Learning to Converse with Noisy Data: Generation with Calibration

Mingyue Shang[1], Zhenxin Fu[1], Nanyun Peng[2], Yansong Feng[1], Dongyan Zhao[1,3], Rui Yan[1,3]*
[1]Institute of Computer Science and Technology, Peking University, China
[2]Information Science Institute, University of Southern California, USA
[3]Beijing Institute of Big Data Research, China
{shangmy, fuzhenxin, fengyansong, zhaodongyan, ruiyan}@pku.edu.cn, npeng@isi.edu

Abstract

The availability of abundant conversational data on the Internet brought prosperity to the generation-based open domain conversation systems. In the training of the generation models, existing methods generally treat all the training data equivalently. However, the data crawled from the websites may contain many noises. Blindly training with the noisy data could harm the performance of the final generation model. In this paper, we propose a generation with calibration framework, that allows high quality data to have more influences on the generation model and reduces the effect of noisy data. Specifically, for each instance in training set, we employ a calibration network to produce a quality score for it, then the score is used for the weighted update of the generation model parameters. Experiments show that the calibrated model outperforms baseline methods on both automatic evaluation metrics and human annotations.

1 Introduction

With the boom of on-line social media and community question-answering platforms, the amount of human-human conversation data available on public websites is growing rapidly. These large volumes of resources stimulate great interests of researchers in building human-computer conversation systems with data-driven approaches. The past few years have seen the prosperity of data-driven human-computer conversation systems in open domain (a.k.a non-task oriented dialogue systems). Previous methods can be roughly categorized into two groups: retrieval-based systems [Hu *et al.*, 2014; Lu and Li, 2013; Yan *et al.*, 2016] and generation-based systems [Ritter *et al.*, 2011; Vinyals and Le, 2015; Shang *et al.*, 2015; Serban *et al.*, 2016].

Inspired by the success of deep learning models for machine translation [Cho *et al.*, 2014; Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014], many researchers adopted recurrent neural networks (RNNs) to build generation-based conversation systems. A widely used framework is sequence-to-sequence (*seq2seq*) [Sutskever *et al.*, 2014], which is composed of two RNNs: an encoder and a decoder. Various methods have been proposed to further improve the performance of seq2seq-based conversation systems. Sordoni *et al.* [2015] considered the conversational context in their generation model. Other work introduced extra information, such as topic [Xing *et al.*, 2017] and personality [Li *et al.*, 2016], into the traditional seq2seq model to promote more informative replies.

Despite considerable efforts made to improve seq2seq models for conversation, previous work generally treats all the conversation data equally, which means each query-reply pair contributes equivalently to the final models. However, the conversations crawled from the Internet contain many noises. As shown in Table 1, in the first case, the reply is completely irrelevant to the query, and even human can not understand the meaning. In the second example, the reply is universal and boring. This universality will be amplified by the seq2seq training to generate safe replies, such as "I don't know" and "I think so", which are suitable for various queries, but are non-informative and uninteresting. Such low quality data will poison the learning process and severely affect the performance of the trained model.

To address the problem of noisy training data, we propose a generation model with calibration mechanism, which automatically evaluates the quality of the training data and promotes the high-quality instances to calibrate the model. We expect the calibrated model to perform better than the naively trained model in terms of the generated responses' quality.

There are two major challenges of this generation with calibration framework: 1) how to measure the quality of data? 2) how to calibrate the generation network efficiently and effectively based on the data quality? The most accurate and straight-forward way to measure the data quality is to have people manually annotate each instance with a quality score. However, obtaining human annotations is both time consuming and expensive. It is critical to come up with automatic measurements of the data quality.

Recently, significant improvements have been made to the automatic evaluation of the conversation systems. Tao *et al.* [2017] and Lowe *et al.* [2017] proposed relatedness-based evaluation metrics that employed matching networks to measure the relatedness between queries and replies. Our generation with calibration network follows Tao *et al.* [2017] to evaluate the quality of the training data. Specifically, we

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn)

Query	Reply
这两天上班忙的都没时间看论坛 I am so busy these days that I have no time to read the on-line forum.	恩，同喜欢摇滚. Yeah, I like Rock and Roll, too.
恐怖电影不太好吧，吓到了就不敢睡觉了 You'd better not watch horror film. You may be too scared to sleep.	哈哈 Aha!

