A Manually Annotated Chinese Corpus for Non-task-oriented Dialogue Systems

Jing Li, Yan Song, Haisong Zhang, Shuming Shi
Tencent AI Lab
{ameliajli,clksong,hansonzhang,shumingshi}@tencent.com

Abstract

This paper presents a large-scale corpus for non-task-oriented dialogue response selection, which contains over 27K distinct prompts more than 82K responses collected from social media.[1] To annotate this corpus, we define a 5-grade rating scheme: bad, mediocre, acceptable, good, and excellent, according to the relevance, coherence, informativeness, interestingness, and the potential to move a conversation forward. To test the validity and usefulness of the produced corpus, we compare various unsupervised and supervised models for response selection. Experimental results confirm that the proposed corpus is helpful in training response selection models.

1 Introduction

Building a dialogue system that can naturally interact with human beings has long been a mission of artificial intelligence ever since the formulation of Turing test (Turing, 1950)[2]. Recently, the breakthrough of artificial intelligence and the availability of big data have jointly brought a surge of interest towards building data-driven dialogue systems. These systems have drawn attentions from not only academia, but also industries, e.g., Apple's Siri[3], Google's smart reply (Kannan et al., 2016), and Microsoft's Xiaoice (Markoff and Mozur, 2015). In terms of functionality, existing dialogue systems can be categorized into two types: task-oriented agents (Young et al., 2013) and non-task-oriented chatbots (Perez-Marin, 2011). Task-oriented agents aim to help people complete a specific task, while non-task-oriented chatbots chitchat on a wide rage of topics, which is particularly popular for serving as friendly conversation partners. For example, in China, Xiaoice has over 20 million registered users and 850 thousand followers on Weibo.[4]

Prompt: *Apple has always been my favorite food!!!*		
Response	**Rating**	**Criterion**
The best RPG ever: URL.	bad	off-topic
Me too!	acceptable	versatile
Me too! One apple a day keeps the doctor away.	excellent	informative

Table 1: Sample responses of the prompt "*Apple has always been my favorite food!!!*" with their ratings and the corresponding criteria.

Training a data-driven dialogue system requires massive data. Conventionally, this prerequisite was fulfilled by collecting conversation alike messages from social media (Wang et al., 2013; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2016; Xing et al., 2017; Shao et al., 2017). However, in doing so, the quality of system output is affected by the noisy, informal, fragmented, ungrammatical nature of social media messages. To illustrate this phenomenon, we list some sample responses for the prompt "*Apple has always been my favorite food!!!*" on Twitter in Table 1. The first response in Table 1, targeting for advertising some RPG game instead of reacting to the prompt, is a bad response for the reason of being completely off-topic and irrelevant. The second one is a versatile response and can safely reply to diverse of prompts. Owing to the prevalence of versatile responses on social media, chatbots trained by so-

[1] This dataset is available at: http://ai.tencent.com/ailab/upload/PapersUploads/A_Manually_Annotated_Chinese_Corpus_for_Non-task-oriented_Dialogue_System.

[2] https://en.wikipedia.org/wiki/Turing_test

[3] https://www.apple.com/ios/siri/

[4] https://en.wikipedia.org/wiki/Xiaoice

cial media responses without distinguishing their quality tend to yield such "one size fits all" responses (Li et al., 2016; Xing et al., 2017). Therefore, versatile responses should be effectively distinguished from good instances. The third response is on-topic, coherent, and informative, can thus serve as an excellent positive instance in training dialogue systems. Given the aforementioned patchy responses, effective models to select or rank responses for dialogue systems are particularly important in distinguishing responses with diverse quality. Therefore, well labeled prompt-response data become prerequisite to train such models, and can be further used to benefit response generation (Shao et al., 2017) and evaluation (Lowe et al., 2017) in dialogue systems.

In this paper, we build a large-scale corpus containing over 27K Chinese prompts with 82K prompt-response pairs[5] with 5-grade human ratings, i.e., bad, mediocre, acceptable, good, and excellent, regarding the criterion such as relativeness, coherence, informativeness, interestingness, etc. Most previous efforts focus on using unannotated (Xing et al., 2017; Shao et al., 2017) and automatically annotated (Wang et al., 2013; Tan et al., 2015; Severyn and Moschitti, 2015) data. To the best of our knowledge, this work is the first attempt to build a Chinese corpus with manual annotations for non-task-oriented dialogue systems. On the annotated corpus, we conduct benchmark experiments comparing various models for dialogue response selection. Experimental results confirm that our corpus is helpful in selecting high-quality responses.

