both response selection and generation tasks can be improved with our proposed models. To the best of our knowledge, it is the first work to build dialogue systems with intents of large-scale multi-turn conversations.

To sum up, our contributions are two-folds: 1) we collect a very large-scale multi-turn E-commerce dialogue corpus with intent labelled, and we release it to the NLP community for free[3]. 2) we propose two intent-aware dialogue models and design experiments to prove that both response selection and response generation tasks can be improved by incorporating intent information.

2 Related Work

There are two lines of research that are closely related to and motivate our work: multi-turn dialogue corpus, and end-to-end dialogue system.

2.1 Multi-turn Dialogue Corpus

The research on chatbots and dialogue systems has kept active for decades. The growth of this field has been consistently supported by the development of new datasets and novel approaches. Especially for popular deep learning based approaches, large scale of training corpus in real scenario becomes decisive. More recently, some researchers construct dialogue datasets from social media networks (e.g., Twitter Dialogue Corpus [11] and Chinese Weibo dataset [18]), or online forums (e.g., Chinese Douban dataset [19] and Ubuntu Dialogue Corpus [9]). Despite of massive number of utterances included in these datasets, they are different from real scenario conversations. Posts and replies are informal, single-turn or short-term related. In terms of dialogue datasets from real

[3] We have the license to redistribute this corpus and third-party users can download it from our official website for research purpose: http://jddc.jd.com.

scenario, ECD [25] corpus is collected from real E-commerce scenario, which keeps the nature of conversation flow and the bi-turn information for real conversation. However, the ECD corpus provides little annotated information for each query, such as intent information of customers. Compared to ECD [25], our E-IntentConv corpus provides extra intent information and description words extracted by an interpretative way to help understanding those intents. These intents contain beneficial information for dialogue systems to understand queries under complicated after-sales circumstances.

2.2 End-to-end Dialogue System

With the development of dialogue datasets, a number of data-driven dialogue systems are designed, divided into retrieval-based models [19, 25, 26], generative-based models [5–7, 13, 24] and hybrid models [14, 21]. For retrieval-based models [19, 25, 26], various text matching techniques are proposed to catch the semantic relevance between context and response. But they ignore the constraint between query and response in the dimension of intent. For generative-based models, modeling dialogue context [24], leveraging external knowledge [6] and promoting language diversity for response [5, 7] have become hot research topics, but all of them neglect the importance of understanding user's intents. Different from previous work, we propose two intent-enhanced dialogue models and verify their effectiveness on E-IntentConv corpus.

3 E-IntentConv Construction

The construction of E-IntentConv includes data collection, intent annotation, and intent description words extraction, which illustrates how we collect the large-scale multi-turn dialogue corpus, how we annotate the intent for each query, and how we mine the intent description words.

Table 1. Overview of E-IntentConv.

Total number of sessions	1,134,487
Total number of utterances	22,495,387
Total number of words	253,523,129
Average number of intents per session	4
Average number of turns per session	20
Average number of words per utterance	11

3.1 Dataset Collection and Statistics

We collect the conversations between customers and customer service staffs from a leading E-commerce website[4] in China. After crawling, we de-duplicate the raw data,

[4] http://www.jd.com

desensitize and anonymize the private information based on very detailed rules. For example, we replace all numbers, order ids, names, addresses with special token <NUM>, <ORDER-ID>, <NAME>, <ADDRESS> correspondingly. Then, we adopt Jieba[5] toolkit to perform Chinese word segmentation. Table 1 summarizes the main information about the dataset. It's observed that the amount of this data is large enough to support building the current mainstream data-driven dialogue models. Meanwhile, the average number of intents per session is 4, which indicates that the customers' intents are constantly changing as the conversations go on. Thus, understanding these changing intents accurately is critical to solve the customers' problems.