Table 1: Conversation data of low quality. Data crawled from the website contains a lot of irrelevant reply (first line) and universal reply (second line).

adopt their matching network as our calibration model to automatically score the query-reply pairs in the training data.

To utilize the quality measurements to calibrate the model training, we employ the calibration model to control the contribution of each instance to the training of the generation model. Specifically, our generation model takes both the instances and their scores produced by the calibration model during training. The instances' scores will control their contribution to the training to promote high-quality instances and demote low-quality ones.

We conduct experiments using data crawled from a Chinese forum named Douban[1]. We implement three generation models under the seq2seq framework as baselines, and two calibrated generation models. Experiments show that the calibration model improves the performance of generation model on both automatic evaluation metrics and human evaluation comparing with the baselines.

The contributions of this paper are summarized as follows:

- We propose a generation with calibration framework to differentiate the impacts of training instances on the generation model based on their qualities. To the best of our knowledge, we are the first to explore the calibration mechanism on generation-based dialog systems.
- We design a neural network architecture for the generation with calibration framework.
- Our experiments demonstrate that modeling the quality of the training data and utilize the information to calibrate the training process is helpful for enhancing the performances of data-driven dialog systems.

2 Generation with Calibration

Most existing data-driven conversation models are trained with massive data crawled from the Internet without explicit modeling of the data quality. However, the low-quality training data such as the examples in Table 1 can be harmful to the training of generation models. To address this problem, we propose a generation with calibration framework which automatically measures the quality of the training instances and incorporates it into the training of the generation model.

2.1 Task Formulation

We assume a training data set of size N, denoted as $D = \{(Q^1, R^1), \ldots, (Q^N, R^N)\}$, where Q denotes an utterance and R denotes the response [2]. Formally, given a query $Q = (q_1, \ldots, q_i, \ldots, q_T)$ and its reply $R = (r_1, \ldots, r_i \ldots, r_{T'})$, where q_i and r_i are the i-th word in the query Q and the reply R, respectively. A generation model is trained to maximize the probability $P(R^j|Q^j)$ of each training instance.

Since the training data is crawled from the Internet, it contains many noises. Our goal is to learn a generation model with awareness of the data quality during training. Therefore, instead of treating each training instance equally, we introduce a calibration model to learn a scoring function $S(R^j, Q^j)$ to rate the quality of the training instances. We then transform the scores to weights $w_j = f_w(S(R^j, Q^j))$, which are incorporated into minimizing weighted loss objective of the training corpora.

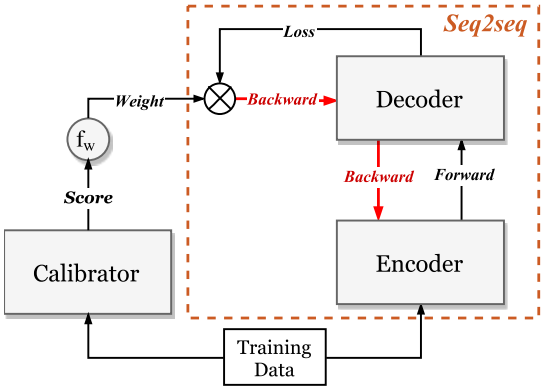

Figure 1: An overview of the generation with calibration framework. The generation network adopts the seq2seq model.

2.2 Model Overview

Figure 1 gives an overview of our proposed model. There are two major components: a calibration network and a generation network. The calibration network is trained to measure the quality of the query-reply pairs, and the generation network takes the scores produced by the calibration network to weight the training instances, such that the high-quality instances have more impact on the generation model while the low-quality ones are less influential. The following sections explain the two networks in details and introduce the model training.

2.3 Seq2seq with Attention for Generation

The generation network of our model is based on the sequence-to-sequence model with attention mechanism [Bahdanau *et al.*, 2014]. Given a query $Q = (q_1, \ldots, q_T)$ and a reply $R = (r_1, \ldots, r_{T'})$, the seq2seq model is trained to maximize the generation probability $P(R|Q)$. Particularly, the encoder transforms Q into an intermediate representation h through a recurrent neural network (RNN), and the decoder generates an output $\hat{R}$ with h as the input.