2 Data and Annotation

2.1 Data Collection

The prompt-response pairs in our corpus are collected from Tieba[6], Zhidao[7], Douban[8], and Weibo[9]. These websites are popular social media platforms in Chinese community, where the conversations on them cover diverse topics. We first extract the topic list from the index pages of each platform, For example, the Weibo topic list includes celebrities, love, military, sports, games, etc. such as celebrities, love, military, sports, games, etc., where the topic lists provided by different platforms are similar. We then use JSoup toolkit[10] to crawl and parse the pages for each topic, and collect the trending prompts and their responses from each topic. As a result, the collected raw data has over 11K prompt-response pairs with over 2M Chinese characters in total.

Rating 1 (bad): The response makes no sense (e.g., [S_1]) or is totally irrelevant with the prompt (e.g., [S_2]).
Rating 2 (mediocre): The response cannot coherently reply to the prompt but mention some keywords in it (e.g., [S_3]), including the cases that only echo with keywords in the prompt (e.g., [S_4]).
Rating 3 (acceptable): The response should be meaningful, relevant, and coherent, but has spacial or temporal limitations (e.g., [S_5]) or is a versatile response (e.g., [S_6]).
Rating 4 (good): The response is coherent and cover relevant content, but is simple and uninformative, which cannot actively move the conversation forward, such as [S_7].
Rating 5 (excellent): The response is not only coherent and relevant, but also informative, interesting, or initiate new and relevant topic that actively leads the conversation to continue, e.g., [S_8].

(a) Schemes for rating from 1 to 5.

Prompt: *"Hey, Beijing, I'm coming"*
[S_1]: <**R1** (Nonsense)> *ddddd*
[S_2]: <**R1** (Irrelevant)> *Maybe not?*
[S_3]: <**R2** (Incoherent)> *Beijing's larger than HK.*
[S_4]: <**R2** (Echoing)> *Beijing*
[S_5]: <**R3** (ST-limited)> *Keep warm! It's snowing.*
[S_6]: <**R3** (Versatile)> *Great!*
[S_7]: <**R4** (Good)> *Enjoy! Beijing is beautiful.*
[S_8]: <**R5** (Excellent)> *Enjoy! Beijing is beautiful. What hotel are you staying?*

(b) Sample responses for the prompt "*Hey, Beijing, I'm coming*". [S_i] is a sample response. **Ri** refers to the rating of the sample response, and Type is our interpreted response type regarding their quality, e.g., ST-limited means spatial or temporal limited responses.

Table 2: Annotation guidelines of our corpus.

2.2 Data Cleaning

Before manual annotation, there are two preprocessing steps for data cleaning. The first step is to filter out sensitive prompt-response pairs that contain dirty words, adult content, and intimate individual details. This step is to avoid any chatbot trained or evaluated based on our corpus to produce uncomfortable content and letting out private information of individuals. In the second step, we identify the knowledge-dependent prompts whose

[5] A prompt-response pair refers to a pair with a prompt and one of its response.
[6] https://en.wikipedia.org/wiki/Baidu_Tieba
[7] https://en.wikipedia.org/wiki/Baidu_Knows
[8] https://en.wikipedia.org/wiki/Douban
[9] https://en.wikipedia.org/wiki/Sina_Weibo
[10] https://jsoup.org/

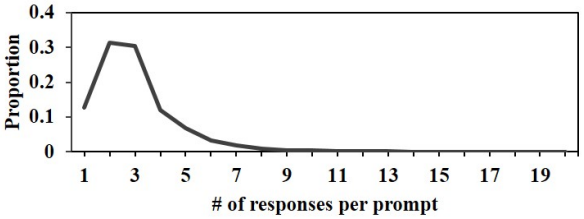

Figure 1: The distribution of response number per prompt.

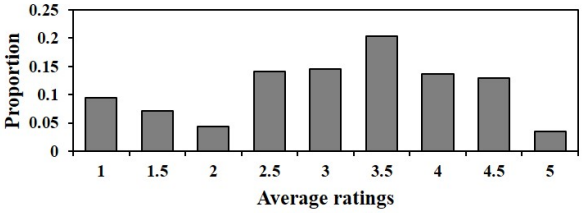

Figure 2: The distribution of the averaged ratings given by two annotators.

responses can be given only with specific knowledge, e.g., "*What's the weather today in Beijing?*", and remove such prompts and all their corresponding responses. The reason behind this step is that a non-task-oriented system does not exploit any domain-specific knowledge. To conduct the two steps, four experienced annotators are hired for manual filtering. As a result, there are $105,825$ pairs remaining in the corpus.