Table 2. An example from E-IntentConv corpus. Best viewed in color.

q_1	可以帮我改下订单的地址吗？ (Could you help me change the address of the order?)
r_1	同一市内可以联系配送员直接修改的哦。 (You can contact the delivery staff directly if the two addresses are in the same city.)
q_2	不在同一个城市，现在地址是上海，但是我明天要回安徽。 (Not the same city. The current address is Shanghai, but I am going to Anhui tomorrow.)
r_2	抱歉，地址在不同城市不能操作的，只能建议您重新下单哦。 (Sorry, you cannot change the address to a different city. In this case, we suggest you place a new order.)
q_3	那我取消订单的话退款多久到账呢？ (How long does it take for the refund to arrive if I cancel the order?)
r_3	微信零钱1个工作日内到账，储蓄卡1-7个工作日内到账，信用卡1-15个工作日内到账的哦！ (For Wechat change, it arrives in 1 working day. For debit card, it arrives in 1-7 working days. And for credit card, it arrives in 1-15 working days.)
q_4	为什么不能改地址，你们这也太不方便了。 (Why can not change the delivery address? That is too inconvenient.)
r_4	非常抱歉，我们物流还有待完善呢。 (I'm sorry. Our logistics system needs to be improved.)
q_5	这也太麻烦了，我还急着用呢。 (That is too troublesome, I'm in a hurry.)
r_5	非常抱歉！如果是我的话我也会很着急的，我们会改进的！ (I'm so sorry! If I were you, I would feel the same. We will do our best to improve it!)
q_6	行吧。 (Fine.)
r_6	谢谢您的理解！还有什么能帮到您的吗？ (Thanks for your understanding! What else can I do for you?)

The E-IntentConv illustrates the complexity of conversations in E-commerce. It covers different kinds of dialogues including: a) task completion: e.g. changing the order address, b) knowledge-based question answering: e.g. asking the warranty and return policy or asking how to use the product, and c) feeling connection with the user: e.g. actively responding to the user's complains and soothe his/her emotion. Therefore, it's totally different from previous dialogue datasets. Table 2 shows a typical example in the corpus, q_1,q_2 in blue refer to task completion, q_3 in red is knowledge-based question and answering while q_4,q_5 in purple require feeling connection.

[5] https://github.com/fxsjy/jieba

3.2 Intent Annotation

As the number of queries in the dataset is huge, it's infeasible to annotate the intents for all queries manually. Here we use a high-quality intent classifier to label the intent for each query automatically. The classifier contains totally 289 classes which are summarized based on the real business of E-commerce customer service. The classes are fine-grained and helpful to understand the user's intents under the after-sales circumstances. To train the intent classifier, we sample 600,000 instances from the corpus and annotate them manually under the user intent taxonomy. Each instance consists of at most three consecutive utterances from users (eg. q_1,q_2,q_3 in Table 2). The former two utterances are context and the last one is the query. Three professional customer service staffs are invited to annotate the training data. The inter-agreement score is 0.723[6] and the final label is decided by voting strategy. At last, totally 578,127 training samples are annotated manually under the user intent taxonomy. Considering the challenging of short text classification and the language understanding in dialogue, we train the intent classifier with Hierarchical Attention Network (HAN) [22] so each utterance in a training sample is weighted differently. The classification accuracy on the test set reaches 93% and the Macro-F1 score is 84.22%, which indicates the predicted intents for the user queries are reliable. By this way, we label the intents for all user queries automatically. Fig. 1 shows the distribution of top 15 intents in E-IntentConv.

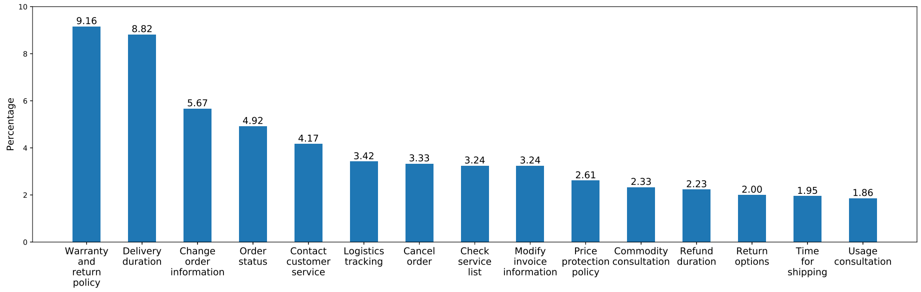

Fig. 1. Distribution of top 15 intents in E-IntentConv.