In this work, we employ a one-layer bidirectional gated recurrent unit (BiGRU) [Cho *et al.*, 2014] as the encoder and

[1] www.douban.com

[2] Q and R can also represent contextual utterances and responses for multi-turn dialogs. In this paper, we focus on single-turn conversation, and leave the extension to multi-turn systems as future work.

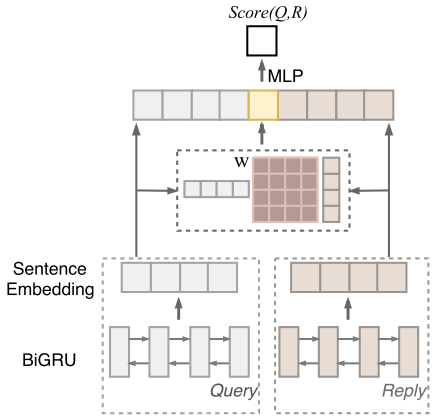

Figure 2: Structure of the calibration network.

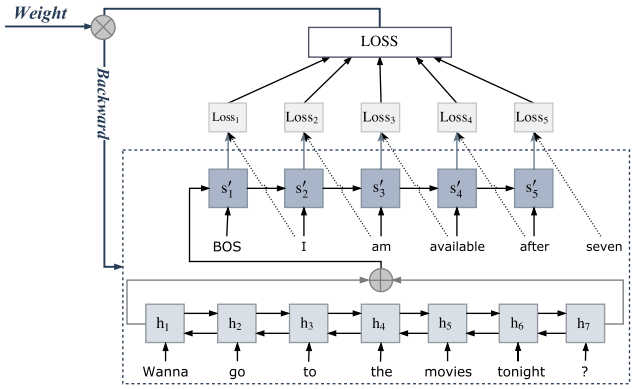

Figure 3: Training process of the calibrated seq2seq model. BOS denotes "begin of sentence".

a one-layer GRU as the decoder. The encoding BiGRU takes the hidden states of both directions and concatenates them to form the final hidden states $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$, where $\overrightarrow{h}_t$ is the forward state and the $\overleftarrow{h}_t$ is the backward state. Each word in a query Q is first mapped to an embedding, then fed into the encoder. The last hidden state h_T of the encoder is taken as the intermediate representation of Q.

During decoding, the attention mechanism is employed to model the alignments between the query and the generated reply. For the generation of each $\hat{r}_i$ in $\hat{R}$, a context vector c_i is computed as the weighted average of the encoder hidden states, where the weights come from the attention. The i-th hidden state of decoder is computed as:

$$s_i = f(r_{i-1}, s_{i-1}, c_i) \tag{1}$$

where r_{i-1} is the $(i\text{-}1)$-th word in the response and $f(.)$ is the GRU function. The probability distribution of the candidate words is computed through softmax. The probability of generating a response R is defined as:

$$P(R|Q) = \prod_{i=1}^{T'} p(r_i|Q, r_1, \ldots, r_{i-1}) \tag{2}$$

2.4 Calibration Network

The goal of the calibration network is to automatically estimate the quality of each query-reply pair in the training data. We follow Tao *et al.* [2017] and design an evaluation-based calibration network to gauge the quality of the conversational data by measuring the appropriateness of a reply given its query. Given a query $Q = (q_1, \ldots, q_T)$ and its reply $R = (r_1, \ldots, r_{T'})$, two one-layer $BiGRU$s are applied to transform each into sentence embeddings, respectively.

$$s_Q = BiGRU_q(Q) \tag{3}$$

$$s_R = BiGRU_r(R) \tag{4}$$

We take the last hidden state s_Q and s_R as representation of the query Q and the reply R. A scoring function is then trained to calculate appropriateness of a reply given its query as follows:

$$S = f_c(Q, R) = MLP\left([s_Q; s_Q^T W_e s_R; s_R]\right) \tag{5}$$

where $MLP(.)$ is a multi-layer perceptron with $\tanh$ activation function. To ensure the output scores are in the range of $[0, 1]$, we add an additional sigmoid function at the last layer of the MLP. W_e is a trainable parameter. $[;]$ denotes concatenation operation. The model structure is shown in Figure 2.

