The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, Bowen Zhou

JD AI, Beijing, China
{chenmeng20, liuruixue, shenlei41, yuanshaozu, zhoujingyan3}@jd.com
{wuyouzheng1, xiaodong.he, bowen.zhou}@jd.com

Abstract

Human conversations in real scenarios are complicated and building a human-like dialogue agent is an extremely challenging task. With the rapid development of deep learning techniques, data-driven models become more and more prevalent which need a huge amount of real conversation data. In this paper, we construct a large-scale real scenario Chinese E-commerce conversation corpus, JDDC, with more than 1 million multi-turn dialogues, 20 million utterances, and 150 million words. The dataset reflects several characteristics of human-human conversations, e.g., goal-driven, and long-term dependency among the context. It also covers various dialogue types including task-oriented, chitchat and question-answering. Extra intent information and three well-annotated challenge sets are also provided. Then, we evaluate several retrieval-based and generative models to provide basic benchmark performance on JDDC corpus. And we hope JDDC can serve as an effective testbed and benefit the development of fundamental research in dialogue task.

Keywords: large-scale dataset, multi-turn dialogues, real E-commerce scenario

1. Introduction

Building a human-like conversational agent is regarded as one of the most challenging tasks in Artificial Intelligence (Turing, 2009). As real scenario human conversation is very complicated, it can be seen as a sequential, continuous, decision-making process, which relies on lots of information to make the conversation go on. For example, dialogue context, intents, external knowledge, common sense, emotions, participants’ background and personas, etc. All these could have an impact on the response in a conversation. Moreover, these uncertainties make dialogue task extremely different from traditional machine learning tasks which usually have an explicit target and clearly defined evaluation metrics.

To tackle this challenging problem, constructing a dialogue dataset is the most essential work. Especially for popular deep learning based approaches, large scale of training corpus in real scenario becomes decisive. However, existing datasets are still deficient. Datasets with structured annotations (e.g., slots and corresponding values) are often small-scale and in a limited capacity. Either traditional domain-specific ones (Allen et al., 1996; Petukhova et al., 2014) or recent multi-domain ones (Bordes et al., 2016; Dodge et al., 2015) are usually built for task-oriented dialogue systems. Another typical branch of work collects the dialogue corpus from movie subtitles, such as OpenSubtitles (Tiedemann, 2009) and Cornell (Danescu-Niculescu-Mizil and Lee, 2011), which contain long sessions (over 100 turns) and some expressions like individual monologues may be not suitable for dialogue systems. More recently, some researchers construct dialogue datasets from social media networks (e.g., Twitter Dialogue Corpus (Ritter et al., 2011) and Chinese Weibo dataset (Wang et al., 2013)), or online forums (e.g., Chinese Douban dataset (Wu et al., 2017) and Ubuntu Dialogue Corpus (Lowe et al., 2015)). Although in large scale, they are different from real scenario conversations, as posts and replies are informal, single-turn or short-term related.

In this work, we construct a large-scale multi-turn Chinese dialogue dataset, namely JDDC (Jing Dong Dialogue Corpus), with more than 1 million multi-turn dialogues, 20 million utterances, and 150 million words, which contains conversations about after-sales topics between users and customer service staff in E-commerce scenario. Different from existing datasets mentioned above, the JDDC dataset illustrates the complexity of conversations in E-commerce. Table 1 presents a typical session in the corpus which contains services including: 1) task completion: changing the order address ($q_1$-$q_2$, text in blue); 2) knowledge-based Question Answering (QA): answering the question about refund period ($q_3$, text in red); and 3) feeling connection with the user: actively responding to the user’s complaints and soothe his/her emotion ($q_4$-$q_6$, text in purple). Therefore, this corpus supports to build a more challenging and comprehensive dialogue system. What’s more, the average conversation turn in a dialogue is 20, so the long-term dependency among the context is an important feature. As the example shown in Table 1 to answer $q_4$, the assistant must look back to $r_2$ for further information. Besides, some contents in a real conversation are redundant or irrelevant to the final request. Taking $q_2$ as an example, the user explains the reason for his/her requirements, which actually has no contribution to solve the problem. Since all the data from our corpus is extracted from the real scenario, it reflects the characteristics in human-human conversations.