3.3 Intent Description Words

Intuitively, intent is a high-level abstraction, so how to make the system understand the intent becomes important. Here, we try to depict the abstract intent with tens of explicit words, which can be seen as descriptions or explanations for each intent. We call those words as **Intent Description Words** (IDW). IDWs should be better the feature words from the perspective of classifier, so they can represent the exact meaning of corresponding intent. Specifically, we utilize the attention mechanism of HAN model

[6] The Fleiss' Kappa score is calculated, and above 0.2 is acceptable

[22] to extract those feature words, which is interpretative as Fig. 2 shows. The words with the top K highest attention weights in each training instance are picked out as IDW candidates. After dealing with all training instances, we collect a set of IDWs for each intent. We rank those words by frequency, and filter the stop words, then top N words are chosen as final IDWs for each intent. IDWs are also provided along with our dataset[7].

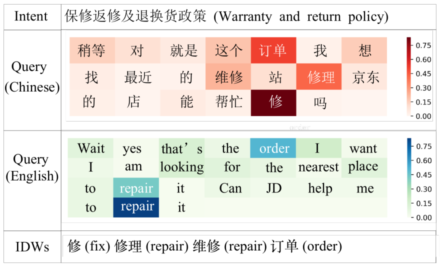

Fig. 2. Visualization of attention weights in HAN. Words with higher attention weights are assigned with deeper colors. English translation is provided for understanding.

4 Methods

Based on the intent-labelled corpus above, in this section, we want to validate the effectiveness of intent information for dialogue modeling. For retrieval-based model, we propose a novel ensemble modeling paradigm and design three intent-aware features, to facilitate the response ranking task. For generation-based model, we integrate a special intent classifier into the encoder-decoder framework to promote the expression of user intent.

4.1 Retrieval-based Model

Existing retrieval-based models mainly focus on calculating the semantic similarity between context and response, and treat the response selection problem as a ranking problem. Both traditional learning to rank models [14, 20] and neural based matching models (e.g., DAM [26], ESIM [2], and BERT [4]) have been proposed. Here, we propose an ensemble modeling approach with GBDT [23] to combine the advantages of existing neural based matching models. By taking BM25 [3], DAM, ESIM and BERT as input features, we build a regression model (denoted as **Ensemble**) to predict final similarity. Meanwhile, we also design several new features to catch the intent consistency between query and response candidate as follows:

[7] Empirically, we set K to 4 and N to 50 in this work

IntentFeat 1: we represent the intent by averaging the word embeddings of all IDWs, and represent the response candidate by averaging all word embeddings in the response utterance, then the cosine similarity of two representations is calculated as the first intent-aware feature.

IntentFeat 2: we calculate the ratio of how many words in the response candidate are 'covered' by the IDWs. A word is considered as 'covered' by the intent as long as the similarity between it and any word in IDWs is greater than the threshold t[8]. Here, the similarity score is also calculated based on the word embeddings.

IntentFeat 3: for each word in the response candidate, the largest similarity score between it and all the IDWs is chosen as the final similarity score. Then we average all similarity scores for all words to represent the similarity between the response candidate and the intent. By adding the extra intent features, we denoted the model as **Ensemble-IDW**.

4.2 Generation-based Model

Popular generation-based models are based on the standard seq2seq model [16] with attention mechanism [1]. To better utilize these IDWs and make the generated response Y more informative and consistent with context C, we propose a model named **S2S-IDW**. First, we represent each intent z as the average word embeddings of corresponding IDWs. Then z is concatenated to each decoder input and used to update the decoder hidden state.