To train the calibration network, we apply negative sampling to release the burden of human annotation. Previous work has shown the effectiveness of negative sampling [Tao *et al.*, 2017; Mou *et al.*, 2016]. Particularly, for query Q and ground truth reply R, we randomly sample a R^- from other replies in training set. The training objective is that score of (Q, R) should be larger than (Q, R^-) with at least Δ threshold. Loss function is given by

$$L_c = \max\left(\Delta - f_c(Q, R) + f_c(Q, R^-),\ 0\right) \tag{6}$$

2.5 Generation with Calibration Model Training

The objective function of the vanilla seq2seq model minimizes the negative log probability of the training data D with size N as is shown in Equation 7.

$$L_{s2s} = \sum_{j=1}^{N} -\log P(R^j|Q^j) \tag{7}$$

which treats each training instance equally. Our proposed generation with calibration network, on the other hand, treats each instance differently based on its quality estimated by the calibration network. Specifically, our model minimizes the weighted negative log probability of the training data:

$$L_{gc} = \sum_{j=1}^{N} -f_w(S(R^j, Q^j)) \log P(R^j|Q^j) \tag{8}$$

where $S(R^j, Q^j)$ is the instance score produced by a pre-trained calibration network, and $f_w(.)$ is a function that transforms the scores into weights, which is defined as:

$$w_i = f_w(S_i) = \frac{S_i}{\frac{1}{b}\sum_{i=k}^{b} S_k} \tag{9}$$

where b is the batch size. Note that the original seq2seq is a special case of our model with all instances' weights equal

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Distinct-1	Distinct-2	Distinct-3	Distinct-4
S2S	9.651	1.049	0.142	0.043	0.022	0.076	0.128	0.183
S2SA	10.629	1.184	0.167	0.042	0.029	0.097	0.172	0.247
S2SA+MMR	7.727	0.788	0.069	0.001	0.067	0.252	0.472	0.669
S2SAC	**11.011**	**1.287**	**0.214**	**0.056**	0.044	0.140	0.242	0.339
S2SAC+MMR	8.782	0.764	0.094	0.029	**0.075**	**0.287**	**0.518**	**0.714**

Table 2: Automatic evaluation result on five models. The upper proportion shows baseline models. In the lower proportion, S2SAC and S2SAC+MMR are our proposed models.

to 1. The goal of the weight transformation in Equation 9 is to control the influence of the quality scores on the gradient updates. Specifically, after the transformation, half of the weights will be larger than 1 and another half smaller, and $\sum_{i=1}^{b} w_i = b$, which indicates the same learning strength as the original seq2seq model. With the instance weighting, the parameters of the generation model are updated by

$$W_{t+1} = W_t + \frac{lr}{b} \sum_{i=1}^{b} w_i \bigtriangledown l_{s2s}(Q^i, R^i) \qquad (10)$$

where lr is the learning rate of the generation model and W_t denotes all the parameters in it at t-th mini-batch. Figure 3 illustrates the training process of our calibrated seq2seq model.

2.6 Parameter Estimation

Our model is implemented in Pytorch[3]. All of our hyperparameters are set based on pilot experiments. We adopt Adam [Kingma and Ba, 2014] optimizer with initial learning rates as 0.0002 for the calibration network and 0.0001 for the generation network. We employ mini-batch training with batch size 64 for both the calibration and the generation network. We adopt the learning rate decay trick for training the generation network to halve the learning rate when the perplexity on validation begins to increase.

For the calibration network, we set the word embedding dimension to 200 and the hidden vector size to 256 for both query encoder and reply encoder. The Δ for calibration network is set to 0.05. For the generation network, the word embedding dimension is 480 and the hidden vector size for both the encoder and the decoder is 512. All the parameters are initialized randomly. We begin to calibrate our generation network from the second training epoch.

3 Experimental Setup

3.1 Dataset

We conduct our experiments on a large dataset crawled from Douban, which is a Chinese discussion forum. We performed Chinese word segmentation for each query-reply pair. There are 1,333,877 pairs in the training set, 10,000 for validation and 1,000 for test. We take the most frequent 80,000 words in the training data as the vocabulary and other words as UNKs.