2.3 Rating Annotation

To annotate each response, we again hired four annotators to label all responses into 5 ordinal ratings. The rating scores from 1 to 5 refers to bad, mediocre, acceptable, good, and excellent responses according to the annotation guidelines listed in Table 2. Each response with its prompt is assigned to two different annotators and annotated independently. The detailed rating scheme is listed in Table 2a. For better understanding the rating scheme, Table 2b shows 8 types of sample responses in illustrating the scheme.

Specifically, bad responses are either meaningless (e.g., $[S_1]$) or completely off-topic (e.g., $[S_2]$). Mediocre responses may be topic-related, but fail in coherently reacting to the prompt (e.g., $[S_3]$), or simply echo with keywords in the prompt (e.g., $[S_4]$). Responses with these two ratings are below expectation and considered as failed cases.

Acceptable ratings are given to boundary cases that are meaningful, coherent, and relevant responses to a prompt. Responses with this rating refers to two typical types, i.e., responses with spacial or temporal limitation, and versatile responses. Responses with spacial or temporal limitations, namely ST-limited responses, are valid only under specific spacial or temporal circumstance. For instance, $[S_5]$ works well in the winter, but looks weird when the conversation takes place in the summer. Since a chatbot should react decently in any situation, such instances should be properly distinguished from good cases in training set. For some fine responses, although they have no spacial or temporal limitation, they are too general to provide specific information for different prompts. Because these versatile responses can be used for multiple prompts, we consider them acceptable instead of good, so as to avoid chatbots from yielding "one size fits all" replies, which is a major drawback of the existing chatbots (Li et al., 2016; Xing et al., 2017) due to the prominence of general responses on social media. For example, $[S_6]$ is an example of versatile responses, which can be used to reply to diverse prompts, such as "*I'm so happy today!*" and "*Coffee Corner is an awesome restaurant.*".

Good responses are natural and sound, with neither ST-limited nor versatile characteristics, such as $[S_7]$. Excellent responses require to be informative or interesting, which helps moving forward the conversation. For example, $[S_8]$ is excellent because it carries on the conversation with a proactive response that initializes a new topic.

	Prompts		Responses		Vocab
	Count	Avg len	Count	Avg len	
Train	21,964	4.05	65,706	7.01	
Dev	2,669	4.02	8,080	6.98	36,035
Test	2,750	4.11	8,224	7.03	

Table 3: Statistics of our corpus. Avg len: the average count of Chinese words after segmentation.

2.4 Corpus Statistics

In the annotated corpus, there are $82,010$ responses whose gaps of the two annotated ratings are ≤ 1. The average scores of the two annotations hence serve as the final ratings. To keep annotation consistency, we remove the rest $23,815$ responses with rating gaps ≥ 2. The final corpus contains $27,383$ distinct prompts and $82,010$ responses. The number of responses for each prompt ranges from 1 to 20. Figure 1 illustrates the distribution of response numbers per prompt.

In the final corpus, the inter-annotator agreement indicated by Cohen's κ (Gwet, 2014) is 80.0%, which implies high degree of consen-

Category	Models	Cut@3			Cut@4			Cut@5		
		P@1	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR
Unsupervised	Cosine sim	84.8	91.1	91.8	67.5	81.3	82.0	40.0	64.9	65.1
	BM25	86.1	**91.4**	**92.4**	70.6	82.7	83.7	53.8	72.3	73.4
Supervised	SVMRank	**86.2**	91.1	92.2	73.0	**84.0**	85.0	64.3	78.6	79.8
	GBDT	85.9	91.0	92.0	71.4	82.9	83.8	55.7	74.4	74.5
	BiLSTM	85.4	90.7	91.8	**73.8**	**84.0**	**85.3**	68.1	81.0	82.2
	CNN	85.5	90.6	91.8	72.5	83.4	84.6	**70.5**	**81.2**	**83.0**

Table 4: Comparison results (%). Higher scores indicate better results. **Cut@N**: responses with rating $\geq N$ are considered as positive, and as negative otherwise. Larger cut indicates a stricter standard. Best results in each column is marked as **bold**.

sus. The distribution of the ratings for all responses is demonstrated in Figure 2. It is observed that 48.9% responses obtain a rating falling in $[2.5, 3.5]$, which indicates the prevalence of rating 3 responses, i.e., versatile responses, or response with spacial or temporal limitations. This observation implies the importance of separating out these two types of responses from other instances. We also found that 35.0% responses obtain a rating ≤ 2.5. This demonstrates the noisiness nature of social media data, and thus it is problematic to assume all user-generated responses to be positive in chatbot training and evaluation.

3 Benchmark Experiments

To evaluate the quality of our corpus, we compare various ranking-based response selection models on it with different settings.