Inspired by [15], we use a intent classifier to enhance the intent expression, and the classification loss is defined as:

$$L_{CLS} = -p(z)\log q(z|Y) \tag{1}$$

$$q(z|Y) = f(W \cdot \frac{1}{T} \sum\nolimits_{t=1}^{T} \mathrm{Ewe}(t; C, z)) \tag{2}$$

where $f(\cdot)$ is the softmax activation function, $p(z)$ is a one-hot vector that represents the desired intent distribution for an instance, and $\mathrm{Ewe}(t; C, z)$ is the expected word embedding at time step t, which is calculated as:

$$\mathrm{Ewe}(t; C, z) = \sum\nolimits_{y_t \in V} p(y_t) \cdot \mathrm{Emb}(y_t) \tag{3}$$

that is, for each decoding step t, we enumerate all possible words in the vocabulary V. Finally, the loss function can be written as:

$$L = L_{CE} + \lambda L_{CLS} \tag{4}$$

where λ is a hyper-parameter that controls the trade-off between cross entropy loss L_{CE} and classification loss L_{CLS}.

[8] Empirically, we set t to 0.6 in our experiments

5 Experiments

In this section, we perform extensive experiments on the E-IntentConv dataset. We firstly introduce the dataset preparation, then show the experimental setup and results for the response selection and generation tasks.

5.1 Dataset Preparation

We first divide the around 1 million conversation sessions into training, validation and testing set with the ratio of 8:1:1. Then we construct I-R pairs from each set into the $\{I, R\} = \{q_1, r_1, q_2, r_2, ..., q_i, r_i, Q, R\}$ format, where $I = \{C, Q\}$ stands for input, C is the dialogue context, Q is the last query, and R represents the response. i is set to 5 so the most recent five rounds of dialogue are kept as context. Finally, there are 2,852,620 I-R pairs for training, 176,600 and 177,412 I-R pairs for validation and testing respectively.

5.2 Response Selection

Following [19] and [25], we randomly sample negative responses for above training, validation, and testing sets. The ratio for positive and negative samples is 1:1 for training set, and 1:9 for validation/testing set. Totally 1 million I-R pairs (Train Set I) are sampled for training the neural matching models. 200k I-R pairs (Train Set II) are sampled for training the ensemble models. It's worth noting that there is no overlap between Train Set I and II. Following [9], recall at position k in n candidates (denoted $R_{10}@1$, $R_{10}@2$, $R_{10}@5$) are taken as evaluation metrics.

Our baselines are as follows:

1) BM25: The standard retrieval technique BM25 [3] is used to rank candidates.

2) DAM: The Deep Attention Matching Network proposed by [26], which matches a response with its multi-turn context using dependency information learned by Transformers.

3) ESIM: The Enhanced Sequential Inference Model proposed by [2], which matches local and composition information by designing a sequential LSTM inference model.

4) BERT: We fine-tune the BERT [4] with I-R pairs and use the predicted similarity score to rank all response candidates.

From Table 3, it can see that, the proposed **Ensemble** model outperforms existing neural semantic matching models significantly, which proves that the different features are complementary to each other. By adding the three intent-aware features, the performance of **Ensemble-IDW** is further improved, which indicates their helpfulness in the response selection task.

5.3 Response Generation

Totally 500k I-R pairs (Train Set III) are sampled to train various generation models. We choose BLEU [10] score, Rouge [8] score, and Distinct-1/2 [7] (denoted as Dist-1/2) to evaluate the quality and the diversity of the generated responses. Here are the

Table 3. The comparison of retrieval-based models. ‡ means statistically significant difference over baseline with $p < 0.05$.