[3]http://pytorch.org/

Model	0-ratio	1-ratio	2-ratio	Avg-score
S2S	0.168	0.718	0.112	0.942
S2SA	0.150	0.612	0.238	1.088
S2SA+MMR	0.320	0.386	0.294	0.974
S2SAC	0.120	0.558	0.322	1.202
S2SAC+MMR	0.208	0.358	0.424	1.206

Table 3: Human evaluation of the five models, regarding the relatedness and fluency of the generate replies. We report the ratio of score $\{0, 1, 2\}$ and the average score on each model, calculated using the annotations from 6 different annotators.

3.2 Baselines and Model Acronyms

We implemented three conversation models as baselines.

S2S represents the standard seq2seq model without attention mechanism.

S2SA denotes the standard seq2seq model with attention mechanism.

S2SA+MMR denotes a seq2seq model (with attention) that incorporates maximum marginal relevance (MMR) into beam search while decoding, which was proposed by [Song *et al.*, 2017] to enrich the diversity of the replies.

We also implemented two variations of the generation with calibration models for comparison.

S2SAC is a S2SA model with calibration mechanism.

S2SAC+MMR is the S2SA+MMR model with calibration mechanism.

3.3 Evaluation Metrics

The evaluation of generative dialog system is notoriously hard, therefore, we adopt several evaluation metrics to measure different aspects of our results.

BLEU: BLEU is a widely used automatic evaluation metric for neural machine translation and dialog systems [Li *et al.*, 2016]. It evaluates dialog systems by measuring word overlap between the generated reply and the ground-truth. We adopt BLEU-1$\sim$4 in this paper.

Distinct: We employ Distinct-1$\sim$4 following Song *et al.*; Li *et al.* [2017; 2015] which measure how informative and diverse the generated replies are. Distinct-i measures the ratio of distinct i-gram in replies.

Human Evaluation: Considering the limitation of the existing automatic evaluation metrics, we also adopt human judgments to evaluate the quality of generated replies. Particularly, we randomly sampled 100 cases and invited six well-educated volunteers to evaluate the generated replies from the

Query	S2SA	S2SA+MMR	S2SAC+MMR
已经有个女儿了好吧, 一岁了 She has a daughter who is one year old.	我也是这么想的 I think so.	好吧，那你是个孩子 Well, you are a child.	我觉得她的女儿好可爱 Her daughter is very lovely.
您好，请问主卧租出去了么？情侣租可以吗? Hello, has the master bedroom been rented out? Is it OK for a couple to rent it?	不好意思，不好意思。 Sorry, sorry!	可以，可以来看房 Yes, you can come and have a look.	房子已经租出去了 Room has been rent out.
每次看到总会想起小时候看的铁胆火车侠 It reminds me of Hikarian everytime I watch it.	我也是，我也觉得 I think you surely are.	我觉得你也是 I think you surely are.	哈哈，小时候看的时候很喜欢 LoL, I liked it very much when I was watching it as a child.

Table 4: Case study, comparing the two best performed baselines with our proposed model. We translate Chinese query-reply pairs to English.

5 models. For each query-reply pair, volunteers are asked to rate it with a score from {0,1,2}. 2 means query and reply are highly related and the reply is fluent; 1 is for the reply that is related but not informative enough; 0 indicates that they are irrelevant or the reply contains grammatical errors. For each model, we calculate the ratio of each score (0, 1 and 2) and its average score as the human evaluation result.

4 Experimental Results

4.1 Automatic Evaluation

Table 2 gives an overview of the automatic evaluation on the five models. Compared with other baselines, S2S shows a poor performance on both BLEU and Distinct. It can be seen from the table that though S2SA+MMR improve the Distinct value comparing with S2SA, its BLEU is much lower then S2SA. It is apparent that the calibrated models (both S2SAC and S2SAC+MRR) significantly outperform the baselines on both BLEU-1∼4 and Distinct-1∼4, which verifies the effectiveness of our proposed approach. Compared with S2SA, S2SAC achieves 33% improvement on BLEU-4 and 37% improvement on Distinct-4. For S2SAC+MMR, it also shows a better performance than S2SA+MMR. The improvement raised by calibration mechanism on both S2SA and S2SA+MMR indicates that the proposed framework is practical to be applied to any other generative models.