3.1 Experiment Setup

For fundamental processing, we use Jieba toolkit[11] for Chinese word segmentation. Later, we split the prompts into 80%, 10%, and 10% as training, development, and test set, respectively. A vocabulary is then built based on the training data. The statistics of the datasets is shown in Table 3.

We test two unsupervised baselines: Cosine Sim and BM25 (Manning et al., 2008). Cosine sim is to rank responses by their cosine similarity of TF-IDF representations to prompts. BM25 model ranks the responses with scores similar to TF-IDF measurement. The document frequency (DF) of all words are calculated based on training set. For supervised models, we test learning to rank models, namely, SVMRank (Joachims, 2002)[12] and gradient boosting decision tree (GBDT) (Friedman, 2001)[13], with manually-crafted features proposed by Wang et al. (2013), e.g., the length of responses and the cosine similarity between a response and its prompt, etc. In addition, we test two state-of-the-art neural models, bidirectional long short-term memory (BiLSTM) (Tan et al., 2015) and convolutional neural networks (CNN) (Severyn and Moschitti, 2015), for answer selection in question-answering research, where in our experiments prompts and responses is mapping to questions and answers, respectively. For all the aforementioned models, hyper-parameters are tuned on development set. For neural models, the hidden size of BiLSTM and CNN encoders are both set as 300. Mean squared error (MSE) is used as the loss function and early-stop strategy (Caruana et al., 2000; Graves et al., 2013) applied in training.

3.2 Comparison Results

We follow the paradigm of question answering to separate responses to be "positive" and "negative" when evaluating the ranked responses given a prompt. In doing so, we set rating thresholds at 3 ,4 and 5, where responses with gold-standard rating $\geq N$ are considered as positive instances, and otherwise as negative instances. Therefore, larger N indicates stricter standard. Given different cut, the result is reported in Table 4, with P@1 (precision@1), mean averaged precision (MAP), and mean reciprocal rank (MRR) scores of all models on the test set. In particular, we remove all responses of a prompt in evaluation if none of them is considered positive under a specific cut, because for this prompt, all models would score 0 no matter how its responses are ranked.

The overall observation is that supervised models perform better than unsupervised models, which indicates the usefulness of our corpus in

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

[12] https://www.cs.cornell.edu/people/tj/svm_light/

[13] https://sourceforge.net/p/lemur/wiki/RankLib/

helping select effective responses. It is also observed that, as the standard becoming stricter by given a larger cut, unsupervised models suffer a larger performance drop, while supervised models yield robust scores. This observation shows that the response selection models trained by our corpus can well distinguish responses of different quality, and can thus produce better ranks, e.g., excellent responses are assigned higher ranking scores than good ones.

4 Conclusion

In this paper, we present a large-scale Chinese corpus containing over 27K distinct prompts and 82K prompt-response pairs. In this corpus, each response is annotated with a 5-grade rating score regarding to its quality in relevance, coherence, informativeness and interestingness. This corpus, to the best of our knowledge, is the first manually annotated Chinese dataset for non-task-annotated dialogue systems. Therefore, it is more reliable than automatic collected data and thus potentially beneficial to chatbot training and evaluation. Benchmark experiments on this corpus comparing various response selection models confirm the usefulness of the proposed corpus for dialogue systems.

References

Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In *Advances in Neural Information Processing Systems*. Denver, Colorado, USA, pages 402–408.

Jerome H Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics* pages 1189–1232.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada, pages 6645–6649.

Kilem L Gwet. 2014. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Thorsten Joachims. 2002. Optimizing Search Engines using Clickthrough Data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada, pages 133–142.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA, pages 955–964.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *In proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, USA, pages 110–119.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, pages 1116–1126.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

John Markoff and Paul Mozur. 2015. For Sympathetic Ear, More Chinese Turn to Smartphone Program. *NY Times* .

Diana Perez-Marin. 2011. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems using Generative Hierarchical Neural Network Models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. Phoenix, Arizona, USA, pages 3776–3784.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Santiago, Chile, pages 373–382.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. 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 of the Asian Federation of Natural Language Processing*. Beijing, China, pages 1577–1586.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating High-quality and Informative Conversation Responses with Sequence-to-sequence Models.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, pages 2210–2219.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. Melbourne, Victoria, Australia, pages 553–562.

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for Non-factoid Answer Selection. *arXiv preprint* arXiv: abs/1511.04108.

Alan M Turing. 1950. Computing Machinery and Intelligence. *Mind* 59(236):433–460.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869* abs/1506.05869.

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

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, California, USA, pages 3351–3357.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based Statistical Spoken Dialog Systems: A Review. In *Proceedings of the IEEE*. volume 101, pages 1160–1179.