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
BM25	0.3618	0.4982	0.7570
DAM	0.7634	0.9067	0.9883
BERT	0.7968	0.9249	0.9926
ESIM	0.8114	0.9330	0.9941
Ensemble	0.8628	0.9598	0.9974
Ensemble-IDW	**0.8657**‡	**0.9610**‡	**0.9975**

Table 4. Case study for retrieval-based models. The user intent is *return policy*.

q_1	手机拆封了还可以七天无理由退吗 (My phone is unpacked, can I still return it within 7 days without reason?)
r_1	这个包装一般是没有关系的 (It doesn't matter if the phone is unpacked.)
q_2	开机激活了，有影响吗？如果退货，邮费是要我自己出的是吧 (I've activated my phone. Is that OK ? Do I need to pay the postage if I return it.)
r_2	支持七天无理由退货，无理由退货的话是需要您承担运费的哦 (You can return it within 7 days without reason but you need to pay the postage.)
Q	好的多少运费 (Ok, how much is the postage)
IDW	运费, 问题, 退货, 货运费, 邮费, 承担, 费用, 快递费, 收费, 退换货 (postage, problems, returns, freight charges, postage charges, responsible for, fees, express charges, charges, return goods)
Ensemble	激活后不支持七天无理由的亲，影响二次销售呢 (You can't return the phone within 7 days without reason if you have already activated it. Because that will influence its secondary sale.)
Ensemble-IDW	这个运费一般是12元，售后审核了会联系告知您的呢。 (The postage is usually 12 RMB. My colleague for after-sale service will contact you after verification.)

baselines:
1) S2S-Attn: The classical seq2seq model with attention mechanism [13], which uses Bi-LSTM as encoder and LSTM as decoder.
2) TF: The Transformer model with multi-head self-attention mechanism [17]. For previous two models, utterances in the context are concatenated into one sentence.
3) HRED: The hierarchical recurrent encoder-decoder model proposed by [12] to consider the hierarchical structure in a dialogue session.

Table 5 shows the results of generation-based models. **HRED** has the best performance on *diversity* however it performs the worst in BLEU and Rouge-L metrics. We argue that the hierarchical modeling of context keeps more information meanwhile brings more noise. The S2S models have the highest *similarity* scores (BLEU and Rouge-L). With the support of IDWs, **S2S-IDW** improves all metrics compared with **S2S-Attn** (Sign Test, with p-value < 0.05).

Table 5. The comparison of generation-based models.

Model	BLEU	Rouge-L	Dist-1	Dist-2
HRED	7.43	16.18	**0.37%**	**3.60%**
TF	8.08	18.03	0.20%	1.21%
S2S-Attn	11.02	22.33	0.12%	0.85%
S2S-IDW	**11.67**	**23.03**	0.24%	1.83%

Table 6. Case study for generation-based models. The user intent is *refund period*.

q_1	这个包我要退款重新下单。咨询订单号ORDER-ID (I want a refund for this bag, and I would like to place a new order for it.)
r_1	亲爱的请您稍等, 小妹会尽快处理您的问题EMOJI (Wait a moment please. I will deal with your problem as soon as possible, EMOJI.)
q_2	好的 (OK.)
r_2	后台显示您的取消申请已提交, 处理时间为2到5个工作日 (Record shows you have already applied for a refund. It will take 2-5 working days to process it.)
q_3	这个钱往哪里退 (How can I get the refund?)
r_3	原返。 (It will go back to your original payment account.)
Q	退款多久到？要用这个退款再下单。 (How long can I get the refund? I need to use the money to place a new order.)
IDW	退款, 取消, 微信, 多久, 银行卡, 还钱, 退钱, 几天, 处理, 白条 (refund, cancellation, Wechat, how long, bank card, repayment, refund, how many days, process, JD IOU)
S2S	1个工作日到账 (It will take one working day.)
S2S-IDW	退款周期储蓄卡1到7个工作日，信用卡15个工作日，微信零钱1个工作日到账，现金支付是退回京东账户余额15个工作日 (Refund period: 1-7 working days for debit card, 15 working days for credit card, 1 working day for Wechat payment, and 15 working days for cash payment)