4.2 Human Evaluation

Table 3 demonstrates the results of the human evaluation. The ratio is calculated by combining all the annotations together. To examine the agreements among all the volunteers, we calculate the Fleiss' kappa of the human annotations on the five models. The results are all around 0.35, which demonstrate the fair inter-human agreements. Compared with baselines, the calibrated models lower the ratio of score 0 and gain a larger ratio on score 2, indicating that our proposed model could generate more informative and diverse replies. The average scores of the five models also demonstrate the calibrated model significantly outperforms the baselines.

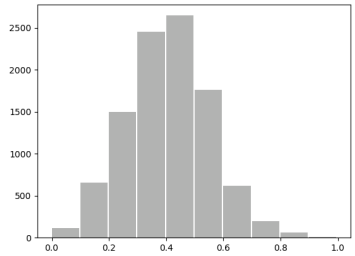

Figure 4: Score distribution of 10,000 query-reply pairs calculated by the calibration network. Horizontal axis denotes the score and the vertical axis means the number of pairs.

4.3 Analysis

The following paragraphs show some analysis on the performance of both calibration networks and the calibrated generation networks. We conduct an experiment to illustrate the score distribution produced by the calibration networks. We then show some cases to help verify the effectiveness of our proposed model.

Calibration score. In our proposed framework, we elaborate a calibration network to measure the quality of query-reply pairs. In this section, we give some analysis on the scores that the calibration network predicts in detail. For better visualization, we normalize the scores s by $s_i = \frac{s_i - min(s)}{max(s) - min(s)}$. The calibration network is designed to give a score in the range $[0, 1]$. We randomly sample 10,000 query-reply pairs from the training set and plot their scores. The score distribution is illustrated in Figure 4. It can be seen that these scores concentrates on range $[0.2, 0.6]$. There are around 10% cases whose score is lower than 0.2, which indicates that almost 10% training cases are of low quality. Thus it is essential to control the effect of those low-quality cases on the final model.

Case study of calibration network. To further analyze the calibration network qualitatively, we randomly choose two cases whose scores are lower than 0.2. As shown in Table

Score	Query	Reply
0.179	可是，我还是很喜欢北京 But I still love Beijing very much.	注定孤独一生？ Destined to be alone forever?
0.154	肤色不错，有点痘痘没事的 Your complexion is good. These pimples doesn't matter.	你参加过什么相亲节目吗？ Have you joined any dating shows?

Table 5: Cases of low scores given by the calibration network.

5, the replies of the two cases are totally irrelevant to their replies. Such cases will mislead the generation network while training.

Case study of generation with calibration network. Table 4 compares S2SAC+MMR with S2SA and S2SAC using some examples. From the cases shown in it, we can see that the S2SA and S2SA+MMR more or less suffer from generating replies that are lack of information or irrelevant to the given queries. As our S2SAC+MMR is trained under the guidance of a calibration network, it benefits more from those high quality data. Thus it is promoted to produce replies that are more informative and relevant.

5 Related Work

5.1 Generation Based Dialog System

With the growth of publicly available conversation data such as social media conversations, data-driven dialog system enjoys significant advances. Ritter *et al.* [2011] proposed a data-driven approach adapted phrase-based statistical machine translation for response generation. Inspired by that, Shang *et al.* [2015] employed seq2seq model for dialog system, which yielded excellent results. Recently, more and more researchers have focused on generation-based conversation system with seq2seq framework [Li *et al.*, 2015; Song *et al.*, 2017]. Their goals are to build interesting, intelligent and flexible dialog systems.

One common drawback of neural networks based dialog system is that the responses are highly repetitive and boring [Li *et al.*, 2015; Song *et al.*, 2017]. To enrich the diversity of response, Li *et al.* [2015] proposed a diversity-promoting objective function. Li *et al.* [2016] trained seq2seq model with user embedding to capture personality speaking style. Xing *et al.* [2017] used topic information to a joint attention to generate topic-related responses. Some other works also paid attention to dialog system with diversity, personality and knowledge-base. However, the existing generation models do not consider the quality differences in data. In this work, we focus on calibrating the generation network according to the quality of training samples.