To bring the dataset more valuable for dialogue research, we labelled the intent for each query in all dialogues with a high-precision in-house classifier, which covers 289 different intents in real E-commerce after-sales scenario. We also prepared three Challenge Sets for evaluating dialogue systems better. In each set, different input information is provided and multiple ground-truth answers are annotated. We plan to annotate more information (e.g., emo-
Table 1: An example from JDDC corpus. Best viewed in color.

| q1  | 可以帮我改下订单的地址吗？ (Could you help me change the address of the order?) |
| q2  | 非常抱歉！我的物流还有待完善呢！ (I’m sorry. Our logistics system needs to be improved.) |
| q3  | 正常，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.) |
| q4  | 然而这也不是很方便哦！ (Why can’t I change my address? That is too inconvenient.) |
| q5  | 这也太麻烦了。我还着急用呢！ (That is too troublesome, I’m in a hurry.) |
| q6  | 行吧。 (Fine.) |
| q7  | 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?) |
Table 2: Existing related large-scale datasets applicable to dialogue systems. Note that '-' represents the number is not mentioned in related papers.

| Dataset                  | Dialogues | Utterances | Words | Description                                      |
|--------------------------|-----------|------------|-------|--------------------------------------------------|
| Twitter Corpus (Ritter et al., 2010) | 1,300,000 | 3,000,000 | -     | English post-reply pairs extracted from Twitter   |
| Weibo Corpus (Shang et al., 2015a)     | 4,435,959 | 8,871,918 | -     | Chinese post-comment pairs extracted from Weibo.com |
| Ubuntu Corpus (Lowe et al., 2015)     | 930,000   | 7,100,000 | 100,000,000 | English post-reply chat logs from Ubuntu Forum. |
| Douban Corpus (Wu et al., 2017)       | 1,060,000 | 7,092,000 | 131,747,880 | Chinese post-reply chat logs from Douban          |
| PERSONA-CHAT (Zhang et al., 2018a)    | 10,907    | 162,064    | -     | English personalizing chit-chat dialogue corpus  |
| DuConv Corpus (Wu et al., 2019)       | 29,858    | 270,399    | 2,872,340 | Chinese knowledge-driven conversation dataset   |
| SGD Corpus (Rastogi et al., 2019)     | 16,142    | 659,928    | 3,217,149 | English multi-domain task-oriented dialogue corpus |
| ECD Corpus (Zhang et al., 2018b)      | 1,020,000 | 7,500,000 | 49,000,000 | Chinese E-commerce dialogue corpus from Taobao   |
| JDDC Corpus                  | 1,024,196 | 20,451,337 | 150,716,172 | Chinese E-commerce dialogue corpus from JD       |

Table 3: Basic statistics of JDDC dataset.

|                     |       |
|---------------------|-------|
| Total sessions      | 1,024,196 |
| Total utterance     | 20,451,337 |
| Total words         | 150,716,172 |
| Average words per utterance | 7.4 |
| Average turns per session | 20 |
| Max turns           | 83    |
| Min turns           | 2     |

Table 3: Basic statistics of JDDC dataset.

3.2. Intent Distribution

Different from open-domain chit-chat or task-oriented dialogue (e.g. booking restaurants or flight tickets), conversations in E-commerce after-sales scenario usually have explicit goals, which can be returning product, changing delivery address, or simply inquiring the warranty policy etc. So knowing the goals is important for modeling this dialogue task. To facilitate the research in the future, we labelled the intent for each query in all dialogues with a high-quality in-house intent classifier. The classifier contains totally 289 intents, and it’s trained with Hierarchical Attention Network (Yang et al., 2016) model so context is also considered. The training data for the classifier includes totally 578,127 instances and all the training instances are annotated by professional customer service staffs. The classification accuracy reaches 93%, so the predicted intents for the JDDC dataset are reliable.