5.4 Case Study

To further illustrate the effectiveness of intent information, we compare the responses of retrieval-based and generation-based models and in Table 4 and Table 6. We can see that, responses enhanced by intents are more informative, diverse and appropriate. Table 4 discusses phone return policy and postage charges. The **Ensemble** model selects the wrong answer due to the misleading context information of *return goods within 7 days without reason*. Meanwhile, the IDW words, *postage*, *returns* and *postage charges* help the **Ensemble-IDW** model to focus accurately on postage charges related answers instead. In Table 6, the user is enquiring information on refund period for his product. As we can see, both models can generate correct answer, however, with IDWs of *refund*, *Wechat* and *JD IOU*, the **S2S-IDW** model generates much more informative and diverse response referring to various payment methods.

6 Conclusion and Future Work

In this paper, we focus on enhancing the dialogue modeling with intent information for E-commerce customer service. To facilitate the research, we firstly construct the

E-IntentConv dataset, which not only includes large-scale, multi-turn dialogues in real scenario but also contains rich and accurate intent information. We also propose two novel dialogue models and verify the effectiveness of intents in both response selection and generation tasks. This work is a first step towards intent-augmented multi-turn dialogue modeling. The work has much limitation and much room for further improvement. For example, the dialogue dataset here is only collected from one company and the intents are also too domain-specific. And the definition of all intents can be more clear and discovered by popular topic modeling approach automatically. In the future, we will improve above aspects, enrich the dataset with more annotations, and explore more effective approaches to utilize these information.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D.: Enhanced lstm for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1657–1668 (2017)
3. Christopher D. Manning, P.R., Schütze, H.: Introduction to information retrieval. Inf. Retr. **13**(2), 192–195 (2010)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
5. Gao, X., Lee, S., Zhang, Y., Brockett, C., Galley, M., Gao, J., Dolan, B.: Jointly optimizing diversity and relevance in neural response generation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1229–1238 (2019)
6. Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M.: A knowledge-grounded neural conversation model. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
7. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: 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. pp. 110–119 (2016)
8. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
9. Lowe, R., Pow, N., Serban, I., Pineau, J.: 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. pp. 285–294 (2015)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
11. Ritter, A., Cherry, C., Dolan, W.B.: Unsupervised modeling of twitter conversations. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA. pp. 172–180 (2010)

12. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. pp. 3776–3783 (2016)
13. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1577–1586 (2015)
14. Song, Y., Li, C.T., Nie, J.Y., Zhang, M., Zhao, D., Yan, R.: An ensemble of retrieval-based and generation-based human-computer conversation systems. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 4382–4388 (2018)
15. Song, Z., Zheng, X., Liu, L., Xu, M., Huang, X.J.: Generating responses with a specific emotion in dialog. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3685–3695 (2019)
16. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. pp. 3104–3112 (2014)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
18. Wang, H., Lu, Z., Li, H., Chen, E.: A dataset for research on short-text conversations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 935–945 (2013)
19. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 496–505 (2017)
20. Yan, Z., Duan, N., Bao, J., Chen, P., Zhou, M., Li, Z., Zhou, J.: Docchat: An information retrieval approach for chatbot engines using unstructured documents. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 516–525 (2016)
21. Yang, L., Hu, J., Qiu, M., Qu, C., Gao, J., Croft, W.B., Liu, X., Shen, Y., Liu, J.: A hybrid retrieval-generation neural conversation model. arXiv: Information Retrieval (2019)
22. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 1480–1489 (2016)
23. Ye, J., Chow, J.H., Chen, J., Zheng, Z.: Stochastic gradient boosted distributed decision trees. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 2061–2064 (2009)
24. Zhang, H., Lan, Y., Pang, L., Guo, J., Cheng, X.: Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. pp. 3721–3730 (2019)
25. Zhang, Z., Li, J., Zhu, P., Zhao, H., Liu, G.: Modeling multi-turn conversation with deep utterance aggregation. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3740–3752 (2018)
26. Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W.X., Yu, D., Wu, H.: Multi-turn response selection for chatbots with deep attention matching network. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 1118–1127 (2018)