5.2 Dialog Evaluation

Our calibration network adapts the automatic evaluation metrics for dialog system to evaluate the quality of data. Therefore, our work is closely related to the automatic evaluation of dialog systems. In this section, we briefly summarize some advanced automatic evaluation metrics.

Liu *et al.* [2016] revealed that traditional evaluation methods based on word-overlap (*e.g., BLEU, METEOR, ROUGE*) correlate weakly with human annotation. To address this problem, some researchers investigated neural network based evaluation metrics. Lowe *et al.* [2017] learned the representation of context, model response and reference response using a pre-trained model, and then optimized a score function to calculate an evaluation result. Tao *et al.* [2017] proposed a referenced and unreferenced blended evaluation routine (RUBER) for dialog system. Both of the abovementioned methods show a high correlation between model scoring and human scoring.

5.3 Instance Weighting

Instance weighting is a line of research that assigns instance-dependent weights to the loss function. Previous research adopted this method to the domain adaption tasks in NLP [Jiang and Zhai, 2007; Rebbapragada and Brodley, 2007; Wang *et al.*, 2017] to address the noise label issue. In domain adaption tasks, noise data is clearly defined and could easily be distinguished between clean data. Unlike domain adaption tasks, the "noise data" for training a conversation system is difficult to define, as the conversation data is of high diversity. Lison and Bibauw [2017] proposed to investigate instance weighting into retrieval-based dialog system. In this work, we elaborate a calibration network to measure the quality of data, and then incorporate the weight given by it into the generation model in an efficient and effective way.

6 Conclusion and Future Work

In this paper, we consider the training data quality for the open-domain dialog systems. To address the noisy training data problem, we propose a generation with calibration framework to measure the qualities of the training instances and utilize the information to improve the training of the generation model. Experiments show that our framework outperforms the traditional generation models on both automatic evaluation and human evaluation metrics.

In the future, we plan to study the effectiveness of the calibration framework on multi-turn dialog systems. Besides, we would like to apply the calibration network to retrieval-based dialog system, and investigate whether the calibration network and the generation network can be mutually beneficial by joint training.

Acknowledgments

We thank Zhengwei Tao for discussions and the anonymous reviewers from IJCAI 2018 for their constructive feedback on this paper. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (No. 61672058), and Contract W911NF-15-1-0543 with the US Defense Advanced Research Projects Agency (DARPA). Rui Yan was sponsored by CCF-Tencent Open Research Fund and MSRA Collaborative Research Program.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Hu *et al.*, 2014] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

[Jiang and Zhai, 2007] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271, 2007.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2015] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[Lison and Bibauw, 2017] Pierre Lison and Serge Bibauw. Not all dialogues are created equal: Instance weighting for neural conversational models. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 384–394, 2017.

[Liu *et al.*, 2016] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[Lowe *et al.*, 2017] Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*, 2017.

[Lu and Li, 2013] Zhengdong Lu and Hang Li. A deep architecture for matching short texts. In *Advances in Neural Information Processing Systems*, pages 1367–1375, 2013.

[Mou *et al.*, 2016] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*, 2016.

[Rebbapragada and Brodley, 2007] Umaa Rebbapragada and Carla E Brodley. Class noise mitigation through instance weighting. In *European Conference on Machine Learning*, pages 708–715. Springer, 2007.

[Ritter *et al.*, 2011] Alan Ritter, Colin Cherry, and William B Dolan. 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, 2011.

[Serban *et al.*, 2016] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784, 2016.

[Shang *et al.*, 2015] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[Song *et al.*, 2017] Yiping Song, Zhiliang Tian, Dongyan Zhao, Ming Zhang, and Rui Yan. Diversifying neural conversation model with maximal marginal relevance. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 169–174, 2017.

[Sordoni *et al.*, 2015] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[Tao *et al.*, 2017] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *arXiv preprint arXiv:1701.03079*, 2017.

[Vinyals and Le, 2015] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[Wang *et al.*, 2017] Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, 2017.

[Xing *et al.*, 2017] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *AAAI*, pages 3351–3357, 2017.

[Yan *et al.*, 2016] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64. ACM, 2016.