Figure 2 gives the distribution of top 20 intents. The top five intents are: ‘Warranty and return policy’, ‘Delivery duration’, ‘Change order information’, ‘Check order status’ and ‘Contact customer service’. Among them, ‘Warranty and return policy’ accounts for 9.2%, which is the most common intent in after-sales circumstance. This distribution
is also consistent with our real experience in E-commerce scenario that we often concern about warranty and delivery cycle, and ask for changing or returning the product.

Challenge Set II, we mask the answers of a multi-turn conversation as \( \{q_1, q_2, ..., q_{i+1}\} \), and the dialogue system is required to generate answers \( \{r_1, r_2, ..., r_{i+1}\} \) according to sequential questions. Particularly, this requires considering not only the input question but also the generated responses in previous turns. This task is more challenging than the previous one because incorrect replies may mislead the next output. Totally 15 dialogues are annotated with 168 questions to be answered in this set.

3.3. Challenge Set

To promote the research of human-machine dialogue systems with massive data in real scenario, we also held large-scale multi-turn dialogue competition with JDDC dataset, namely, JingDong Dialogue Challenge in 2018 and 2019. Aiming to fully evaluate the dialogue systems submitted in the competitions, we released 3 challenge sets with different input information, and also annotated multiple ground-truth answers for each task. To further clarify these challenge sets, diagrams shown in Figure 3 illustrates the difference between 3 tasks.

**Challenge Set I.** The dialogue system is required to output the final response \( r_{i+1} \) by utilizing provided multi-turn dialogue context in the format of \( \{q_1, r_1, q_2, r_2, ..., q_{i+1}\} \), where \( q \) represents question and \( r \) means response. This task is designed for long context modeling. Totally 300 dialogues are annotated, so we have 300 questions to be answered in this set.

http://jddc.jd.com/

Figure 2: Distribution of intents in JDDC dataset.

![Figure 3: The explanation of our 3 challenge sets. The responses (r) in red color are required to be answered by the dialogue system.](http://jddc.jd.com/)

| Challenge Set | Description |
|---------------|-------------|
| Challenge Set II | We mask the answers of a multi-turn conversation as \( \{q_1, q_2, ..., q_{i+1}\} \), and the dialogue system is required to generate answers \( \{r_1, r_2, ..., r_{i+1}\} \) according to sequential questions. Particularly, this requires considering not only the input question but also the generated responses in previous turns. This task is more challenging than the previous one because incorrect replies may mislead the next output. Totally 15 dialogues are annotated with 168 questions to be answered in this set. |
| Challenge Set III | We combine the characteristics of the former two tasks, for which the model needs to generate the response \( \{r_i, r_{i+1}\} \) sequentially under circumstance of several rounds of dialogue context and sequential questions \( \{q_1, r_1, q_2, r_2, ..., q_i, r_i\} \). The questions here are mainly long-tailed and hard questions compared with Challenge Set I and II. Totally 108 dialogues are annotated and 500 questions need to be answered in this set. |

4. Experiments

In this section, we conduct experiments on the JDDC dataset. We focus on two categories of models used in data-driven dialogue systems: retrieval-based models based on BM25 and BERT (Devlin et al., 2019) and generative models (Song et al., 2018; Gu et al., 2016). We will introduce some empirical settings, including dataset preparation, baseline methods, parameter settings. Then we introduce the experimental results on this dataset.

4.1. Experimental Setup

We first divide the around 1 million conversation sessions into training, validation and testing set. Then we split them into the \( \{I, R\} = \{q_1, r_1, q_3, r_2, Q, R\} \) format, where \( I = \{C, Q\} \) stands for input, \( C = \{q_1, r_1, q_2, r_2\} \) is the dialogue context and \( Q \) represents the last query. Please notice that we only kept most recent 2 rounds of dialogue as context. We also filtered some too short/long dialogues during experiment preparation. The statistics of the pre-processed dataset for experiment are shown in Table 4.

| Train | Valid | Test |
|-------|-------|------|
| Sessions | 963,358 | 4,992 | 4,992 |
| I-R Pairs | 1,522,859 | 5,000 | 5,000 |

Table 4: JDDC dataset division in experiments.

To construct the dataset for retrieval-based models, we randomly pick 500,000 I-R pairs from the dataset as positive samples. For every pair, we also select one response randomly to construct the negative I-R pair. So totally 1 million I-R pairs are used to fine-tune the BERT model. Our model implementation for BERT is based on Google’s work...
and follows the hyper-parameter settings in the original model.

For generative models, we first clean the dataset to decrease the portion of short responses (shorter than 3 Chinese characters) and control the portion of general responses (e.g., “What else can I do for you?”). Then we randomly pick 1 million I-R pairs for training. Our code implementation is based on the machine translation toolkit OpenNMT (Klein et al., 2017). In all generative experiments, we set 100,000 for vocabulary size and 200 for word embedding dimension. The source length is 128 and target length is decreased to 40 to avoid generating too long response. Other training settings are the same as default.

4.2. Comparable Models
In this subsection, we will introduce the detailed information on retrieval-based models and generative models used for our experiment.

4.2.1. Retrieval-based Models

BM25 To make the retrieval baseline more efficient, we firstly index all the Input-Response pairs in the training set using ElasticSearch. Then we use BM25 to retrieve the top 20 candidates for further matching. Response from top 1 candidate is used for evaluation. The equation of BM25 is defined as:

\[
S(I_{test}, I_{doc}) = \sum_{i}^{n} W_i \cdot S_1(w_i, I_{doc}) \cdot S_2(w_i, I_{test})
\]

where \( I_{test} \) stands for the test input including context C and query Q, \( w_i \) is the i-th word in the \( I_{test} \), \( I_{doc} \) is the document input in the repository, \( W_i \) represents the weight of \( w_i \) (such as inverse document frequency), and \( S(\cdot) \) calculates the relevance score of the two elements. Therefore, \( S(I_{test}, I_{doc}) \) is the similarity score between the test input and the existing I-R pairs in the repository.

BERT-Retrieval The retrieval method above only use one lexical feature to calculate the similarity. To capture more semantic information, we fine-tuned the pre-trained BERT model (Devlin et al., 2019) and add a dense layer with softmax as classifier to get the semantic similarity score for every \( (I_{test}, R_{doc}) \) pair. Then we use BERT score to re-rank the top 20 candidates from ElasticSearch and return the final top 1.

4.2.2. Generative Models

Vanilla Seq2Seq We implement the vanilla Sequence-to-Sequence (Seq2Seq) model (Shang et al., 2015b) with 512-unit 4-layer Bi-LSTMs for both the encoder and decoder. The input is the concatenated context and query, while the output is the response.

Attention-based Seq2Seq To improve our baseline, we applied attention mechanism (Luong et al., 2015) in the Seq2Seq model. This model is regarded as our second baseline and will be referred as Seq2Seq-Attention.

Attention-based Seq2Seq with Copy The context-query input is usually long and contains a lot of rare terminologies like “京东白条” (Jing Dong IOU(I owe you)), which may be OOV (out of vocabulary) words. Therefore, we add the copy mechanism (Gu et al., 2016) to the attention-based Seq2Seq baseline (Seq2Seq-Copy). The copy mechanism can explicitly extract words or phrases like certain entities from the input.

4.3. Evaluation Measures
In order to provide comparable baseline results for future research, we use some quantitative metrics for automatic evaluation. BLEU and ROUGE scores, which are widely used in NLP and multi-turn dialogue generation tasks (Tian et al., 2017; Luo et al., 2015; Shen et al., 2019), are used to measure the quality of generated responses via the comparison with the ground truths. The recently proposed Distinct (Distinct-1/2) (Li et al., 2016), are used to evaluate the degree of diversity by calculating the ratio of unique unigrams and bigrams in the generated responses.

4.4. Experimental Results
In this section, we analyze different baselines’ performance based on automatic evaluation measures and present in-depth case study.

4.4.1. Automatic Evaluation Results

|                  | BLEU | Rouge-L | Dist-1 | Dist-2 |
|------------------|------|---------|--------|--------|
| BM25             | 9.94 | 19.47   | 5.03%  | 28.89% |
| BERT-Retrieval   | 10.27| 19.90   | 5.23%  | 30.85% |
| Vanilla Seq2Seq  | 9.02 | 17.11   | 1.49%  | 4.25%  |
| Seq2Seq-Attention| 14.15| 22.17   | 1.79%  | 6.31%  |
| Seq2Seq-Copy     | 14.27| 23.62   | 1.79%  | 6.14%  |

Table 5: Automatic evaluation results. Dist-1/2 for Distinct-1/2.

The results of automatic evaluation are shown in Table 5. To further study the responses given by these models, we conduct some statistical analysis on the result sets. In Table 6 we present the response diversity (the portion of unique responses) in the Ground Truth set, BERT-Retrieval result set, and Seq2Seq-Copy result set. The portions of top three most common responses in the result sets are also listed in the table. Our observations can be summarized as follows:

1. For retrieval-based models, BERT-Retrieval performs better than BM25, which shows the strong ability of pre-trained model in semantic matching task. For generative models, Seq2Seq-Copy performs the best, which shows the effectiveness of using attention and copy mechanism.

2. The generative model has better performance in the similarity metrics (BLEU and Rouge-L) with the ground truth. There are mainly two reasons. One reason is generative model can generate new answers
3. The retrieval-based model performs much better in response diversity (Dist-1/2 in Table 5 and the response diversity in Table 6). While the performance of the generative model is very poor since it prefers to generate similar generic responses repeatedly (shown in Table 6), which is the common disadvantage of the generative models (Li et al., 2016).

4.4.2. Case Study
We also find two representative cases shown in Table 7 to illustrate the difference between the two approaches intuitively. The retrieval model fails in the first case because it gives wrong information (talking about the “SSD”, not the account), and the generative model fails in the second case for giving a generic response rather than useful information. From the cases above, we can see that the retrieval model tends to give responses with specific information (like “SSD” and “tax number”), while the generative model usually gives generic answers (such as “yes” or “what else can I do for you?”).

For some quite specific questions which may never appear in the repository, the retrieval model may give a wrong answer, even though they are both talking about a similar topic. However, for frequently asked questions like invoicing, order-cancellation, etc., it can perform well. For generative models, the generated generic responses may be lack of information, but they rarely make mistakes. Even not that satisfied, the users can still accept the generic responses sometimes. However, in general, both retrieval and generative models above are still not good enough for this task. This shows the task complexity in JDDC corpus.

### Table 6: Response Diversity statistics.

| Response Diversity                     | Ground Truth | BERT-Retrieval | Seq2Seq-Copy |
|----------------------------------------|--------------|----------------|--------------|
| “Yes”                                  | 88.74%       | 93.63%         | 28.08%       |
| “What else can I do for you?”          | 4.10%        | 0.94%          | 3.74%        |
| “Wait a moment, I’ll check for you right away” | 3.42%        | 0.88%          | 24.48%       |
|                                         | 0.50%        | 0.18%          | 3.90%        |

### Table 7: Examples of case study.

**Example 1**

| q1  | 帮我查下这个商品 (Please check this item for me) |
|-----|-----------------------------------------------|
| r1  | 好的，请问有什么可以帮您 (Ok, what can I do for you?) |
| q2  | 你好 (Hi)                                      |
| r2  | 你好 (Hi)                                      |
| Q   | 我要换这个摄像机，我这个上面绑定的账号能不能换绑？ (I want to change this camera, may I change the bound account on this either?) |

BERT-Retrieval: 这个摄像机没有储存卡就不能回放。 (You can’t play back the video on the camera if you don’t have the SSD.)

Seq2Seq-Copy: 可以的哦。 (Yes, it’s ok.)

**Example 2**

| q1  | 电子发票报销方便吗? (Is it convenient to reimburse with electronic invoice?) |
|-----|-----------------------------------------------------------------------------|
| r1  | 方便的，电子发票也是一样的 (Yes, electronic invoice is the same.)         |
| q2  | 稍等我问问会计 (Wait a moment for me to ask the accountant.)                |
| r2  | 好的 (No problem.)                                                          |
| Q   | 那就开电子票吧 (Ok, send me electronic invoice please.)                    |

BERT-Retrieval: 好的，请问您税号多少呢? (Ok, what’s your tax number please?)

Seq2Seq-Copy: 好的，请问还有其他可以帮到您的吗? (Ok, what else can I do for you?)

5. Conclusions and Future Work
In this work, we construct the Chinese JDDC dataset which is large-scale, multi-turn and collected in real scenario. We also contribute three high-quality human-annotated challenge sets for better evaluation. Over 200 intents are also labelled for each query in the dataset. We also evaluate several mainstream models on this dataset. The experimental results indicate either retrieval or generative models still have a long way to go in order to solve the real scenario conversation problem. More in-depth researches on context modeling, controllable response generation, question and answering, and reinforcement learning are needed in the future. Moreover, we will enrich the dataset annotations (e.g., emotions, and external knowledge) from various aspects in future work. Our dataset is available at: [https://jddc.jd.com/](https://jddc.jd.com/) and we hope it can serve as an effective testbed and benefit future research in dialogue systems.
6. Bibliographical References

Al-Rfou, R., Pickett, M., Snaider, J., Sung, Y.-h., Strope, B., and Kurzweil, R. (2016). Conversational contextual cues: The case of personalization and history for response ranking. arXiv preprint arXiv:1606.00372.

Allen, J. F., Miller, B. W., Ringger, E. K., and Sikorski, T. (1996). A robust system for natural spoken dialogue. In Proceedings of the 34th annual meeting on Association for Computational Linguistics, pages 62–70. Association for Computational Linguistics.

Bordes, A., Boureau, Y.-L., and Weston, J. (2016). Learning copying mechanism in sequence-to-sequence learning. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 5016–5026.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogues. In Proceedings of the 2nd workshop on cognitive modeling and computational linguistics, pages 76–87. Association for Computational Linguistics.

Devinin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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), pages 4171–4186.

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., Szlam, A., and Weston, J. (2015). Evaluating prerequisite qualities for learning end-to-end dialog systems. arXiv preprint arXiv:1511.06931.

El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrrota, R., and Suleman, K. (2017). Frames: a corpus for adding memory to goal-oriented dialogue systems. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pages 207–219.

Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. arXiv preprint arXiv:1603.06393.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. ACM Trans. Inf. Syst., 2(1):26–41, January.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In Proceedings of NAACL-HLT, pages 110–119.

Li, J., Song, Y., Zhang, H., and Shi, S. (2018). A manually annotated chinese corpus for non-task-oriented dialogue systems. arXiv preprint arXiv:1805.05542.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285–294.

Luo, L., Xu, J., Lin, J., Zeng, Q., and Sun, X. (2018). An auto-encoder matching model for learning utterance-level semantic dependency in dialogue generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 702–707.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Petukhova, V., Gropp, M., Klakow, D., Schmidt, A., Eigner, G., Topf, M., Srb, S., Motlicek, P., Potard, B., Dines, J., et al. (2014). The dbbox corpus collection of spoken human-human and human-machine dialogues. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), number CONF. European Language Resources Association (ELRA).

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2019). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. arXiv preprint arXiv:1909.05855.

Ritter, A., Cherry, C., and Dolan, W. B. (2010). 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, pages 172–180.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In Proceedings of the conference on empirical methods in natural language processing, pages 583–593. Association for Computational Linguistics.

Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., and Heck, L. (2018). Building a conversational agent overnight with dialogue self-play. arXiv preprint arXiv:1801.04871.

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

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

Shen, L., Feng, Y., and Zhan, H. (2019). Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. arXiv preprint arXiv:1906.07429.

Song, Y., Shi, S., Li, J., and Zhang, H. (2018). Directional
skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., and Zhao, D. (2017). How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236.

Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer.

Wang, H., Lu, Z., Li, H., and Chen, E. (2013). A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945.

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

Wu, W., Guo, Z., Zhou, X., Wu, H., Zhang, X., Lian, R., and Wang, H. (2019). Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). 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*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018a). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhang, Z., Li, J., Zhu, P., Zhao, H., and Liu, G. (2018b). Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752.